REFINING COMPARATIVE PROTEOMICS BY SPECTRAL COUNTING TO ACCOUNT FOR SHARED PEPTIDES AND MULTIPLE SEARCH ENGINES

By

Yao-Yi Chen

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Biomedical Informatics

August, 2012

Nashville, Tennessee

Approved:

Professor David L. Tabb

Professor Bing Zhang

Professor Ming Li

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. David L. Tabb, for his patience, invaluable support, supervision and helpful suggestions through my research. I am also grateful to my other dissertation committee members, Dr. Ming Li and Dr. Bing Zhang, who were very supportive and provided valuable advice on my project.

I am grateful to members in Tabb group, Dr. Surendra Dasari, Dr. Zeqiang Ma, Matt Chambers, Dong Wang and Jay Holman for their tremendous help in my research. I would like to thank Dr. Lorenzo Vega Montoto for his biostatistical advice and Dr. Patrick Halvey for providing data sets and helpful suggestions for my work.

I am also grateful to the Department of Biomedical Informatics for their help and cooperation over the last year. I would like to especially thank Dr. Cynthia Gadd and our program coordinator Rischelle Jenkins for their support during my transfer from Biochemistry Department and during my graduate study so far.

Finally, I would like to thank my friends and family for being ever understanding and supportive.

Thanks to National Cancer Institute grant U01 CA152647 for supporting my research work.

ABBREVIATIONS

CPTAC	Clinical Proteomic Tumor Analysis Consortium
FDR	False Discovery Rate
MS	Mass Spectrometry
MS/MS	Tandem Mass Spectrometry
ppm	parts per million
PSM	Peptide Spectrum Match
PTM	Post-Translational Modification
SDS-PAGE	Sodium Dodecyl Sulfate PolyAcrylamide Gel Electrophoresis
MM	MyriMatch
SQ	Sequest
TR	TagRecon
XT	X!Tandem

ABSTRACT

Spectral counting has become a widely used approach for comparing protein abundance in label-free shotgun proteomics. However, when analyzing complex samples, the ambiguity of matching between peptides and proteins greatly affects the assessment of peptide and protein differentiation. Meanwhile, the configuration of database searching algorithms that assign peptides to MS/MS spectra may produce different results. Here, I present three strategies to improve comparative proteomics through spectral counting. I show that comparing spectral counts for peptide groups rather than for protein groups forestalls problems introduced by shared peptides. I present four models to combine four popular search engines that lead to significant gains in spectral counting differentiation. Among these models, I demonstrate a powerful vote counting model that scales well for multiple search engines. I also show that semi-tryptic searching outperforms tryptic searching for comparative proteomics. Overall, these techniques considerably improve protein differentiation on the basis of spectral count tables.

TABLE OF CONTENTS

ACKNOWLEDGMENTS....ii

ABBREVIATIONS....iii

ABSTRACT....iv

LIST OF FIGURES AND TABLES....vii

Chapter

1. INTRODUCTION 1

1.1 Overview 1
1.2 Shotgun Proteomics 3
1.3 Comparative Shotgun Proteomics 4
1.4 Proteomic Data Analysis 7
1.4.1 Peptide Identification 8
1.4.2 Protein Inference 14
1.4.3 Tryptic Search vs. Semi-Tryptic Search 18

2. MATERIALS AND METHODS 20

2.1 Data Sources.... 20
2.1.1 ABRF Data 20
2.1.2 CPTAC Data 21
2.1.3 HNSCC Data 21
2.1.4 ASW480 Data 22
2.2 Database Search Pipeline 23
2.3 Model for Peptide Group-Based Spectral Count Differentiation 24
2.4 Models for Combining Search Engines 26

3. RESULTS AND DISCUSSION 29

3.1 Peptide Group-Based Spectral Count Differentiation Improves Protein Differentiation 29
3.2 Combining Multiple Search Engines Improves Protein Differentiation 33
3.3 Semi-Tryptic Search Outperforms Tryptic Search in Protein Differentiation 37

4. CONCLUSIONS 42

REFERENCES 45

Appendix 51

A. Supplementary Materials 1 51

B. Supplementary Materials 2 59

MyriMatch Configurations 59
Sequest Configurations 61
TagRecon Configurations 63
X!Tandem Configurations 65

C. Supplementary Materials 3 68

R Code Implementation for Vote Counting Model 68

D. Supplementary Material 4 73

Peptides Corresponding to Peptide Groups in ASW480 dataset Table S2 73

E. Supplementary Material 5 75

Peptides Corresponding to Peptide Groups in HNSCC dataset Table S3 75

LIST OF FIGURES AND TABLES

Figure 1. The workflow of shotgun proteomics 4

Figure 2. Workflow of shotgun proteomics data analysis. 7

Figure 3. Four peptide identification strategies 9

Figure 4. Overview of database search. 10

Figure 5. Protein inference in shotgun proteomics. 15

Figure 6. Protein assembly, protein groups and peptide groups in IDPicker 17

Figure 7. Number of true positive proteins out of the Top 250 of proteins of corresponding combination of search engines in ABRF *E.coli* dataset. 34

Figure 8. Number of true positive proteins in top 50 differentiated proteins using different combination of search engines in CPTAC Study6 dataset. 35

Figure 9. Comparison of fully tryptic and semi-tryptic searching in ABRF dataset. 38

Table 1. Fully tryptic versus semi tryptic search in Yeast Sample with Spiked Human Proteins 39

Table S1. Data sets, Search Engines, Protein Sequence Databases used in this study. 51

Table S2. Peptide-to-protein Table of Desmin and Vimentin in ASW480 Dataset 52

Table S3. Peptide-to-protein Table of Myosin 14 in HNSCC Dataset 52

Table S4. Average number of fully and semi-tryptic peptides confidently identified (rank1) by both searches 53

Figure S1. The “Red/Yellow” iPRG 2009 LC-MS/MS data set with “Blue/Green” LC-MS/MS answer keys. 54

Figure S2 Venn diagram of differentially expressed proteins at protein level and peptide group level in two datasets 55

Figure S3 Venn diagram of differentially expressed protein at protein and peptide group level in ASW480 replicates of the same groups 56

Figure S4. Correlation coefficient between protein and peptide group-based p-values……….56

Figure S5. Venn diagram comparing the differential proteins identified by different search engines in ABRF dataset 57

Figure S6 Venn diagram comparing the differential proteins identified by different search engines in CPTAC dataset 58

CHAPTER I

INTRODUCTION

1.1 Overview

Shotgun proteomics based on tandem mass spectrometry has become a widespread method for analyzing complex biological mixtures. It begins by digesting protein mixtures and separating the resulting peptides by liquid chromatography. After peptide MS/MS spectra are acquired, they are matched to database peptide sequences by search engines such as Sequest[1], Mascot[2], X!Tandem[3], and MyriMatch[4]. Proteins are assembled from these raw identifications by validation tools [5-9] that convert arbitrary search scores into statistical measures [10]. Proteins can then be filtered by customized criteria for further analysis. Because shotgun analyses can represent complex proteomes in considerable depth, a key question is how one can compare shotgun proteome inventories to reveal molecular characteristics of biologically distinct phenotypes to discover clinically important biomarkers. Improvement in protein differentiation broadly benefits the identification and validation of molecular markers that relate to various biological or medical outcomes, thus improving the current state of the art in biological research and clinical practice.

In shotgun proteomics, the link between peptides and proteins is lost through the digestion of protein mixture. Determining which protein these shared peptides arose from is a challenge in comparative proteomics. A particular peptide may correspond to multiple potential protein sources. In systems where proteins of multiple species are present, such as xenograft models of cancer, shared peptides are very common, and so a difference in one protein may masquerade as a difference in a second protein that shares peptides with the first.

Moreover, search results differ from one search engine to the next, depending on both the type of mass spectrometer used and the configuration of the search. In biological samples, often the most interesting proteins are lowest in abundance, and meaningful changes in protein abundance may be small in magnitude. Detecting these differences may be visible by one search engine but not another because of differences in match scoring. Even if the search engine is held constant, the way in which the tool is configured may significantly impact the set of identifications produced. Deciding between a “fully tryptic” search and a “semi-tryptic” search would seem to primarily impact the amount of time required, but this decision has been shown to significantly alter the set of peptides identified from a mixture.

Here, I characterize three strategies for improving comparative proteomics through spectral counting. First, I will demonstrate that the problem of shared peptides can be resolved through comparison for peptide groups rather than proteins, giving examples of differences that would be confused by standard approaches. Then, I will examine the gains achieved for spectral counting when collating search results from a set of four high-performance peptide identifiers. I will also determine the impact of tryptic and semi-tryptic searching for spectral count tables to frame recommendations for best practices. Taken together, these techniques enable higher quality differentiation on the basis of spectral count tables.

In this chapter, I will provide an introduction to shutgun proteomics, comparative shotgun proteomics, and the workflow of proteomics data analysis, including peptide identification, protein inference and protein assembly.

1.2 Shotgun Proteomics

Shotgun proteomics is currently the most commonly used approach for identification and quantification of large number of proteins. It has been proved to be successful in post-translational modification identification, protein quantification, and protein-protein interaction[1]. The workflow of shotgun proteomics is illustrated in Figure 1. First, taking the sample of a mixture of proteins, reduce the complexity by SDS-PAGE or two-dimensional gel electrophoresis. Then the proteins are digested into peptides by sequence-specific proteolysis. Trypsin is the most commonly used protease that cleaves peptide at the C-terminal side of arginine and lysine. The peptide mixtures are then separated by liquid chromatography and ionized in a mass spectrometry. Peptides are isolated in mass spectrometer and characterized by tandem mass spectrometry (MS/MS), which involves breaking the peptide into smaller fragments and measuring the mass spectrum of these fragments. During data analysis, the peptides are identified from the tandem mass spectra. Then proteins are assembled from the peptides.

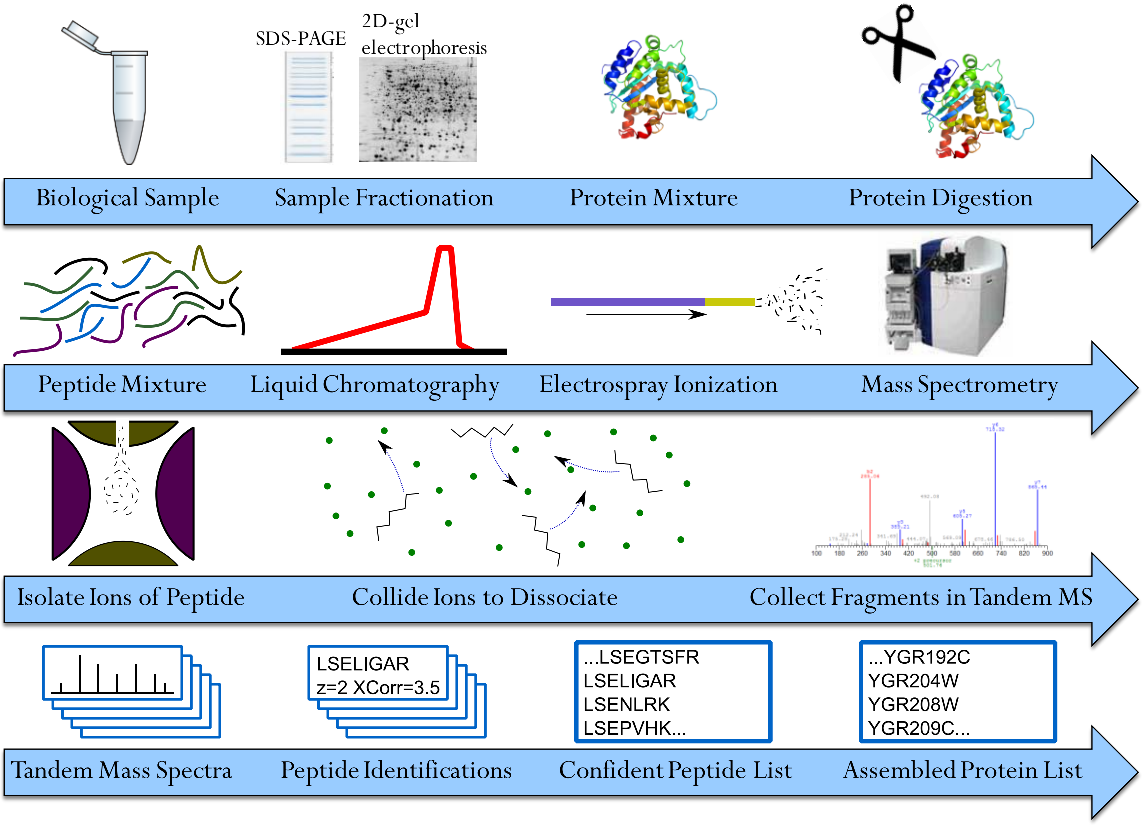


Figure 1. The workflow of shotgun proteomics

1.3 Comparative Shotgun Proteomics

An important goal in proteomics is to globally profile changes in protein abundances in biological systems, thus discovering protein expression state in response to biological perturbation, disease progression or drug treatment. In general, protein quantification by mass spectrometry is performed by stable isotope labeling or a label-free approach. A number of methods of stable isotope labeling of proteins or peptides, including Isotope-Code Affinity Tag (ICAT)[2], Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC)[3], Isobaric Tags for Relative and Absolute Quantification (iTRAQ)[4] are used. However, label based quantification methods have common limitations including requirements for higher amounts of biological samples, increased complexity in experiments, and high costs for the labeling reagents. Therefore, label-free shotgun proteomics have emerging and developed as an alternative for protein quantification and differentiation. Compared to label-based quantification, label-free quantification has the following advantages: (a) There is no limit to the number of experiments to be compared while label-based approach, for example, SILAC is limited to 2-8 experiments that can be directly compared; (b) label-free methods provide higher dynamic range of quantification, thus will benefit analysis when large and global protein changes between experiments are observed[5]. Currently, two label-free quantification methods are used: (a) precursor intensity approach: measuring and comparing mass spectrometric signal intensity of peptide precursor ions of a given protein; (b) spectral counting approach: counting the number of MS/MS spectra matched to peptides of a given protein. To use peptide precursor intensities, for every peptide, ion chromatograms are extracted from MS/MS run and their precursor MS peak areas or peak intensities are integrated over the chromatographic time or collected. The intensity for each peptide from one experiment can be compared to other experiments to yield relative quantitative information. Prior work has demonstrated that a frequency-based analysis approach using the number of observed spectral counts for each protein provides a rough measure of protein levels in complex protein mixtures, especially for more abundant proteins[6-8]. Thus, relative abundance can be calculated by comparing the number of spectra between multiple experiments.

Comparing spectral counting and precursor intensity method, previous studies have shown that spectral counting can be as sensitive as ion peak intensities considering detection range, correlation and linearity. Spectral counting is more sensitive for detecting changes in protein abundance, whereas peak area provides more accurate estimates of protein ratios[9].

Comparative proteomics spans two complementary goals. First, researchers may seek to *differentiate* the proteomes of two sample cohorts, seeking the proteins that appear in one sample to a significantly greater degree than in another. Second, researchers may seek to *quantify* the extent to which proteins change in magnitude between sample cohorts. Here, I consider the first of these goals, leaving quantification as a topic for experimental methodologies better designed for this purpose, such as selected reaction monitoring[10] . The evidence produced for a protein in "shotgun" experiments is the result of a high-throughput sampling process. As a result, which spectra are captured from a particular protein digest will vary among experiments [11]. Spectral counts attributed to a particular protein group may vary due to random sampling or due to differences in protein quantity. In general, one expects to collect both more spectra from individual peptides (potentially varying in charge state or modification state) as well as more peptides from a particular protein group as the concentration of that protein rises compared to the sample background. As a result, finding significant differences requires the ability to compare variation in replicates to variation between cohorts. Here I focus on refining spectral counting method to achieve better protein differentiation.

1.4 Proteomic Data Analysis

Usually, hundreds of thousands of tandem mass spectra are collected in a single shotgun proteomics experiment. Thus bioinformatics tools are required for proteomics data analysis.

The typical proteomics data analysis workflow is shown in Figure 2. First, experimental spectra are interpreted as peptides with strategies discussed in next section. Next, peptide identifications are validated to estimate the false discovery rate of the confidence of the assignments. Peptides identified with high confidence are used to assemble proteins.

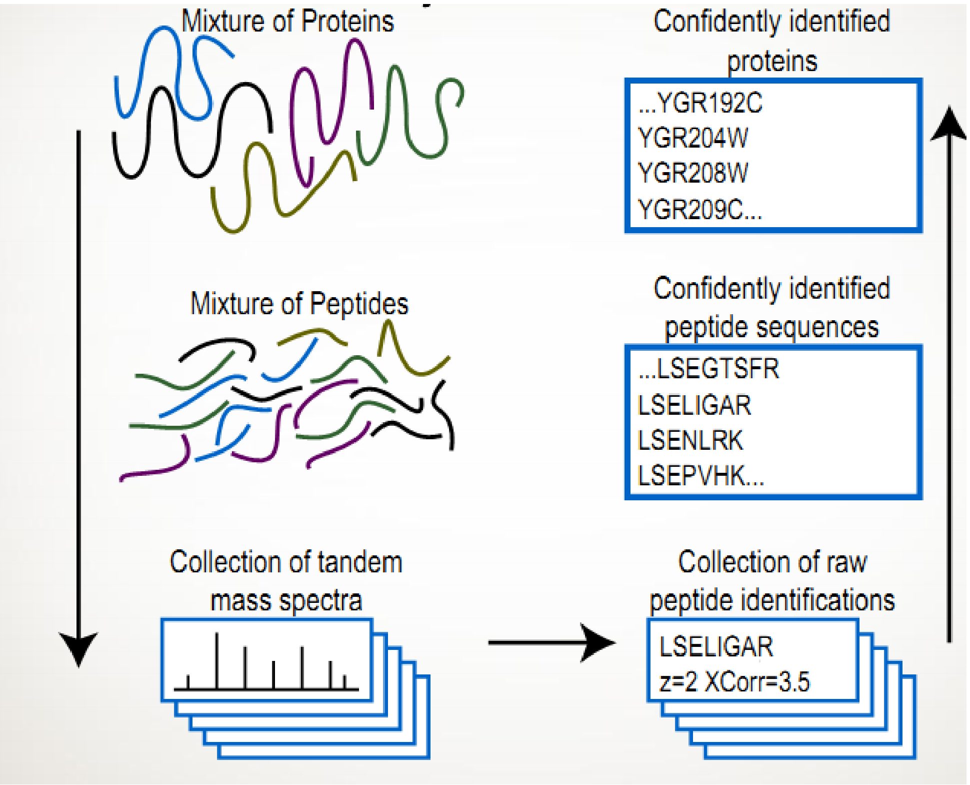

Mixture of Proteins
Mixture of Peptides
Collection of tandem mass spectra
Collection of raw peptide identifications
LSELIGAR
z=2 XCorr=3.5
Confidently identified peptide sequences
...LSEGTSFR
LSELIGAR
LSENLRK
LSEPVHK...
Confidently identified proteins
...YGR192C
YGR204W
YGR208W
YGR209C...

Figure 2. Workflow of shotgun proteomics data analysis.

1.4.1 Peptide Identification

Database search algorithms are essential to link tandem mass spectra (MS/MS) to peptide sequences from protein database. There are many search algorithms that are currently in use. Here I will introduce four major peptide identification approaches (Figure 3), with focus on four of the most popular search engines- Sequest[12], X!Tandem[13], MyriMatch[14] and TagRecon[15,16].

Sequence Database Search

A general approach in Sequest, MyriMatch and X!Tandem are known as sequence database (DB) search. The overview of sequence database search is illustrated in Figure 4. First, the experimental precursor ion mass of a peptide is compared with calculated peptide mass; database search tools perform an *in-silico* digestion of the protein database to enumerate all candidate peptide sequences that has the mass within the mass tolerance range. Then, tandem mass spectra are then matched to the fragment ion mass values. Corresponding mass values are counted or scored in a way that allows the identification of peptides best matches the data[17]. Database search tools then select the top ranked peptides of each spectrum for subsequent analysis.

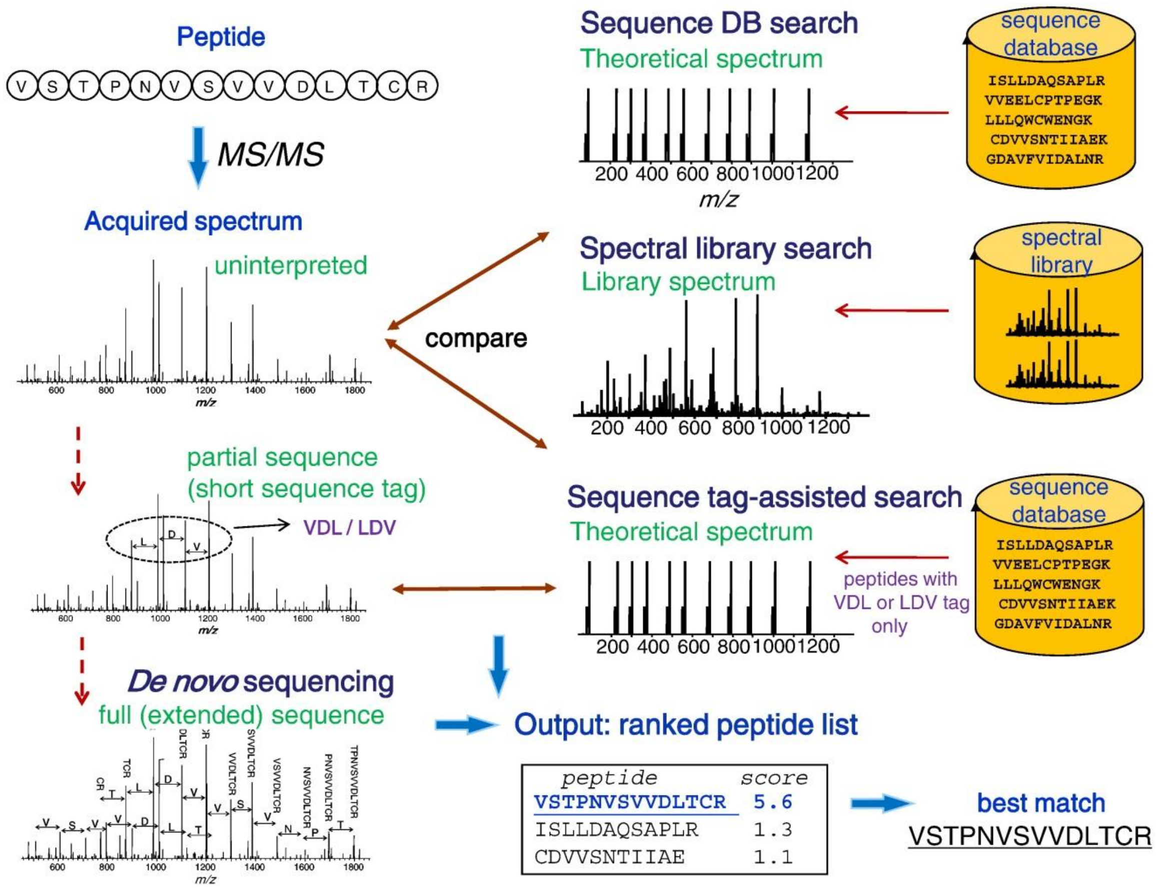


Figure 3. Four peptide identification strategies

Sequence DB search, spectral library search, sequence tag-assisted search and *De novo* sequencing search are used for peptide identification [18].

The scoring function is critical for measure the similarity between experimental and theoretical spectra. Scoring functions such as correlation fuctions (cross correlation in Sequest and dot product in X!Tandem) and probability-based function (ion score and identity score in Mascot and multivariate hypergeometric distribution score in MyriMatch) are used to evaluate the peptide-spectrum-matches (PSMs).

Sequest uses a cross-correlation function to provide a measurement of similarity between mass-to-charge ratios for the fragment ions from the observed tandem mass spectrum and the fragment ions predicted from amino acid sequences obtained from database. Difference between normalized cross-correaltion function of the first and second-ranked search results shows the confidence of match between sequence and spectrum.

MyriMatch first stratifies peaks into multiple intensity classes, and then scores peptide matches based upon the multivariate hypergeometric (MVH) distribution on the basis of peak intensity. The scoring of peptides pays greater emphasis on matching intense peaks, which in result gains considerable discriminative power.

X!Tandem generates theoretical spectra for the peptide sequences using knowledge of the intensity patterns associated with particular amino acid residues, and calculate an empirical E-value to access the significance of a peptide match. Peptide candidate score distributions are utilized for thresholding or E-value extrapolation.

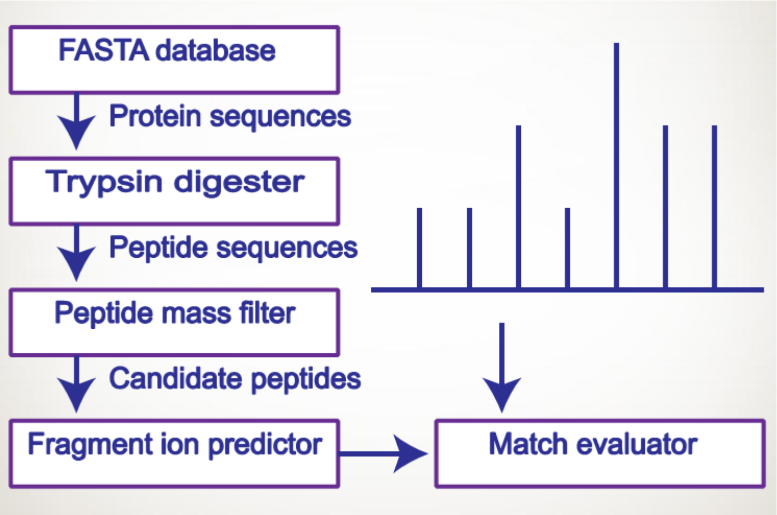


Figure 4. Overview of database search.

In these DB search tools, search parameters have great impact on search results. First, precursor mass tolerance determines the peptide candidates to be compared to the experimental spectrum. Mass tolerance window varies by instruments for collecting MS data, high mass accuracy instruments such as Orbitrap allow a very narrow mass window ~10ppm, while low mass accuracy instruments such as LTQ require a broader mass window ~3 Da. Narrower mass window reduces search time and decreases number of false matches. Second, enzyme digestion specifications constraint also controls the number of candidate peptides to be analyzed. Tryptic search eliminates identification of peptides that undergo unexpected cleavages and spend less time than non-tryptic searches or semi-tryptic searches. I will discuss the difference further in the following sections. Other search parameters such as post-translational modifications and reference protein database can also affect search results.

In spite of the wide spread usage of database search, database search tools rely heavily on protein databases, in which some of the genome sequences and annotation may not be accurate. Especially, mutations and modifications are often ignored by database search tools. Moreover, database search is very time-consuming process for the large number of comparisons between observed spectra and theoretical spectra.

Sequence tagging-based database search

Sequence tagging-based database search first infers short peptide sequences ("tags") from spectra. The tags are then used to match candidate peptides via database search. Sequence tagging-based database search is particularly useful in the identification of mutation and modifications. Tools such as InsPecT [19] and TagRecon[16] are examples of sequence tagging-based database search.

TagRecon works with DirecTag as an integrated bioinformatics pipeline. DirecTag infers sequence tags from MS/MS spectra. TagRecon detects a sequence tag that matches a peptide sequence reconciling mass differences, and then compares the mass of flanking regions of both spectrum and peptide sequence to determine whether the masses matches is within a specified mass tolerance[16]. TagRecon uses two probabilistic subscores: an intensity-based probabilistic MVH score and a nonprobablistic fast cross-correlation (XCorr) score to indicate the confidence of the PSMs.

De novo sequencing search and spectral library search

De novo sequencing infers peptide sequences directly from experimental spectra. The inferred peptides then mapped to proteins by downstream tools such as MS-BLAST[20]. This approach is especially useful when the organisms have unsequenced or partial sequenced genomes.

Spectral library search matches MS/MS scans to a spectral library, a large collection of observed spectra that are confidently identified in previous experiments. It is a very efficient and accurate way for peptide identification. However, the assignment of peptides to observed spectra is largely affected by the completeness and accuracy of assembled spectral libraries.

Combining Search Engines

The real proteomic samples are complex. Often the most interesting proteins are at low abundance and could not be discriminated due to the inference of other proteins and noises, especially when dealing with low quality spectra. Therefore, two search engines can provide very different results for the same sample [21].

Despite improvements in mass spectrometry instruments and peptide identification algorithms, a significant number of high quality MS/MS spectra left unassigned to peptide sequences or have scores below confidence thresholds[18]. This can be partially due to the deficiencies of scoring schemes implemented in the software to rank candidate peptides and select the best match for each experimental spectrum, resulting in loss of sensitivity in complex samples. A peptide may be identified by one search engine but blind to another due to their different scoring systems. Spectral counting depends upon identification, and yet little evaluation of its dependence on search engines has appeared in the peer-reviewed literature. Integrating results from search algorithms is a promising strategy to improve peptide and protein identification confidence by reducing noise and utilizing complementary strengths. Several approaches have been proposed for integrating search results. Alves *et al.* proposed combination of independent p-values from multiple search engines into a meta-analytic p-value for each peptide[22]. Searle *et al*. proposed a framework to combine the results of multiple search engines using Bayesian rules and the expectation maximization learning algorithm[23]. However, a peptide-centric model for combination of different search tools suffers from the difficulties from the lack of a common statistical standard.

Kwon et al. proposed a probabilistic approach by first converting raw search scores from search engines into a probability score for every possible PSM accounting for the correlation between scores, and control an unified false discovery rate for data integrated from different search engines[24]. Stepping beyond a single search scenario, researchers have demonstrated that collating results from multiple search engines improves sensitivity for inferring protein inventories [25,23,24], so long as false positives are kept under control. It would seem that the improved coverage available through multiple search engines would be a boost for differentiation, as well. How to leverage the increased information yield, however, has not yet been described.

Here I will first compare the search results from different search engines on the same datasets in identifying differential proteins, then propose and compare four new approaches to combine search results at protein level, and examine the gains achieved for spectral counting when collating search results from a set of four high-performance peptide identifiers, thus providing new insights into integrating search tools to achieve better protein differentiation.

1.4.2 Protein Inference

Identification of peptides resulting from proteolytic digestion of proteins is only an intermediate step to identify and quantify proteins. The ultimate goal of a study is to identify and quantify proteins in the analyzed sample. One of the problems for protein differentiation arose from shared peptide in the task of assembling the sequences of identified peptides to infer protein content of the sample (Figure 5).

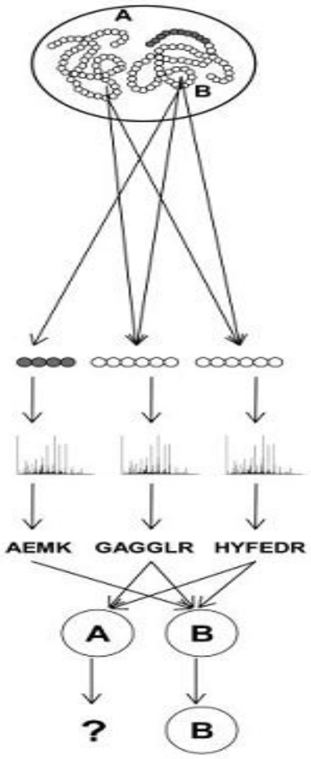


Figure 5. Protein inference in shotgun proteomics.

In this example, the sample contains two proteins, A and B, which share sequence homology. The three identified peptides, AEMK, GAGGLR, and HYFEDR are present in protein B, and GAGGLR, HYFEDR are present in protein A. in the shotgun approach, the connectivity between peptides and proteins are lost. No information on the number or properties of proteins in the samples is available. It is not possible to conclude the presence of A for B can account for all observed peptides[26].

For these peptides that are shared between two or more proteins, the abundance of the peptide is a combinational effect of multiple proteins, leading to ambiguities in protein differentiation by the redundancy. When using protein-based spectral count differentiation, determining which protein these shared peptides arose from is a challenge to comparative proteomics. After peptide validation, incorrect PSMs can still be accepted for protein inference. A commonly used approach for protein inference and error estimation is to use a target-decoy strategy for database search and apply various filters to control output proteins at a specific protein-level FDR. Protein filters including minimum number of distinct peptides to infer a protein, minimum number of spectra per protein and peptide FDR are usually used to remove incorrect proteins.

Protein parsimony is widely accepted by proteomic community. The central idea is to present results of large scale shotgun experiments in terms of minimal lists of protein identifications. Nesvizhskii *et. al* illustrated differentiation of proteins on the basis of identified peptides[26]. Zhang *et. al* modeled peptide-protein relationships in a bipartite graph and identified protein clusters with shared peptides and to derive the minimal list of proteins[27]. The software- IDPicker, a protein assembly tool [28,27], organizes peptides into groups when they match identical sets of proteins, and it similarly organizes proteins into groups when they match identical sets of peptides (Figure 6). This structure enables the development of methods to differentiate proteomes in units of “peptide groups” that do not overlap with each other. This approach is based on the assumption that a high degree of similarity exists in the relative expression level of different proteins in the same protein group [29].

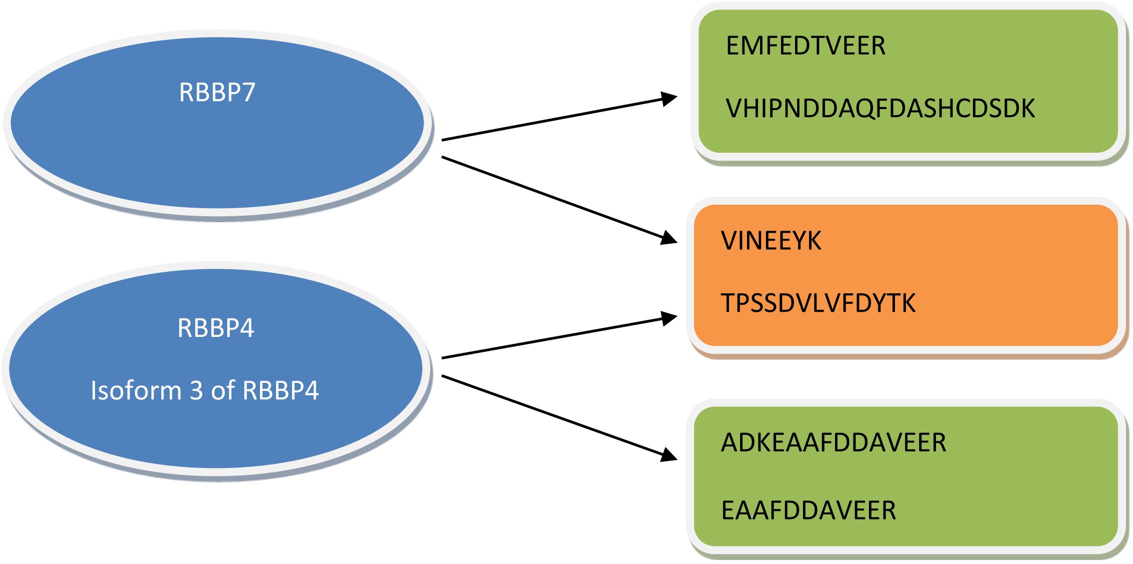


Figure 6. Protein assembly, protein groups and peptide groups in IDPicker
In this diagram, three peptide groups are associated with two protein groups. IDPicker groups

peptides, such as the two peptides in the orange box, to "peptide groups" when they match to exactly the same proteins- Histone-binding protein RBBP7 and RBBP4. "Protein Groups" are sets of proteins such as RBBP4 and isoform 3 of RBBP4 that are indiscernible on the basis of the observed peptides. Peptide groups that only associate with one protein group are called unique peptide groups (green box). Peptide groups that associate with more than one protein group are called shared peptide groups (orange box).Both the protein and peptide groups are shown in IDPicker reports.

Several approaches have been proposed to improve the quantification and differentiation of proteins [30,27,29,31]. One approach is to discard shared peptides during protein quantification. However, previous studies have found that eliminating shared peptides from analysis eliminates protein inference but may significantly decrease the number of proteins for which relative abundance can be obtained[29]. Fermin *et. al* describes a method to adjust spectral counts to accurately account for peptides shared across multiple proteins by spectral counts of unique peptides. However, this approach has the risk of attempting to apportion large numbers of spectra on the basis of relatively small sets of differentiating spectra[30]. Here, I propose a new approach for protein differentiation based on peptide groups, which forestalls problems introduced by shared peptides.

1.4.3 Tryptic Search vs. Semi-Tryptic Search

In shotgun proteomics, proteins are usually digested by trypsin followed by liquid chromatography mass spectrometry. One of the problems is the low coverage of peptides when analyzing complex protein samples. It has been shown that only ~10-15% of all tryptic peptides from a given protein sample can be identified [32] by search engines. Non-tryptic or semi-tryptic peptides are generated from the truncation of regular tryptic peptides before separation. In semi-tryptic search, one end, but not both of the peptide is allowed to diverge from the expected cleavage site. Peptide truncation can be caused by several factors such as *in vivo* biological mechanism or various chemical mechanisms during sample preparation, handling and storage. When searching peptide tandem mass spectra against sequence database, peptides identified are conformed to the search parameters-fully tryptic, semi-tryptic or non-tryptic, where a trade-off of false positives and false negatives will be yielded. Deciding between a “fully tryptic” search and a “semi-tryptic” search would seem to primarily impact the amount of time required, but this decision has been shown to significantly alter the set of peptides identified from a mixture [28]. Many research groups consider only peptides that under a rigorous "fully tryptic" cleavage rules for protein database search, whereas other groups allow "non-tryptic" or "semi-tryptic"peptides. It has been shown that non-tryptic peptide search overdoes the noice and specificity is impaired. Furthermore, it is much more complex to compute and requires a large and counter-productive increase in search time. Olsen *et. al* used the high mass accuracy of a linear ion-trap-FTICR mass spectrometer to exclude precursor ions with less than 1 p.p.m mass accuracy and found that trypsin cleaves solely C-terminal to arginine and lysine[33]. This work provided evidence to support fully tryptic search. However, the rigorous mass filter excluded the non-tryptic peptides, which composed a large portion of the overall experimental peptides. Moreover, this result is not applicable to lower sensitivity and mass accuracy experiments. It is found that although for a given protein, semi-tryptic peitdes might be generated at lower probability than the tryptic peptides, a high concentration protein often contribute large numbers of semi-tryptic peptides comparative to tryptic peptides of low concentration proteins in being selected for fragmentation[28]. However, the impact of trypsin specificity configurations on protein differentiation has not been considered in depth. Deciding between a "fully tryptic" search and a "semi-tryptic" search has been shown to significantly alter the set of peptides identified from a mixture [28,34]. Here, I compare the fully tryptic and semi-tryptic search in protein differentiation, and generalize the conclusion in two datasets with different sensitivity and mass accuracy.

CHAPTER 2

MATERIALS AND METHODS

2.1 Data Sources

2.1.1 ABRF Data

I used a dataset from the Association of Biomolecular Research Facilities (ABRF) iPRG 2009 study. In that study, two samples of *E.coli* lysates (labeled “red” and “yellow”) were digested with trypsin then analyzed with LC-MS/MS on an LTQ-Orbitrap with five technical replicates for each sample. The Red and Yellow replicates were derived from the same *E. coli* lysate sample running on two halves of one gel with a single region excised from each half (The “Green” and “Blue” proteomic data sets) . Proteins in the changing region of red and yellow cohorts were enriched in Blue and Green cohorts respectively (for more information see Figure S1 in Supplementary Information). A differential protein key list was built by comparing the differentially expressed proteins between the less complex Blue and Green cohorts with significance level 0.05. 85% of the proteins in the key list corresponded with the mass regions excised from the gel. The proteins significantly expressed in the Blue cohort that were also significantly expressed in the Red cohort were considered as true positives. Similarly, the proteins significantly expressed in the Green cohort that were also significantly expressed in the Yellow cohort were considered as true positives.

2.1.2 CPTAC Data

I used a dataset created by the Clinical Proteomic Technology Assessment for Cancer (CPTAC) program [11]. In the study, a yeast lysate was spiked with a mixture of 48 human proteins (Sigma-Aldrich UPS1) at several levels of concentrations. Each sample was analyzed with triplicates on seven independent instruments of four models (Thermo Fisher LTQ, LTQ-XL, LTQ-XL-Orbitrap, and LTQ-Orbitrap). Groups A, B, C, D, E were yeast spiked with UPS-1 at 0.25, 0.74, 2.2, 6.7, and 20 fmol/ul respectively. Data were processed using a FASTA database combining the yeast and human proteomes. Search parameters are provided in Supplemental File1.

2.1.3 HNSCC Data

The Head and Neck Tissue Repository [35] collected 20 head and neck squamous cell carcinomas (HNSCC) from all patients undergoing surgery in head and neck area at Vanderbilt University. These cancerous samples can be compared to 20 normal tonsillectomy tissues which were collected from pediatric tonsillectomies performed at Vanderbilt Children's Hospital. Tissues were snap-frozen in liquid nitrogen and kept at -80 °C until processing. Tumor samples were macrodissected to achieve a minimum of 70% tumor cells in the specimen to be analyzed. Epithelial cells were dissected away from lymphoid cells in normal specimen. The tissues were embedded in polyvinyl alcohol, which was then removed with wash with deionized water. Peptides were separated by isoelectric focusing and cut into 20 fractions. Each of these fractions was analyzed by liquid chromatography, followed by MS/MS analysis on a LTQ-Orbitrap.

2.1.4 ASW480 Data

Adenomatous polyposis coli (APC) is a negative regulator of Wnt signaling. Mutation of APC occurs in up to 60% of colorectal cancer (CRC) tumors. Halvey *et. al* in Vanderbilt University has examined the proteomics of two colon tumor cell lines- SW480APC (APC restored), SW480Null (mutant APC). Cells were grown in RPMI 1640 medium, supplemented with 10% fetal bovine serum, 1% penicillin/streptomycin and genetecin (1.5 mg/ml), then lysed at ambient temperature. Proteins were reduced and alkylated with 40 mM tris(2-carboxyethyl)phosphine (TCEP)/100 mM dithiothreitol (DTT) and 50 mM iodoacetamide (IAM), respectively. Samples were diluted in 50 mM AmBic, pH 8.0 and tyrpsinized overnight at 37 °C (1:50, w:w). Subsequently, peptides were lyophilized overnight. Peptides were desalted as described [36], and separated by isoelectric focusing (IEF) using immobiline IPG strips (24 cm, pH 3.5-4.5) (GE Healthcare) as described.[36,37] . LC-MS-MS shotgun proteomic analyses were performed on LTQ XL mass spectrometer (Thermo Fisher Scientific) equipped with an Eksigent NanoLC AS1 autosampler and Eksigent NanoLC 1D Plus pump, Nanospray source, and Xcalibur 2.0 SR2 instrument control. Peptides were separated on a packed capillary tip (Polymicro Technologies, 100 mm × 11 cm) with Jupiter C18 resin (5 mm, 300 Å, Phenomenex) using an in-line solid-phase extraction column (100 mm × 6 cm) packed with the same C18 resin using a frit generated with liquid silicate Kasil 1. Mobile phase A consisted of 0.1% formic acid and mobile phase B consisted of 0.1% formic acid in 90% acetonitrile. A 90-min gradient was carried out with a 30-min washing period (100% A) to allow for solid-phase extraction and removal of any residual salts. Following the washing period, the gradient was increased to 25% B by 35 min, followed by an increase to 90% B by 50 min and held for 9 min before returning 95% A. MS-MS spectra of the peptides are acquired using data-dependent scanning in which one full MS spectrum (mass range 400-2000 m/z) is followed by five MS-MS spectra. MS-MS spectra are recorded using dynamic exclusion of previously analyzed precursors for 60 s with a repeat of 1 and a repeat duration of 1. MS/MS spectra were generated by collision-induced dissociation of the peptide ions at normalized collision energy of 35% to generate a series of b- and y-ions as major fragments. Biological samples from 3 independent cell cultures were injected in duplicate for a total of 6 replicate measurements for the SW480null and SW480APC cell lines. A subset of proteins found to be differentially expressed by LC-MS/MS were validated by targeted proteomics (LC-MRM-MS). For all 22 proteins that were validated by targeted proteomics, label free shotgun proteomics data and LC-MRM data were broadly concordant, and identical trends in protein expression were observed between the two platforms[38].

2.2 Database Search Pipeline

MS/MS scans were converted to mzML file format by the msConvert tool in the ProteoWizard[39] library to provide input files for TagRecon (TR)[16], MyriMatch (MM) and X!Tandem (XT) search. These files were then converted to DTA format by ScanSifter[40,16] to enable Sequest (SQ) search. All protein databases contained sequences in both forward and reverse orientations for estimation of protein and peptide identification error rates .For LTQ data, MM, TR, and XT applied a precursor tolerance of 1.25 m/z, while SQ applied a 2.5 Da mass tolerance. For Orbitrap data, MM and XT applied a precursor tolerance of 10 or 40 ppm, while TR applied 0.01 m/z tolerance and SQ applied a 0.1 Da mass tolerance. The search results were processed by IDPicker to yield a 5% or 2% False Discovery Rate (FDR). Peptides passing these thresholds were considered as legitimate identifications. IDPicker assembled protein identifications from peptides using parsimony rules [27,28].

Statistically significant differences in protein spectral counts between different groups were calculated using quasi-likelihood Generalized Linear Modeling (GLM) by QuasiTel[35]. Proteins with p-values less than 0.05 were considered as differential proteins. Differentially expressed proteins were mapped to genes and compared for enrichment of defined classes against a reference set of all identified proteins. Search configurations, dataset information, and identified peptides are shown in Table S1, S4 and Supplementary material 2.

2.3 Model for Peptide Group-Based Spectral Count Differentiation

IDPicker generates tables reporting the number of spectral counts for each peptide group (Figure 6). I used Fisher's Exact Test instead of GLM to compute a p-value for each peptide group because the GLM includes additional covariates in the comparisons which may diminish accuracy for peptide groups with low spectral counts. I also used Fisher's Exact Test to compute a p-value for each protein group as a comparison method. I employed the Benjamini-Hochberg FDR method to correct p-values for multiple hypothesis testing [41]. Statistical techniques for the peptide group-based analysis differed from those employed in the search algorithm combination and semi-tryptic evaluations. These latter examinations employed the standard QuasiTel GLM for differentiation.

Common data analysis practices in comparative proteomics reflect the belief that FDR (multiple hypothesis testing corrected p-value) is good both as a qualitative and a quantitative indicator of the overall significance of the results. The use of FDR based meta-analysis was previously demonstrated in ChIP-chip meta-analysis [42]. The corrected p-values of peptide groups corresponding to the same protein group were combined using Stouffer's z-reverse normal transform method [43] to estimate the significance level of changes at the protein group level.

The weighted Stouffer's inverse normal transform method I built, described in equation (1) and (2), took peptide p-value, sample size and effect direction (4) into consideration to compute a protein p-value. Optimal weights for the weighted Z method were given by the square root of the spectral counts of peptide groups divided by their occurrence in protein groups (3). By this strategy, unique peptide groups are assigned higher weights than the shared peptides.

$$S_{pro} = \sum_{s=1}^{Ns} w_s d_s \, \phi^{-1}(1 - \frac{p_{pep}(s)}{2}) \tag{1}$$

$$p_{pro} = 2(1 - \phi\big(|S_{pro}|\big) \tag{2}$$

$$w_s = \sqrt{\frac{peptide\ SpC(s)/occurence(s)}{\sum_{i=1}^{Ns} peptide\ SpC(i)/occurence(i)}} \tag{3}$$

$$\begin{aligned} d_s = &+1\ for\ increased, \\ &-1\ for\ decreased, \\ &0\ for\ unchanged\ from\ one\ sample\ to\ another \end{aligned} \tag{4}$$

* ϕ and ϕ^{-1} denote the standard normal cumulative distribution function and its inverse.

2.4 Models for Combining Search Engines

I present statistical models to combine search results from four search engines. Heterogeneity among search engines results from factors including spectral pre-processing, theoretical spectrum prediction, and match scoring algorithms. As a result, FDR-based meta-analysis was necessary to summarize results. In the first model, spectral counts from each search engine were added together prior to differentiation. The combined spectral counts were analyzed by QuasiTel and corrected by the FDR method to compute p-values. In the second model, I computed FDR corrected p-values of protein spectral counts separately by search engine. These p-values were then combined for each protein using Stouffer's Z-transform probability test [44]. In the third model, I ranked the proteins by FDR corrected p-values from individual search engines (from smallest to largest). The ranks were then added together to compute a super rank for each protein. In the “Stouffer p-combo Model” and “p-Rank Sum Model,” proteins that were not identified by any included search engine were excluded in the comparison. Vote counting is well described for use in microarrays and peptide identifications [45,46]. Rhodes et al used a comparative meta-profiling which assesses the overlap of gene expression differentiation from a diverse collection of microarray datasets. Several modifications enable its use for protein spectral count differentiation. Briefly, the spectral count data were analyzed by QuasiTel, and p-values from individual search engines were FDR corrected. I then defined a significance threshold –α ($\alpha_{DEFAULT}$=0.05) and the number of top proteins I wanted to select –N_{SELECT} For these thresholds, I then ranked proteins by the number of search engines that find each significant; this positions each search engine as a “voter.” Within each class of proteins with the same vote-counts, I then ranked proteins by the minimum of their p-value from the combining search engines (minimum p-value, increasing).This process ranked potential protein differences, with the most substantial changes at the top.

Assessing FDR for vote counts and best p-values followed a permutation strategy. First, I counted the proteins for each possible number of vote counts (N_1, N_2…N_S). Permuting the p-values per search engine among proteins generated a set of randomly produced differences. I counted these differences for each possible number of vote counts (E_1, E_2…E_S). The minimum meta-false discovery rate (mFDRmin) can then be calculated by:

$$mFDRmin = minimum(\frac{[E_i + 1]}{[N_i]})\ for\ i = 0\ to\ S$$

Then I assess the validity of α with the following criteria: If mFDRmin<α, these proteins were found to be differentially expressed at the threshold α. If not, I repeated the enumeration of votes with the value of α lowered by 20% at each iteration until either a valid α is defined or the number of differential proteins detected in two or more search engines reaches 0. A valid α should not fall so far that the number of proteins with at least one vote was less than N_{SELECT}. Furthermore, to be strict in the significant level of the threshold, I should find the smallest (most significant), valid α setting by lowering α by 20% and repeat the previous validity testings iteratively. The algorithm was implemented in R (Supplementary Material 3). This model is tuned for the best performance when voter turnout is large, i.e. more search engines are deployed for each data set.

CHAPTER 3

RESULTS AND DISCUSSION

3.1 Peptide Group-Based Spectral Count Differentiation Improves Protein Differentiation

Peptide group-based spectral count differentiation better evaluates the impact of unique and shared peptide groups on protein differentiation, thus effectively reducing false positives. This method is most effective in reducing false positives when working with proteomic samples of higher organisms where a lot of shared peptide groups exist. Therefore, I tested the technique in ASW480 and HNSCC human proteomic datasets. In the ASW480 dataset, 6042 peptide groups were identified, mapping to 7325 proteins in 5215 protein groups. I compared the cell line with and without the APC vector with protein group-based and peptide group-based techniques after MyriMatch search and IDPicker filtering. Of the differentiating proteins discovered by peptide group analysis, 95% were also discovered through protein group analysis. Correspondingly, 81% of the differential proteins from protein-based differentiation were also identified by peptide group-based differentiation (Figure S2, S4). At first, this would seem to imply higher sensitivity to differences in protein group analysis, perhaps due to more aggressive p-value correction in the more numerous peptide group comparisons. Only 5 proteins were identified exclusively by peptide group-based differentiation, while 21 proteins were differentiated by protein-based but not peptide-based techniques. I examined these 21 proteins with a critical eye. In the example of protein groups for Desmin and Vimentin, four peptide groups were shared between Desmin and Vimentin and six other protein groups (Table S2). The p-values of Desmin and Vimentin from protein-based spectral count differentiation were 0.0230 and <0.0001 respectively, signifying that these two proteins were both differentially expressed. However, I found that the spectral count of the unique peptide group of Desmin had not significantly changed (p-value>0.05). Desmin and Vimentin share four peptide groups that were also shared by 2-4 other protein groups, causing cross-talk between these proteins and others that were legitimately changing. The spectral count of these peptide groups greatly impacted the total spectral count of Desmin. These data demonstrate that shared peptides can cause unchanging proteins to become false positive differences.

The p-value for Desmin was 0.1551 when differentiation was performed at the level of peptide groups with combination via Stouffer's inverse normal method[43]. Separating peptides by protein association revealed that the expression level of Desmin had not significantly changed. On the other hand, the change of Vimentin level remained significant (p-value <0.0001). In fact, the lack of change for Desmin was reinforced by microarray (p-value of 0.9875) [38]. On the other hand, the enrichment analysis of proteomic data revealed that targets of transcription repressor ZEB1 were measured at lower levels in the SW480 Null cell line, implying elevated ZEB1 activity in this cell line. Others have shown that disruption of the ZEB1/SMARCA4 binding causes an increase in CDH1 expression and a decrease in Vimentin [47]. I also compared the two methods between replicates of APC or control group which I knew should not show any differential proteins. Peptide-group based differentiation reduced the false positive differentiation by 20-41% (Figure S3). These facts have shown that peptide group-based differentiation is robust against false positives induced by shared peptides.

Peptide group based-spectral count differentiation is also more sensitive to changes in unique peptide groups. In the HNSCC dataset, 4011 proteins were assembled to 2569 protein groups, with 2941 peptide groups mapping to them. 100 differential proteins were identified by peptide group-based differentiation (Figure S2). As a test, I evaluated the biomarker set resulting from a comparison using only the peptide groups that mapped to a single protein group; limiting the information to this set of peptides, however, reduced detection of differentiating proteins by 22% (Figure S2). Of the proteins found to be differences from the peptide group-based technique employing all peptides, 94% were also found through the protein-based technique. Of the protein-based difference set, 82% were also observed through peptide-group differentiation. A majority of protein changes found by peptide group-based differentiation shared peptides with other protein groups. Myosin 14 was among the differences found by peptide group-based but not protein group-based techniques. This non-muscle myosin, which appears to play a role in cytokinesis and cell shape, was matched to five peptide groups (Table S3). Protein-based spectral count differentiation could not provide enough evidence (p-value =0.2580>0.05) to show that myosin14 was differentially expressed in cancer group versus control group. However, when I look closely into each peptide group, I find that the peptide group that contains sequences specific to this form of myosin changes significantly in spectral counts, increasing from 39 to 110 (2.82

fold, p-value=0.0004<0.05). By peptide group based spectral count differentiation, the difference is significant (p-value=0.0016<0.05). Previous studies have shown that overexpression of myosin14 inhibits cell growth [48], which coincides with the heightened expression in normal samples. Without peptide group-based comparison, this difference would be masked by other myosin forms.

Generally, protein and peptide group-based differentiation are highly concordant with each other (Figure S4). The correlation coefficient for the p-values of ASW480 proteins was 0.9470, while the HNSCC set yielded a 0.9420 correlation. After finding the differential proteins by p-values, the fold change of a protein can be estimated by averaging the fold change of its peptides. Because there are more peptide groups than protein groups for an assembly, multiple testing adjustment reduces the count of significant differences more strongly for peptide groups than for protein groups. For example, of the 20 proteins that were disagreements between the two differentiation techniques in the ASW480 dataset, three proteins (CD2 antigen cytoplasmic tail-binding protein 2, Envoplakin, Heat shock protein beta-1) are proteins with only one peptide group. As a result, the set of spectral counts compared in protein group and peptide group techniques are the same. Once multiple testing correction has been applied, though, Envoplakin shifts to a 0.0446 p-value from protein group evaluation or to an insignificant 0.0581 p-value from peptide group evaluation. Whether this constitutes the removal of a false positive difference or losing sensitivity for real differences cannot be resolved from the data on hand.

3.2 Combining Multiple Search Engines Improves Protein Differentiation

Protein differentiation is considerably affected by search algorithms. In the ABRF iPRG *E. coli* dataset, 1275 proteins in total were identified by the four search engines, while only 662 proteins were shared between all four search engines. The ability to identify truly differentiated proteins also varied among different search engines. MM, TR, XT, and SQ each identified 228, 225, 226, 207 truly differentiated proteins, respectively (Figure S5). Most truly differential proteins (derived from identifications in the "blue" and "green" samples) reach agreement between two or more search engines with consistent fold change directions. These results highlighted the necessity of combining search engines to detect more correct differences and reduce false discoveries. I applied four distinct models (see Methods) to combine different search engines. These models have shown their unique advantages to achieve better protein differentiation. I ranked the proteins by p-values from the "Count Sum Model" and "Stouffer p-combo Model" and by super rank of "p-Rank Sum Model" from smallest to largest, or by vote-counts from the "Vote Counting Model" from largest to smallest and chose the top 250 proteins (approximately the length of the key list) for true positive and false positive analyses. As shown in Figure 7, generally, combinations of search engines outperform individual search engines. For the pairing of SQ and TR, the "Stouffer p-combo Model" increased AUC by 12.7%, from 79.6% to 89.7%, and identified 18 more true positive proteins than TR by itself. Combining all four search engines by the "p-Rank Sum Model" identified 3%-13% more true positive proteins than for any individual search engine; this combination revealed that adding all possible search engines is not guaranteed to outperform a well-selected set of search engines, since the MM+TR+SQ combination was more effective. Of the search engine pairs, XT and SQ appeared least effective at complementing each other.

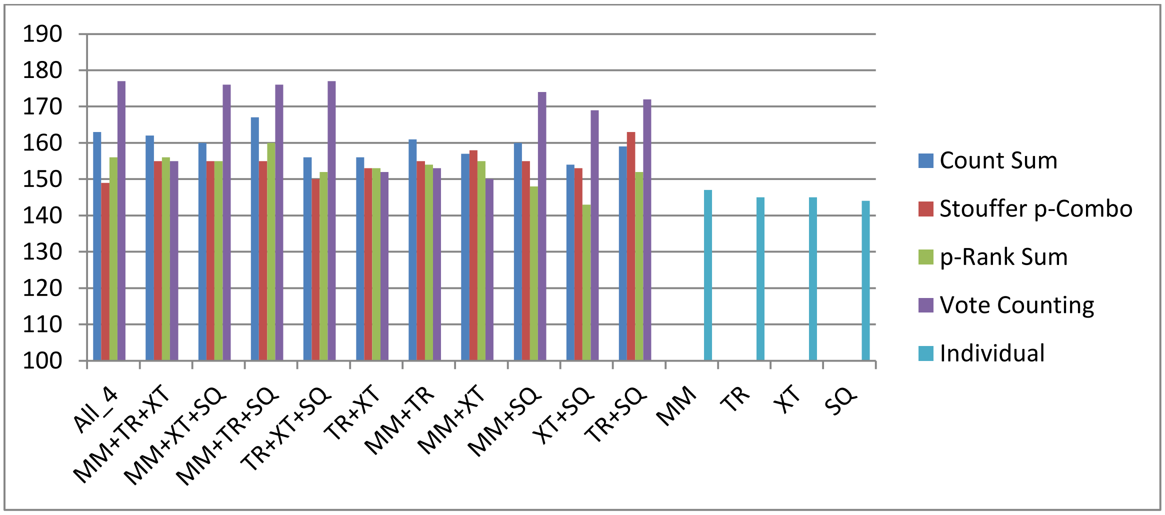


Figure 7. Number of true positive proteins out of the Top 250 of proteins of corresponding combination of search engines in ABRF *E.coli* dataset.

In "Count Sum Model", results from different search engines are combined by adding spectral counts together, in "Stouffer p-combo Model", p-values form different search engines are combined by Stouffer's method. In "p-Rank Sum Model", proteins were ranked by p-value from individual search engine (from smallest to largest). The ranks are added together to compute a super rank for each protein. In the "Vote Counting Model", proteins were ranked by the number of search engines deeming them significant along with the best individual search engine p-value.

Combining all the four search engines with the “Vote Counting Model” produced the best true positive ratio and lowest false positive ratio, with 177 true positives out of top 250 differences, while the best number of true positives of other models is only 167. The “Vote Counting Model” identified 20.5%, 22.1% 22.1% and 22.9% more true positive proteins than searching by MM, TR, XT or SQ individually. Combining three search engines such as MM+TR+SQ is also effective.

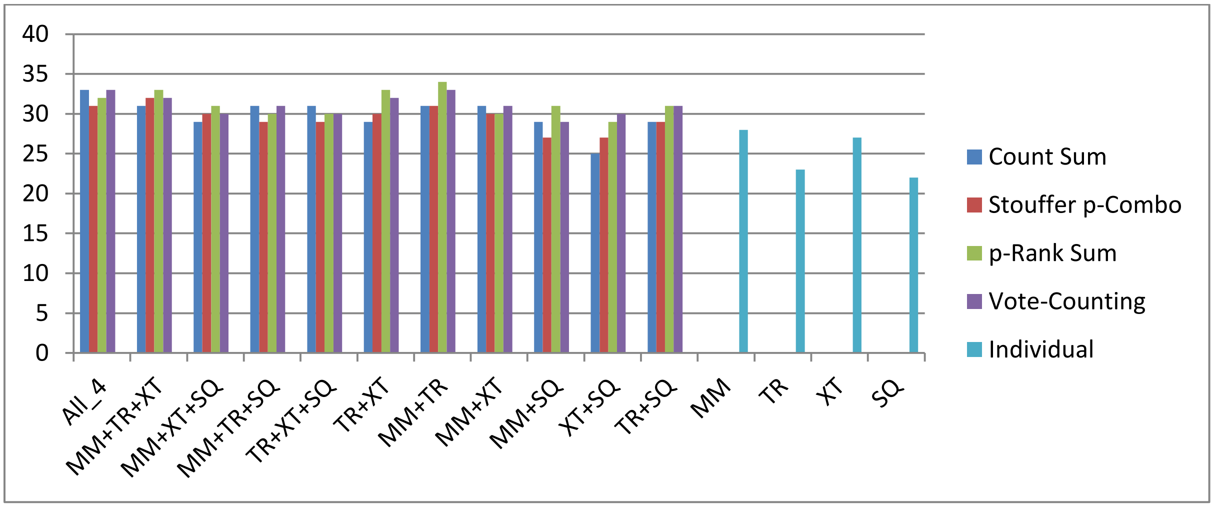


Figure 8. Number of true positive proteins in top 50 differentiated proteins using different combination of search engines in CPTAC Study6 dataset.

Combinations of search engines by these models were also evaluated in the context of the CPTAC LTQ dataset. I used data from C and E cohorts (a 9 fold difference of UPS-1 spike concentration). In total, 45 out of 48 UPS-1 proteins were identified by the four search engines. MM, TR, XT, SQ identified 42,42,41,40 UPS-1 proteins respectively

(Figure S6). I ranked the proteins by p-values, super rank or vote-counts and analyzed the top 50 proteins with the four models. Numbers of true positives among the top 50 proteins were compared in Figure 8. Again, combinations of search engines outperformed individual search engines for revealing protein differences. Combining TR and SQ with the “Stouffer p-combo Model” generated 32% more true positives than SQ individually. Combining all four search engines by the “Count Sum Model” identified 17.9%-50.0% more true positive proteins than individual search engines.

Again, combining all the four search engines with the “Vote Counting Model” produced one of the best true positive ratio and lowest false positive ratio, with 33 true positives out of top 50 differences. The “Vote Counting Model” identified 17.9%, 43.5% 22.2% and 50.0% more true positive proteins than searching by MM, TR, XT or SQ individually. The advantage is not distinctive here because of the small number of proteins in the “answer key.” Combining only two search engines was helpful for one data set but not the other; the voting model benefits from a larger pool of votes (Figure 7, 8). For example, MM+XT, TR+XT, MM+TR only identified around 150 true positive proteins. In the ABRF data set, only the combinations that included Sequest gave the highest performance, though this algorithm working alone had yielded the lowest number of true differences.

The four models for combining search engines have different strengths and weaknesses. In simply adding spectral counts for a protein identified by multiple search engines, a single spectrum might be counted multiple times. Although the multiple counting increases the confidence of identification and spectral count differentiation, it will get extreme p-values because of the correlation between search results. In the “Stouffer p-combo Model”, combining p-values among algorithms increases the sensitivity of the collective analysis, but has risks of bias towards idiosyncratically significant p-values of one search engine. In the “p-Rank Sum Model”, the super rank comprises a non-parametric assessment of the results from individual engines. Drawing conclusions about which of these techniques is best would over-generalize from the two sample sets evaluated in this study, though combination is clearly beneficial. The “Vote Counting Model” was most powerful when combining more search engines. Overall, combining search engines improves protein differentiation by not only increasing the protein inventories, but also increasing the pool of information available to differentiate each protein. Each combination of search engines allows for better discrimination than any individual search engine.

3.3 Semi-Tryptic Search Outperforms Tryptic Search in Protein Differentiation

A given search engine may yield different performance depending on its configuration. Bioinformaticists have argued for years that semi-tryptic searching, which allows the identification of peptides that differ from canonical trypsin specificity on one terminus, improves the inventories possible from proteomics[49]. I tested this parameter for its impact on comparative proteomics. Table S4 reports the number of identified peptides by fully-tryptic and semi-tryptic searches.

I first compared Red/Yellow cohorts in the iPRG *E.coli* dataset. All the other configurations and analysis were identical. Figure 9 shows the ROC curve of differentiated protein expression with semi-tryptic or fully tryptic searches. Semi-tryptic search achieved better sensitivity and specificity than fully tryptic search, with AUC increased by 6% (from 83.77% to 88.56%). Similarly, when comparing true positive and false positive proteins at the cut point of p-value 0.05, semi-tryptic search greatly increases true positive proteins by 7.07% for the same number of false positives. The improvement reveals that semi-tryptic search achieves better sensitivity and specificity than fully tryptic search for a sample in which many proteins offer stark differences between cohorts.

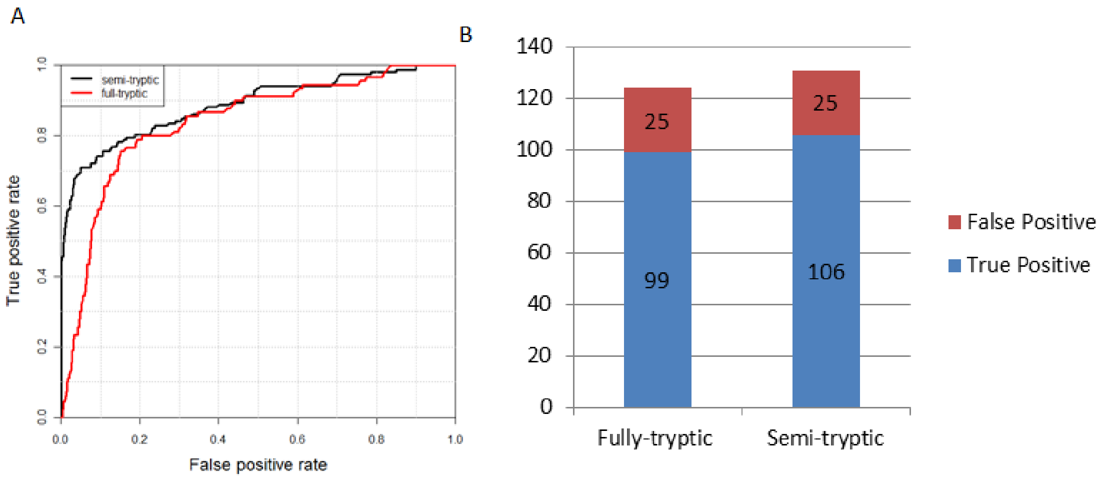


Figure 9. Comparison of fully tryptic and semi-tryptic searching in ABRF dataset.

{A} ROC curve of differentiated proteins expression using semi-tryptic and fully tryptic search in the iPRG dataset. {B} True positives and false positives at p-value of 0.05 discovered through semi-tryptic and fully tryptic search.

I next analyzed the CPTAC dataset (where the spiked proteins differed by a factor of three between each pair of five levels) with fully tryptic and semi tryptic search. I compared the spectral counts of proteins in these cohorts in pairs (Table 1). I chose a sampling of the possible fold changes, preferring samples where spike concentrations were greater. Semi- tryptic search generally outperformed fully tryptic search in AUC. Especially in D and E cohorts, where UPS1 proteins were most dominant, semi-tryptic search increased AUC by 5.5% (from 86.76% to 91.50%).

Table 1. Fully tryptic versus semi tryptic search in Yeast Sample with Spiked Human Proteins

	Fully tryptic			Semi tryptic		
	True positive/false positive	Average Spectral Count Ratio	Area Under Curve	True positive/false positive	Average Spectral Count Ratio	Area Under Curve
	27-fold difference					
B versus E	30/20	10.74	0.9786	32/18	12.36	0.9827
A versus D	25/25	9.67	0.9939	27/23	12.62	0.9944
	9-fold difference					
C versus E	29/21	4.84	0.9777	31/19	4.33	0.9779
B versus D	22/28	2.45	0.9799	25/25	6.76	0.9765

	3-fold difference					
D versus E	21/29	1.95	0.8676	24/26	2.12	0.9150
C versus D	21/29	2.53	0.8290	24/26	2.48	0.8841
B versus C	9/41	3.25	0.8859	14/36	3.21	0.8942

*The amount of UPS-1 (Sigma-Aldrich) proteins that spiked in A, B, C, D, E are 0.24, 0.67, 2.7, 6.7, 20 (fmol/µg yeast) respectively.

*Geometric average is calculated by geometric mean of the ratio : $(spectral\ count\ in\ group\ 1+1)/(spectral\ count\ in\ group\ 2+1)$, group 1 and group 2 indicates the comparison pairs. For spike in protein that is not found in the search results, their spectral counts are set as zero.

The top 50 (approximately the number of proteins in the gold standard) most differentiated proteins for each pairwise comparison were evaluated against the list of proteins known to change, and the numbers of true positives and false positives were computed (TP/FP). At different spike levels, semi-tryptic search detects more true positive proteins along with fewer or unchanged false positive proteins. Especially in B vs C, which contained only small amounts of spiked proteins with a three-fold concentration difference, semi tryptic search identified 55% more true positives than fully tryptic search. Generally, semi tryptic search provides better sensitivity and specificity than fully tryptic search, especially when comparing groups with small spike-in protein concentration changes (D vs E, C vs D).

Why would adding semi-tryptic peptide improve protein differentiation? When an algorithm fails to identify a spectrum, a semi-tryptic search will typically assign a semi-tryptic peptide to the spectrum (because random semi-tryptic peptides outnumber fully-tryptic peptides by more than an order of magnitude). Software that separates correctly identified spectra from incorrectly identified ones exploits this information to identify a larger set of peptides, even if no semi-tryptic peptides are present. The most abundant proteins in a mixture are, in turn, more likely to produce semi-tryptic peptides in addition to fully-tryptic peptides. As the concentration of UPS-1 proteins increases from group A to group E, the percentage of semi-tryptic peptides from these UPS-1 proteins was 0% in group A and group B. The percentage increased to 6.9% -7.0% in group C and group D, and reached the highest-10.6% in group E. The increased identification of semi-tryptics from dominant proteins increases the power of semi-tryptic search in protein differentiation and expands the dynamic range of differentiation.

CHAPTER 4

CONCLUSIONS

Spectral count differentiation benefits from a peptide group-based evaluation strategy, new models for combining database search engines, and care in search configuration. Peptide group-based spectral count differentiation helps to resolve the protein inference problem, giving particular power when untangling complex protein-peptide clusters. It can be used as an alternative or complementary differentiation method when working with complex comparative proteomic samples where a lot of shared peptide groups exist. In systems where proteins of multiple species are present, such as xenograft models of cancer or other samples that contain proteins from multiple eukaryotes, the method has great potential in improving protein differentiation. Due to the influence of multiple testing adjustment, this method may lose power for proteins near the p-value threshold.

Three of the four tested models for combining search engines for differentiation proved to be effective. The "Count Sum Model" can be easily implemented for almost any workflow and delivers solid performance, though false positives may prove problematic. The "p-Rank Sum Model" may be more robust against idiosyncratic performance for individual search engines. These two models can be used when combining two or three search engines; in this examination, MM+TR+SQ yielded the best performance. With the increased ability of incorporating three or more search engines, the "Vote Counting

Model" is very robust against idiosyncratic results for individual search engines. Its steady, high performance in these datasets suggested great potential for fielding many search engines at once.

These models may be most useful in biomarker discovery, where some proteins of interest are at low abundance. The use of multiple engines can broaden the pool of information available to differentiate proteins present at small quantities. These models can also apply to samples with large genomes when low-resolution mass analyzers have measured precursor masses; these searches compare very large numbers of candidate sequences to every spectrum, thus losing discrimination. In the future, these models may be developed by recognizing the unique contribution of each search engine. The search engines that provide more confident IDs with better sensitivity and specificity, such as MM in the two datasets above, should be afforded more importance. In the "Count Sum Model," excluding the overlapping peptide spectrum matching by different search engines can also be used to reduce the type I error.

In both datasets, semi-tryptic peptide search outperforms fully tryptic peptide search in protein differentiation studies in multiple aspects including higher discovery rate, better specificity, and better sensitivity. Semi-tryptic search is more sensitive to small protein concentration changes. Ignoring the contributions of semi-tryptic peptides would sacrifice discrimination for levels of abundant proteins. If endogenous proteases are present in a sample, semi-tryptic search is obviously the choice for better protein differentiation, but the improved inventories are feasible through this option even in samples dominated by fully-tryptic peptides. In the future, more general conclusions can be drawn by in-depth analysis of trypsin specificity configurations by search engines other than MM.

In conclusion, these three strategies yield higher quality differentiation based on spectral counting. These strategies are each generic enough to enable their incorporation in many bioinformatics pipelines. Since the spectral counting strategy was introduced in 2004, it has become a standby for many laboratories. These advances will enable its application to samples where proteins share peptides in complex relationships, discrimination of correct peptides requires multiple pipelines, and a wide dynamic range of proteins is interrogated.

REFERENCES

1. Aebersold R, Mann M (2003) Mass spectrometry-based proteomics. Nature 422 (6928):198-207

2. Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat Biotechnol 17 (10):994-999

3. Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, Mann M (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Mol Cell Proteomics 1 (5):376-386

4. Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, Hattan S, Khainovski N, Pillai S, Dey S, Daniels S, Purkayastha S, Juhasz P, Martin S, Bartlet-Jones M, He F, Jacobson A, Pappin DJ (2004) Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. Mol Cell Proteomics 3 (12):1154-1169

5. Bantscheff M, Schirle M, Sweetman G, Rick J, Kuster B (2007) Quantitative mass spectrometry in proteomics: a critical review. Anal Bioanal Chem 389 (4):1017-1031

6. Liu H, Sadygov RG, Yates JR, 3rd (2004) A model for random sampling and estimation of relative protein abundance in shotgun proteomics. Anal Chem 76 (14):4193-4201

7. Zybailov B, Coleman MK, Florens L, Washburn MP (2005) Correlation of relative abundance ratios derived from peptide ion chromatograms and spectrum counting for quantitative proteomic analysis using stable isotope labeling. Anal Chem 77 (19):6218-6224

8. Fu X, Gharib SA, Green PS, Aitken ML, Frazer DA, Park DR, Vaisar T, Heinecke JW (2008) Spectral index for assessment of differential protein expression in shotgun proteomics. J Proteome Res 7 (3):845-854

9. Old WM, Meyer-Arendt K, Aveline-Wolf L, Pierce KG, Mendoza A, Sevinsky JR, Resing KA, Ahn NG (2005) Comparison of label-free methods for quantifying human proteins by shotgun proteomics. Mol Cell Proteomics 4 (10):1487-1502

10. Keshishian H, Addona T, Burgess M, Kuhn E, Carr SA (2007) Quantitative, multiplexed assays for low abundance proteins in plasma by targeted mass spectrometry and stable isotope dilution. Mol Cell Proteomics 6 (12):2212-2229

11. Tabb DL, Vega-Montoto L, Rudnick PA, Variyath AM, Ham AJ, Bunk DM, Kilpatrick LE, Billheimer DD, Blackman RK, Cardasis HL, Carr SA, Clauser KR, Jaffe JD, Kowalski KA, Neubert TA, Regnier FE, Schilling B, Tegeler TJ, Wang M, Wang P, Whiteaker JR, Zimmerman LJ, Fisher SJ, Gibson BW, Kinsinger CR, Mesri M, Rodriguez H, Stein SE, Tempst P, Paulovich AG, Liebler DC, Spiegelman C (2010) Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. J Proteome Res 9 (2):761-776

12. Eng JKM A, Yates JR. (1994) An approach to correlate tandem mass-spectral data of peptides with amino acid sequences in a protein database. J Am Soc Mass Spectrom (5):976–989.

13. Craig R, Beavis RC (2004) TANDEM: matching proteins with tandem mass spectra. Bioinformatics 20 (9):1466-1467

14. Tabb DL, Fernando CG, Chambers MC (2007) MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. J Proteome Res 6 (2):654-661

15. Dasari S, Chambers MC, Codreanu SG, Liebler DC, Collins BC, Pennington SR, Gallagher WM, Tabb DL (2011) Sequence tagging reveals unexpected modifications in toxicoproteomics. Chem Res Toxicol 24 (2):204-216

16. Dasari S, Chambers MC, Slebos RJ, Zimmerman LJ, Ham AJ, Tabb DL (2010) TagRecon: high-throughput mutation identification through sequence tagging. J Proteome Res 9 (4):1716-1726

17. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20 (18):3551-3567

18. Ning K, Fermin D, Nesvizhskii AI (2010) Computational analysis of unassigned high-quality MS/MS spectra in proteomic data sets. Proteomics 10 (14):2712-2718

19. Tanner S, Shu H, Frank A, Wang LC, Zandi E, Mumby M, Pevzner PA, Bafna V (2005) InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. Anal Chem 77 (14):4626-4639

20. Shevchenko A, Sunyaev S, Loboda A, Bork P, Ens W, Standing KG (2001) Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching. Anal Chem 73 (9):1917-1926

21. Figeys D, Boutilier K, Ross M, Podtelejnikov AV, Orsi C, Taylor R, Taylor P (2005) Comparison of different search engines using validated MS/MS test datasets. Anal Chim Acta 534 (1):11-20

22. Alves G, Wu WW, Wang G, Shen RF, Yu YK (2008) Enhancing peptide identification confidence by combining search methods. J Proteome Res 7 (8):3102-3113

23. Searle BC, Turner M, Nesvizhskii AI (2008) Improving sensitivity by probabilistically combining results from multiple MS/MS search methodologies. J Proteome Res 7 (1):245-253

24. Kwon T, Choi H, Vogel C, Nesvizhskii AI, Marcotte EM (2011) MSblender: A probabilistic approach for integrating peptide identifications from multiple database search engines. J Proteome Res 10 (7):2949-2958

25. Jones AR, Siepen JA, Hubbard SJ, Paton NW (2009) Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines. Proteomics 9 (5):1220-1229

26. Nesvizhskii AI, Aebersold R (2005) Interpretation of shotgun proteomic data: the protein inference problem. Mol Cell Proteomics 4 (10):1419-1440

27. Zhang B, Chambers MC, Tabb DL (2007) Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. J Proteome Res 6 (9):3549-3557

28. Ma ZQ, Dasari S, Chambers MC, Litton MD, Sobecki SM, Zimmerman LJ, Halvey PJ, Schilling B, Drake PM, Gibson BW, Tabb DL (2009) IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. J Proteome Res 8 (8):3872-3881

29. Jin S, Daly DS, Springer DL, Miller JH (2008) The effects of shared peptides on protein quantitation in label-free proteomics by LC/MS/MS. J Proteome Res 7 (1):164-169

30. Fermin D, Basrur V, Yocum AK, Nesvizhskii AI (2011) Abacus: a computational tool for extracting and pre-processing spectral count data for label-free quantitative proteomic analysis. Proteomics 11 (7):1340-1345

31. Nesvizhskii AI, Keller A, Kolker E, Aebersold R (2003) A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem 75 (17):4646-4658

32. States DJ, Omenn GS, Blackwell TW, Fermin D, Eng J, Speicher DW, Hanash SM (2006) Challenges in deriving high-confidence protein identifications from data gathered by a HUPO plasma proteome collaborative study. Nat Biotechnol 24 (3):333-338

33. Olsen JV, Ong SE, Mann M (2004) Trypsin cleaves exclusively C-terminal to arginine and lysine residues. Mol Cell Proteomics 3 (6):608-614

34. Picotti P, Aebersold R, Domon B (2007) The implications of proteolytic background for shotgun proteomics. Mol Cell Proteomics 6 (9):1589-1598

35. Li M, Gray W, Zhang H, Chung CH, Billheimer D, Yarbrough WG, Liebler DC, Shyr Y, Slebos RJ (2010) Comparative shotgun proteomics using spectral count data and quasi-likelihood modeling. J Proteome Res 9 (8):4295-4305

36. Sprung RW, Jr., Brock JW, Tanksley JP, Li M, Washington MK, Slebos RJ, Liebler DC (2009) Equivalence of protein inventories obtained from formalin-fixed paraffin-embedded and frozen

tissue in multidimensional liquid chromatography-tandem mass spectrometry shotgun proteomic analysis. Mol Cell Proteomics 8 (8):1988-1998

37. Slebos RJ, Brock JW, Winters NF, Stuart SR, Martinez MA, Li M, Chambers MC, Zimmerman LJ, Ham AJ, Tabb DL, Liebler DC (2008) Evaluation of strong cation exchange versus isoelectric focusing of peptides for multidimensional liquid chromatography-tandem mass spectrometry. J Proteome Res 7 (12):5286-5294

38. Halvey PJ, Zhang B, Coffey RJ, Liebler DC, Slebos RJ (2012) Proteomic consequences of a single gene mutation in a colorectal cancer model. J Proteome Res 11 (2):1184-1195

39. Kessner D, Chambers M, Burke R, Agus D, Mallick P (2008) ProteoWizard: open source software for rapid proteomics tools development. Bioinformatics 24 (21):2534-2536

40. Ma ZQ, Tabb DL, Burden J, Chambers MC, Cox MB, Cantrell MJ, Ham AJ, Litton MD, Oreto MR, Schultz WC, Sobecki SM, Tsui TY, Wernke GR, Liebler DC (2011) Supporting tool suite for production proteomics. Bioinformatics 27 (22):3214-3215

41. Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I (2001) Controlling the false discovery rate in behavior genetics research. Behav Brain Res 125 (1-2):279-284

42. Pyne S, Futcher B, Skiena S (2006) Meta-analysis based on control of false discovery rate: combining yeast ChIP-chip datasets. Bioinformatics 22 (20):2516-2522

43. Whitlock MC (2005) Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. J Evol Biol 18 (5):1368-1373

44. Stouffer SA, Suchman, E.A., DeVinney, L.C., Star, S.A. & Williams, R.M. Jr. (1949) The American soldier: Adjustment during army life. Princeton University Press Princeton

45. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM (2004) Large-scale meta-analysis of cancer microarray data identifies common

transcriptional profiles of neoplastic transformation and progression. Proc Natl Acad Sci U S A 101 (25):9309-9314

46. Edwards N, Wu X, Tseng C-W (2009) An Unsupervised, Model-Free, Machine-Learning Combiner for Peptide Identifications from Tandem Mass Spectra. Clinical Proteomics 5 (1):23-36

47. Sanchez-Tillo E, Lazaro A, Torrent R, Cuatrecasas M, Vaquero EC, Castells A, Engel P, Postigo A (2010) ZEB1 represses E-cadherin and induces an EMT by recruiting the SWI/SNF chromatin-remodeling protein BRG1. Oncogene 29 (24):3490-3500

48. Wan D, Gong Y, Qin W, Zhang P, Li J, Wei L, Zhou X, Li H, Qiu X, Zhong F, He L, Yu J, Yao G, Jiang H, Qian L, Yu Y, Shu H, Chen X, Xu H, Guo M, Pan Z, Chen Y, Ge C, Yang S, Gu J (2004) Large-scale cDNA transfection screening for genes related to cancer development and progression. Proc Natl Acad Sci U S A 101 (44):15724-15729

49. Keller A, Nesvizhskii AI, Kolker E, Aebersold R (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem 74 (20):5383-5392

APPENDIX A

Supplementary Materials 1

Table S1. Data sets, Search Engines, Protein Sequence Databases used in this study.

Dataset	replicates	Instrument	average No. of MS2 scans	Sequence databases	MyriMatch	Tagrecon	X!Tandem	Sequest
					Precursor mz tolerance/fragment mz tolerance	Precursor mz tolerance/fragment mz tolerance	Parent monoisotopic mass error /fragment monoisotopic mass error	Peptide mass tolerance/fragment ion tolerance
ASW480	6	LTQ	12124	Uniprot-Human-20110701	1.25/0.5	\	\	\
HNSCC	1	Orbitrap	28230	Uniprot-Human-20110701	0.1/0.5	\	\	\
ABRF-Ecoli	5	Orbitrap	17496	UniProt-ECOLI-20110208	10ppm /0.5	0.01/0.5	40 ppm/0.5 daltons	0.1/0.0
CPTAC-Yeast	3	LTQ	261485	Uniprot-Yeast and Human	1.25/0.5	1.25/0.5	+3.0 -0.5 daltons/0.5 Daltons	2.5/0.0

Table S2. Peptide-to-protein Table of Desmin and Vimentin in ASW480 Dataset

Peptide Group	1	2	3	4	5	6	Total Spectral Count of Proteins	p-value	
								Protein group based p-value	Peptide group based p-value
Occurence	Only in Desm	Shared by DESM, VIM and 2 other protein groups	Shared by DESM, VIM and 2 other protein groups	Shared by DESM, VIM and 4 other protein groups	Shared by DESM, VIM	Only in VIM			
Spectra Count APC/Null	7/8	14/19	10/7	23/40	19/42	338/613			
P-value of individual peptide group	1	1	1	0.4699	0.0879	4.30E-21			
Desmin(DESM)	X	X	X	X	X		73/116	0.0232	0.1551
Vimentin(VIM)		X	X	X	X	X	404/721	<0.0001	<0.0001

Peptides corresponding to each peptide group are shown in Supplemental Material 4.

Table S3. Peptide-to-protein Table of Myosin 14 in HNSCC Dataset

Peptide Group	1	2	3	4	5	Total Spectral Count of Protein	p-value	
							Protein group based p-value	Peptide group based p-value

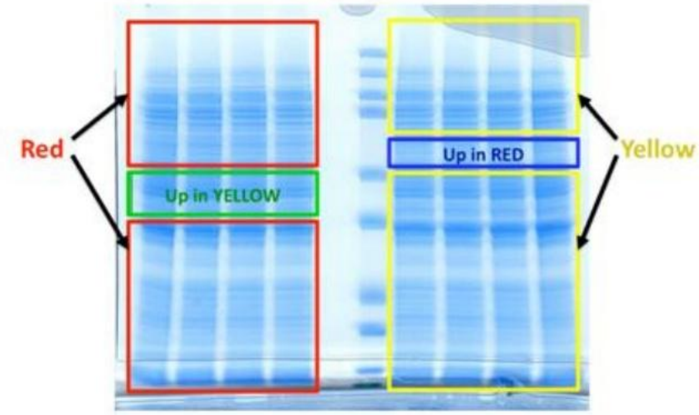
Occurence	Unique peptide group of myosin 14	Shared by myosin 14, and 3 other protein groups	Shared by myosin 14, and 2 other protein groups	Shared by myosin 14 and 2 other protein groups	Shared by myosin 14 and myosin 9			
Spectra Count (Cancer/Ctrl)	39/110	23/19	14/9	4/3	4/4	84/145		
P-value of individual peptide group	0.0004	0.8260	0.7185	1	1			
Myosin 14	X	X	X	X	X		0.2584	0.0016

Peptides corresponding to each peptide group are shown in Supplemental Material 4.

Table S4. Average number of fully and semi-tryptic peptides confidently identified (rank1) by both searches

	ABRF Dataset		CPTAC Dataset	
	Fully-tryptic search	Semi-tryptic search	Fully-tryptic search	Semi-tryptic search
Number of Peptides	16581	16988	41209	49199

Figure S1. The “Red/Yellow” iPRG 2009 LC-MS/MS data set with “Blue/Green” LC-MS/MS answer keys.



source: http://www.abrf.org/ResearchGroups/ProteomicsInformaticsResearchGroup/Studies/iPRG2009_presentation.pdf

Figure S2. Venn diagram of differentially expressed proteins at protein level and peptide group level in two datasets

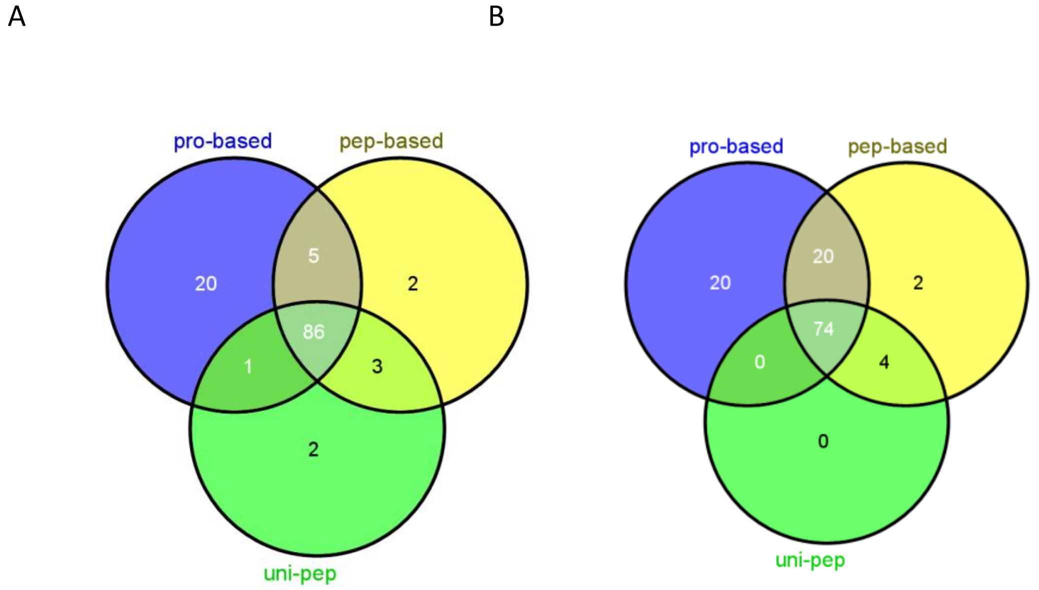


Figure S2 Venn diagram of differentially expressed proteins at protein level, peptide group level and with only unique peptides in {A} ASW480, {B} HNSCC dataset. With our method of protein differentiation at peptide group level, I identified 96 differential proteins in ASW480 dataset and 100 proteins in HNSCC dataset with 94-95% overlapping with protein based differentiation. When using peptide-based differentiation with only unique peptides, I will lose 7-22% differential proteins with few gains.

Figure S3. Venn diagram of differentially expressed protein at protein and peptide group level in ASW480 replicates of the same groups

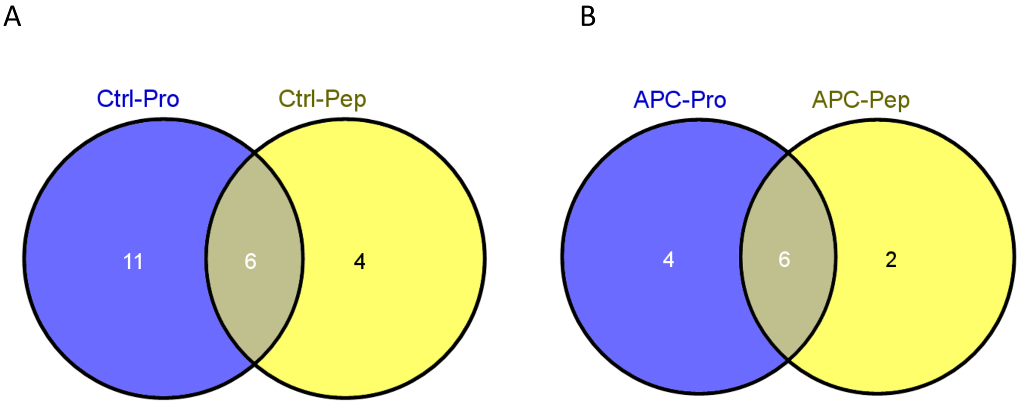


Figure S3 Venn diagram of differentially expressed protein at protein and peptide group level in ASW480 dataset comparing replicates of {A} control groups (Ctrl-Pro, Ctrl-Pep) or {B} APC groups (APC-Pro, APC-Pep) respectively. There should not be differential proteins between replicates of either control or APC group, thus these identified proteins are false discoveries. I can see that peptide group based differentiation effectively reduces false positives.

Figure S4. Correlation coefficient between protein and peptide group-based p-values

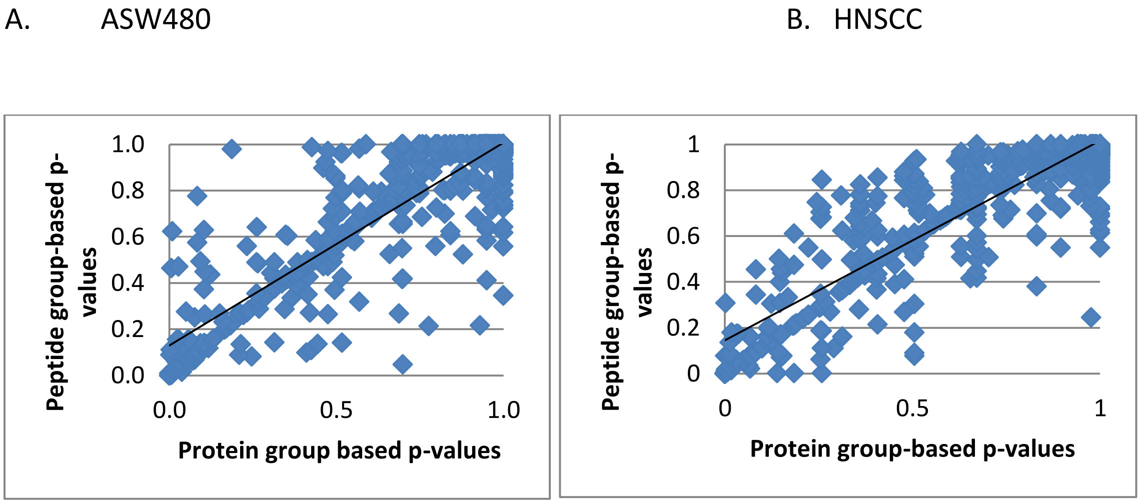


Figure S4. The Correlation coefficient between protein and peptide group-based p-values are {A} 0.9472 (ASW480 Dataset) and   {B} 0.9422 (HNSCC Dataset);

Figure S5. Venn diagram comparing the differential proteins identified by different search engines in ABRF dataset

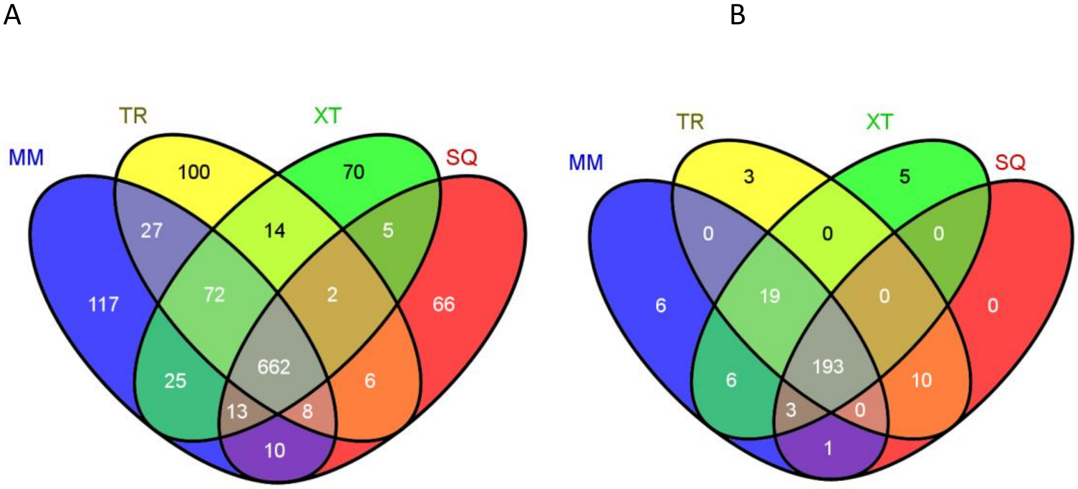


Figure S5: Venn diagram comparing the differential proteins identified by different search engines –Myrimatch (MM), X!tandem(XT) Tagrecon (TR) and Sequest (SQ) in ABRF dataset. {A}Venn diagram of proteins that identified by four different search engines. The number of proteins identified by MM, TR, XT, SQ are 934, 891, 863, 772 respectively.   {B} Venn diagram of truly differential proteins identified by four different search engines. Of the 193 truly differential proteins identified by all four search engines, only 3 proteins showed inconsistent fold change directions among search engines. Average pairwise correlation of fold changes between search engines is 0.9109.

Figure S6. Venn diagram comparing the differential proteins identified by different search engines in CPTAC dataset

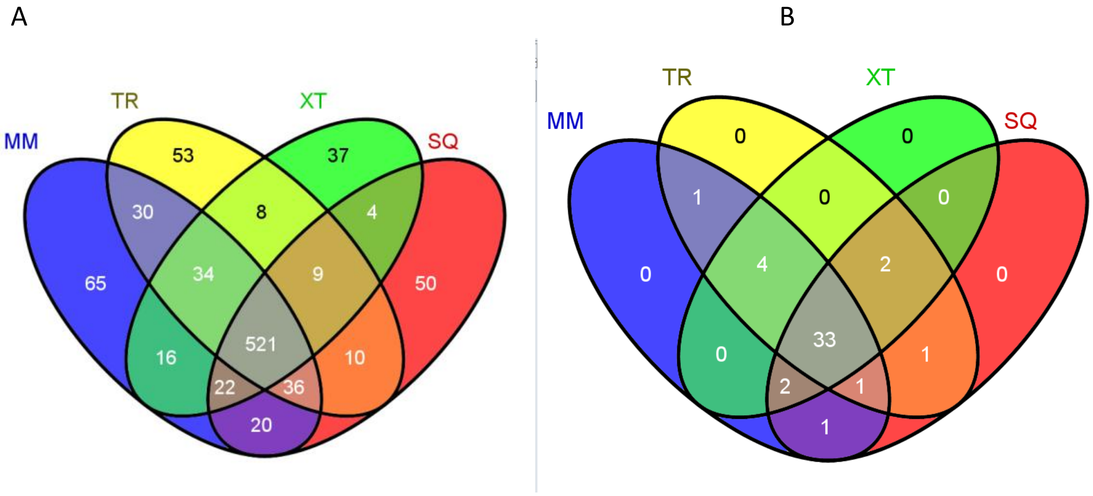


Figure S6: Venn diagram comparing the differential proteins identified by different search engines –Myrimatch (MM), X!tandem(XT) Tagrecon (TR) and Sequest (SQ) in CPTAC study 6 dataset. {A} Venn diagram of proteins that identified by four different search engines. The number of proteins identified by MM, TR, XT, SQ are 744, 701, 651, 672 respectively. {B} Venn diagram of UPS-1proteins identified by four different search engines. MM, TR,XT,SQ have identified 42,42,41,40 UPS-1 proteins respectively. All the 33 truly differential proteins identified by all four search engines showed consistent fold change directions among search engines.

APPENDIX B

Supplementary Materials 2

MyriMatch Configurations

ABRF and HNSCC Dataset Configurations

```
PrecursorMzTolerance= 10 ppm
FragmentMzTolerance = 0.5
FragmentMzToleranceUnits = daltons

AdjustPrecursorMass    = true
MinPrecursorAdjustment = -1.008665
MaxPrecursorAdjustment = 1.008665
PrecursorAdjustmentStep = 1.008665
NumSearchBestAdjustments = 3

DuplicateSpectra = true
UseChargeStateFromMS = true
NumChargeStates = 4
UseSmartPlusThreeModel = true
TicCutoffPercentage    = 0.95

CleavageRules =  "trypsin"
NumMaxMissedCleavages =  2
NumMinTerminiCleavages =  1 (for semi tryptic search  or 2 for fully tryptic search)
UseAvgMassOfSequences = false
MinCandidateLength =  5

DynamicMods = "M ^ 15.9949 (Q * -17.026"
MaxDynamicMods = 2
StaticMods = "C 57.0215"
```

```
ComputeXCorr = true
```

CPTAC Dataset Configuration

```
PrecursorMzTolerance= 1.25
PrecursorMzToleranceUnits = daltons
FragmentMzTolerance = 0.5
FragmentMzToleranceUnits = daltons

AdjustPrecursorMass = false
MinPrecursorAdjustment = -1.008665
MaxPrecursorAdjustment = 1.008665
PrecursorAdjustmentStep = 1.008665
NumSearchBestAdjustments = 3

DuplicateSpectra = true
UseChargeStateFromMS = false
NumChargeStates = 3
UseSmartPlusThreeModel = true

CleavageRules =  "trypsin"
NumMaxMissedCleavages =  2
NumMinTerminiCleavages =  1
UseAvgMassOfSequences = true
MinCandidateLength =  5

DynamicMods = "M ^ 15.9949 (Q * -17.026 C @ 57.021"
MaxDynamicMods = 3
StaticMods = ""

ComputeXCorr = true

DecoyPrefix = "rev_"

MaxResults = 5
```

ASW480 Dataset Configuration

```
PrecursorMzTolerance= 1.25
FragmentMzTolerance = 0.5

DuplicateSpectra = true
UseChargeStateFromMS = false
NumChargeStates = 3
```

```
UseSmartPlusThreeModel = true
TicCutoffPercentage = 0.98

CleavageRules =  "trypsin"
NumMaxMissedCleavages =  2
NumMinTerminiCleavages =  2
UseAvgMassOfSequences = true
MinCandidateLength =  5

DynamicMods = "M ^ 15.9949 (Q @ -17.026 ( $ 42.015"
MaxDynamicMods = 3
StaticMods = "C 57.0215"

MaxResults = 5
ComputeXCorr = true
```

Sequest Configurations

ABRF Dataset Configuration

```
database_name =/hactar/home/yaoyi/fasta/20110208-UniProt-ECOLI-Cntms-reverse.fasta
first_database_name =/hactar/home/yaoyi/fasta/20110208-UniProt-ECOLI-Cntms-reverse.fasta
second_database_name =
peptide_mass_tolerance = 0.1
create_output_files = 1             ; 0=no, 1=yes
ion_series = 0 1 1 0.0 1.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0
fragment_ion_tolerance = 0.0        ; leave at 0.0 unless you have real poor data
num_output_lines = 5                ; # peptide results to show
num_description_lines = 5           ; # full protein descriptions to show for top N peptides
num_results = 500                   ; # of results to process
show_fragment_ions = 0              ; 0=no, 1=yes
print_duplicate_references = 1      ; 0=no, 1=yes
enzyme_number = 0     # 0.  No_Enzyme     1.  Trypsin_Strict   KR          2.  Trypsin KRLNH
diff_search_options = 15.994915 M 57.021464 C
term_diff_search_options = 0.000 0.000; c term, n term diff mods
max_num_differential_AA_per_mod = 3   ; max # of modified AA per diff. mod in a peptide
```

```
nucleotide_reading_frame = 0          ; 0=proteinDB, 1-6, 7=forward three, 8=reverse
three, 9=all six
mass_type_parent = 1                 ; 0=average masses, 1=monoisotopic masses
mass_type_fragment = 1               ; 0=average masses, 1=monoisotopic masses
remove_precursor_peak = 0            ; 0=no, 1=yes
ion_cutoff_percentage = 0.0          ; prelim. score cutoff % as a decimal number i.e. 0.30
for 30%
protein_mass_filter = 0 0            ; enter protein mass min & max value ( 0 for both =
unused)
max_num_internal_cleavage_sites = 2   ; maximum value is 5; for enzyme search
match_peak_count = 0                 ; number of auto-detected peaks to try matching (max
5)
match_peak_allowed_error = 1          ; number of allowed errors in matching auto-
detected peaks
match_peak_tolerance = 1.0           ; mass tolerance for matching auto-detected peaks
partial_sequence =
```

CPTAC Dataset Configuration

```
database_name =/hactar/fasta/20080131-SGD-BSA-Cntm-Human-reverse.fasta
first_database_name =/hactar/fasta/20080131-SGD-BSA-Cntm-Human-reverse.fasta
second_database_name =
peptide_mass_tolerance = 2.5
create_output_files = 1              ; 0=no, 1=yes
ion_series = 0 1 1 0.0 1.0 0.0 0.0 0.0 0.0 0.0 1.0 0.0
fragment_ion_tolerance = 0.0          ; leave at 0.0 unless you have real poor data
num_output_lines = 5                 ; # peptide results to show
num_description_lines = 5            ; # full protein descriptions to show for top N peptides
num_results = 500                      ; # of results to process
show_fragment_ions = 0               ; 0=no, 1=yes
print_duplicate_references = 1        ; 0=no, 1=yes
enzyme_number = 1                    ; # 0.  No_Enzyme       1.  Trypsin_Strict   KR          2.
Trypsin   KRLNH
diff_search_options = 15.9949 M
term_diff_search_options = 0.000 0.000; c term, n term diff mods
max_num_differential_AA_per_mod = 3    ; max # of modified AA per diff. mod in a
peptide
nucleotide_reading_frame = 0          ; 0=proteinDB, 1-6, 7=forward three, 8=reverse
three, 9=all six
mass_type_parent = 0                 ; 0=average masses, 1=monoisotopic masses
mass_type_fragment = 1               ; 0=average masses, 1=monoisotopic masses
remove_precursor_peak = 0            ; 0=no, 1=yes
ion_cutoff_percentage = 0.0          ; prelim. score cutoff % as a decimal number i.e. 0.30
for 30%
```

```
protein_mass_filter = 0 0             ; enter protein mass min & max value ( 0 for both = unused)
max_num_internal_cleavage_sites = 2   ; maximum value is 5; for enzyme search
match_peak_count = 0                  ; number of auto-detected peaks to try matching (max 5)
match_peak_allowed_error = 1          ; number of allowed errors in matching auto-detected peaks
match_peak_tolerance = 1.0            ; mass tolerance for matching auto-detected peaks
partial_sequence =
sequence_header_filter =


add_C_Cysteine = 57.0215              ; added to C - avg. 103.1388, mono. 103.00919
```

TagRecon Configurations

ABRF Dataset Configuration

```
PrecursorMzTolerance= 0.01
FragmentMzTolerance = 0.5
NTerminusMzTolerance =  0.5
CTerminusMzTolerance =  0.5

AdjustPrecursorMass =  false

DuplicateSpectra = true
UseChargeStateFromMS = true
NumChargeStates = 4
UseSmartPlusThreeModel = true
TicCutoffPercentage = 0.98

CleavageRules =  "trypsin"
NumMaxMissedCleavages = 2
NumMinTerminiCleavages =  1
UseAvgMassOfSequences = false

DynamicMods = "M ^ 15.9949 C @ 57.021"
MaxDynamicMods = 3
StaticMods = ""
```

```
ExplainUnknownMassShiftsAs = "preferredptms"
PreferredDeltaMasses = "( 42.015 (Q -17.026 N -17.023 [DES] 21.981 [WYF] 15.996 W
3.994 W 31.989  K 92.105 R 185.628 C 88.62 C 47.73 (C 39.902"
MaxNumPreferredDeltaMasses = 2

Blosum = “blosum62.fas"
UnimodXML = "unimod.xml"
BlosumThreshold = -4

ComputeXCorr = true
MinCandidateLength =  5

MaxResults = 5
```

CPTAC Dataset Configuration

```
PrecursorMzTolerance= 1.25
FragmentMzTolerance = 0.5
NTerminusMzTolerance =  1.5
CTerminusMzTolerance =  1.25

DuplicateSpectra = true
UseChargeStateFromMS = false
NumChargeStates = 3
UseSmartPlusThreeModel = true
TicCutoffPercentage = 0.98f

CleavageRules =  "trypsin"
NumMaxMissedCleavages = 2
NumMinTerminiCleavages =  1
UseAvgMassOfSequences = true
MinCandidateLength =  5

DynamicMods = "M ^ 15.9949 (Q * -17.026 C @ 57.021"
MaxDynamicMods = 3
StaticMods = ""
# Path to the unimod.xml and blosum.fas files. These files are packaged with the
installation.
UnimodXML = /hactar/home/dasaris/bumbershoot/src/tagrecon/unimod.xml
Blosum = /hactar/home/dasaris/bumbershoot/src/tagrecon/blosum62.fas

ComputeXCorr = true

MaxResults = 5
```

X!Tandem Configurations

ABRF Dataset Configuration

```
        <enzymatic_search_constraint enzyme="trypsin"
max_num_internal_cleavages="1" min_number_termini="1" />
        <aminoacid_modification aminoacid="C" massdiff="57.0215" mass="160.0307"
variable="N" />
        <aminoacid_modification aminoacid="C" massdiff="-17.0265" mass="143.0042"
variable="Y" symbol="^" /><!--X! Tandem n-terminal AA variable modification-->
        <aminoacid_modification aminoacid="E" massdiff="-18.0106" mass="111.0320"
variable="Y" symbol="^" /><!--X! Tandem n-terminal AA variable modification-->
        <aminoacid_modification aminoacid="M" massdiff="15.9949" mass="147.0354"
variable="Y" />
        <aminoacid_modification aminoacid="Q" massdiff="-17.0265" mass="111.0321"
variable="Y" symbol="^" /><!--X! Tandem n-terminal AA variable modification-->

        <!-- Input parameters -->
        <parameter name="output, histogram column width" value="30"/>
        <parameter name="output, histograms" value="no"/>
        <parameter name="output, maximum valid expectation value" value="1"/>
        <parameter name="output, parameters" value="yes"/>
        <parameter name="output, path" value="C:\chen\ABRF-iPRG-
2009\xmls\sh_072808p_E_coli_ABRF_red.xml"/>
        <parameter name="output, path hashing" value="no"/>
        <parameter name="output, performance" value="yes"/>
        <parameter name="output, proteins" value="yes"/>
        <parameter name="output, results" value="all"/>
        <parameter name="output, sequences" value="yes"/>
        <parameter name="output, sort results by" value="protein"/>
        <parameter name="output, spectra" value="yes"/>
        <parameter name="output, title" value="Orbi X!Tandem"/>
        <parameter name="protein, C-terminal residue modification mass" value="0.0"/>
        <parameter name="protein, N-terminal residue modification mass" value="0.0"/>
        <parameter name="protein, cleavage semi" value="yes"/>
        <parameter name="protein, cleavage site" value="[RK]|[7]"/>
        <parameter name="protein, taxon" value="UniprotHuman"/>
        <parameter name="refine" value="no"/>
        <parameter name="refine, maximum valid expectation value" value="0.1"/>
        <parameter name="refine, spectrum synthesis" value="yes"/>
        <parameter name="residue, modification mass" value="57.0215@C"/>
        <parameter name="residue, potential modification mass" value="15.9949@M"/>
        <parameter name="scoring, maximum missed cleavage sites" value="1"/>
        <parameter name="scoring, minimum ion count" value="4"/>
        <parameter name="spectrum, dynamic range" value="100.0"/>
        <parameter name="spectrum, fragment monoisotopic mass error" value="0.5"/>
        <parameter name="spectrum, fragment monoisotopic mass error units"
value="Daltons"/>
        <parameter name="spectrum, maximum parent charge" value="4"/>
        <parameter name="spectrum, minimum fragment mz" value="150.0"/>
        <parameter name="spectrum, minimum parent m+h" value="500.0"/>
        <parameter name="spectrum, minimum peaks" value="15"/>
        <parameter name="spectrum, parent monoisotopic mass error minus"
value="10"/>
        <parameter name="spectrum, parent monoisotopic mass error plus" value="10"/>
        <parameter name="spectrum, parent monoisotopic mass error units"
value="ppm"/>
        <parameter name="spectrum, parent monoisotopic mass isotope error"
value="yes"/>
        <parameter name="spectrum, path type" value="mzxml"/>
        <parameter name="spectrum, threads" value="1"/>
        <parameter name="spectrum, total peaks" value="50"/>
        <parameter name="spectrum, use contrast angle" value="no"/>
        <parameter name="spectrum, use noise suppression" value="no"/>
```

CPTAC Dataset Configuration

```
    <enzymatic_search_constraint enzyme="trypsin" max_num_internal_cleavages="1"
min_number_termini="1" />
        <aminoacid_modification aminoacid="C" massdiff="57.0215" mass="160.0307"
variable="N" />
        <aminoacid_modification aminoacid="C" massdiff="-17.0265" mass="143.0042"
variable="Y" symbol="^" /><!--X! Tandem n-terminal AA variable modification-->
        <aminoacid_modification aminoacid="E" massdiff="-18.0106" mass="111.0320"
variable="Y" symbol="^" /><!--X! Tandem n-terminal AA variable modification-->
        <aminoacid_modification aminoacid="M" massdiff="15.9949" mass="147.0354"
variable="Y" />
        <aminoacid_modification aminoacid="Q" massdiff="-17.0265" mass="111.0321"
variable="Y" symbol="^" /><!--X! Tandem n-terminal AA variable modification-->

        <!-- Input parameters -->
        <parameter name="protein, C-terminal residue modification mass" value="0.0"/>
        <parameter name="protein, N-terminal residue modification mass" value="0.0"/>
        <parameter name="protein, cleavage semi" value="yes"/>
        <parameter name="protein, cleavage site" value="[RK]|[7]"/>
        <parameter name="protein, taxon" value="20110726-Yeast-Human"/>
        <parameter name="refine" value="no"/>
        <parameter name="refine, maximum valid expectation value" value="0.1"/>
        <parameter name="refine, spectrum synthesis" value="yes"/>
        <parameter name="residue, modification mass" value="57.0215@C"/>
        <parameter name="residue, potential modification mass" value="15.9949@M"/>
        <parameter name="scoring, maximum missed cleavage sites" value="1"/>
        <parameter name="scoring, minimum ion count" value="4"/>
        <parameter name="spectrum, dynamic range" value="100.0"/>
        <parameter name="spectrum, fragment monoisotopic mass error" value="0.4"/>
        <parameter name="spectrum, fragment monoisotopic mass error units"
value="Daltons"/>
        <parameter name="spectrum, maximum parent charge" value="4"/>
        <parameter name="spectrum, minimum fragment mz" value="150.0"/>
        <parameter name="spectrum, minimum parent m+h" value="500.0"/>
        <parameter name="spectrum, minimum peaks" value="15"/>
        <parameter name="spectrum, parent monoisotopic mass error minus"
value="0.5"/>
        <parameter name="spectrum, parent monoisotopic mass error plus" value="3.0"/>
        <parameter name="spectrum, parent monoisotopic mass error units"
value="Daltons"/>
        <parameter name="spectrum, parent monoisotopic mass isotope error"
value="no"/>
        <parameter name="spectrum, path type" value="mzxml"/>
        <parameter name="spectrum, threads" value="2"/>
        <parameter name="spectrum, total peaks" value="50"/>
        <parameter name="spectrum, use contrast angle" value="no"/>
        <parameter name="spectrum, use noise suppression" value="no"/>
```

APPENDIX C

Supplementary Materials 3

R Code Implementation for Vote Counting Model

```
data<-read.table("input path",sep="\t", header=T)

##The input file should be with the columns: label   Protein[MM    TR       XT       SQ
(whatever search engines to combine)]
protcount <- length(data[,1])
summary(data)



votesum_TP<-function(alpha,no.proteins,searchEngines){
#this function takes a p-value threshold, number of top proteins you want to choose,
#and for this trained dataset, it also calculate number of true positives in the top
proteins
##The input is like:
##votesum_TP(0.05,250,cbind(data$MM,data$XT,data$TR,data$SQ))
##The output is  like:
##votesum==0  votesum==1  votesum==2  votesum==3   votesum==4   True positives in
top N proteins'


no_engines=length(searchEngines[1,])
searchEng_p=matrix(rep(NA,protcount*no_engines),nrow=protcount,ncol=no_engines)
vot_sum=rep(NA,protcount)

min_searchEngine=rep(NA,protcount)

for (i in 1:protcount){
        for(j in 1:no_engines){
        if(!is.na(searchEngines[i,j])&&searchEngines[i,j]<=alpha)
                searchEng_p[i,j]=1
        else
                searchEng_p[i,j]=0
}
vot_sum[i]=sum(searchEng_p[i,])
if(sum(is.na(searchEngines[i,]))<no_engines){
min_searchEngine[i]=min(searchEngines[i,],na.rm=TRUE)}
}
list=data.frame(data,vot_sum)
sort.vot <- list[order(vot_sum,-min_searchEngine,decreasing = TRUE) , ]
TP_Top=sum(sort.vot$label[1:no.proteins])

vot_result=rep(NA,no_engines+1)
for (i in 0:no_engines){
vot_result[i+1]=length(subset(vot_sum,vot_sum==(i)))}

return (c(vot_result,TP_Top))
}

Calc_FDRmin=function(alpha,no.proteins,searchEngines){
##this function simulates permutations of p-values of each search engine among
proteins, and calculates a minimum FDR

no_engines=length(searchEngines[1,])

N_sum=votesum_TP(alpha,no.proteins,searchEngines)

sim=cbind(data$label,data$Protein)
for (i in 1:no_engines){
sim=cbind(sim,sample(searchEngines[,i]))}

E_sim_sum=votesum_TP(alpha,no.proteins,sim[,3:(3+no_engines-1)])
ratio=(E_sim_sum+1)[1:(no_engines+1)]/N_sum[1:(no_engines+1)]
mFDRmin=min(ratio)
num_threshold_engines=which(ratio==mFDRmin)-1
return(c(mFDRmin,num_threshold_engines))
}



eval_alpha=function(alpha,no.proteins,searchEngines){
##This function evaluates the validity of threshold with three criteria
##(1) If mFDRmin<a, these proteins were found to be differentially expressed at the
threshold a.
##(2) If not, I repeated the enumeration of votes with the value of a lowered by 20% at
each iteration until either a valid a is defined or the number of differential proteins
detected in two or more search engines reaches 0.
##(3) A valid a should not fall so far that the number of proteins with at least one vote
```

was less than N SELECT.

```
no_engines=length(searchEngines[1,])
if(sum(votesum_TP(alpha,no.proteins,searchEngines)[3:(3+no_engines-2)])==0){
##two or more signatures reached 0
        return(0)
}
actual_alpha=Calc_FDRmin(alpha,no.proteins,searchEngines)
if(actual_alpha[1]<alpha){
    ## if none of these proteins are at this threshold can yield no. of proteins
          ##This criteria is to avoid too small alpha
                if(sum(votesum_TP(alpha,no.proteins,searchEngines)[2:(2+no_engines-
1)])<no.proteins){
                #not enough proteins have significant p-values at this threshold, please
increase alpha'
                return(0)                }
                else{
                #SUCCESS
                return(alpha)            }}

##this is to avoid too big alpha
else{
                eval_alpha(alpha*.8,no.proteins,searchEngines)
  }
}

find_alpha=function(alpha,no.proteins,searchEngines){
##this function finds the smallest (most strict) p-value threshold that is valid through
eval_alpha validity test
while(eval_alpha(alpha*0.8,no.proteins,searchEngines)!=0){
alpha=alpha*0.8}
return(alpha)
}

#########################################################
##User input area


no.proteins=50
start_alpha=0.1

 searchEngines=cbind(data$MM,data$XT,data$SQ,data$TR)
```

```
# searchEngines=cbind(data$MM,data$XT,data$TR)
# searchEngines=cbind(data$MM,data$XT,data$SQ)
# searchEngines=cbind(data$MM,data$TR,data$SQ)
# searchEngines=cbind(data$XT,data$TR,data$SQ)

# searchEngines=cbind(data$XT,data$TR)
# searchEngines=cbind(data$MM,data$TR)
# searchEngines=cbind(data$MM,data$XT)
# searchEngines=cbind(data$MM,data$SQ)
# searchEngines=cbind(data$XT,data$SQ)
# searchEngines=cbind(data$TR,data$SQ)


###THE OUTPUT
PART********************************************************
no_engines=length(searchEngines[1,])
#must specify the number of protein list that you want from the begining

result_alpha=find_alpha(start_alpha,no.proteins,searchEngines)
print(paste('The best p-value threshold for highest sensitivity and specificity within the
top ',no.proteins, 'proteins is ',result_alpha))
result_alpha
no_TP=votesum_TP(result_alpha,no.proteins,searchEngines)
no_TP[length(no_TP)]


##output with the data and protein information with the vote sum

searchEng_p=matrix(rep(NA,protcount*no_engines),nrow=protcount,ncol=no_engines)
vot_sum=rep(NA,protcount)

for (i in 1:protcount){
        for(j in 1:no_engines){
        if(!is.na(searchEngines[i,j])&&searchEngines[i,j]<=result_alpha)
                searchEng_p[i,j]=1
        else
                searchEng_p[i,j]=0
}
vot_sum[i]=sum(searchEng_p[i,])
}


list=data.frame(data,vot_sum)
sort.vot <- list[order(vot_sum,-searchEngines[,1],-searchEngines[,2],decreasing =
TRUE) , ]
write.table(sort.vot,"output path",sep="\t",row.names =F)
```

APPENDIX D

Supplementary Material 4

Peptides Corresponding to Peptide Groups in ASW480 dataset Table S2

Peptide Group	1	2	3	4	5	6
Peptides	ELYEEELR	EYQDLLNVK	KLLEGEESR	LLEGEESR	TNEKVELQELNDR	DGQVINETSQHHDDLE
	NISEAEEWYK				VELQELNDR	DNLAEDIMR
	VYQVSRTSGGAGGLGSLRASR					EEAENTLQSFR
						EKLQEEMLQR
						ELRRQVDQLTNDK
						EMEENFAVEAANYQDTIGR
						EM1EENFAVEAANYQDTIGR
						ETNLDSLPLVDTH
						ETNLDSLPLVDTHSK
						FADLSEAANR
						FANYIDK
						FAVEAANYQDTIGR
						GTNESLER
						ILLAELEQLK
						ISLPLPNFSSLNLR
						KVESLQEEIAFLK
						LGDLYEEEMR
						LGDLYEEEM1R
						LHEEEIQELQAQ
						LHEEEIQELQAQIQEQH
						LLQDSVDFSLADAINTEFK
						LQDEIQNMK
						LQDEIQNM1K
						LQDEIQNMKEEM
						LQDEIQNMKEEM
						LQDEIQNMKEEMAR
						LQDEIQNMKEEMAR

LQEEMLQR
MALDIEIATYR
NLQEAEEWYK
QDVDNASLAR
QESTEYR
QQYESVAAK
QVDQLTNDK
QVQSLTCEVDALK
SLTCEVDALK
SLYASSPGGVYATR
SVSSSSYR
SYVTTSTR
TCEVDALK
TYSLGSALRPSTSR
VESLQEEIAFLK
VEVERDNLAEDIMR
VQIDVDVSKPDLTAALR
SLYASSPGGVYATR

Supplementary Material 5

Peptides Corresponding to Peptide Groups in HNSCC dataset Table S3

Peptide Group	1	2	3	4	5
Peptides	AEAELCAEAEETR	EDQSILCTGESGAGK	EEELQAALAR	ALELDPNLYR	IFEYIDR/IFEYLDR
	ALEEEQEAR	EDQSILCTGESGAGKTENTK	QLLQANPILEAFGNAK		
	AQAELENVSGALNEAESK	FDQLLAEEK			
	AQVTELEDELTAAEDAK	KFDQLLAEEK			
	DLGEELEALR				
	DLQGRDEAGEER				
	EAEALTQR				
	EAQAALAEAQEDLESER				
	EEIFSQNR				
	ELQTAQAQLSEWR				
	ELSSTEAQLHDAQELLQEETR				
	EQLEEEAAAR				
	EVGELQGR				
	EVVLQVEEER				
	FEEDLLLLEDQNSK				
	FEEDLLLLEDQNSKLSK				
	GELEDTLDSTNAQQELR				
	GLEAEVLR				
	KFEEDLLLLEDQNSK				
	LAEFSSQAAEEEEK				
	LALEAEVSELR				
	LAQAEEQLEQETR				
	LAQLEEER				
	LELQLQEVQGR				
	LGEEDAGAR				
	LLGLGVTDFSR				
	LQEELAASDR				
	QDEVLQAR				
	QDEVLQAR				
	QDEVLQARAQELQK				

	QEEEAGALEAGEEAR
	QLEEAEEEASR
	RQEEEAGALEAGEEAR
	RQLEEAEEEASR
	TLEEETR
	VAEQAANDLR
	VAQLEEER
	VGEEEECSR