BIOMEDICAL INFORMATICS

ENHANCED LC-MS/MS PROTEOMIC DIFFERENCE TESTING VIA INTEGRATION OF PEPTIDE ION INTENSITIES WITH SPECTRAL COUNTS

PETER STEVEN STRAUB

December 2010

Thesis under the direction of Professor David L. Tabb

Shotgun liquid chromatography/tandem mass spectrometry (LC-MS/MS) technology provides data sets rich in the type of information required proteomic quantitation; however, these data are not fully exploited by existing tools. We present a statistical model for combining MS precursor intensity data with MS/MS spectral count data and obtaining a single p-value using Fisher's Method of combining p-values. Our model is demonstrated using a new tool, IDPQuantify, which generates MS/MS spectral count data and MS persistent peptide isotopic distribution (PPID) intensity data for peptide group-level difference testing. Using the iPRG 2009 ABRF *E. coli* data set with known differences in protein content between cohorts, we compared the performance of existing candidate statistical tests using either spectral counts or PPIDs alone. We then compared the performance of our combined model with our candidate tests. Spectral count-based tests showed lower sensitivity but higher specificity than PPID-based tests. In comparison, our combined model yielded a slight drop in sensitivity coupled with an enormous improvement in specificity compared to the PPID-based test alone. We also observed that shared peptide groups tended to yield erroneous rejections of the null hypothesis more often than unshared peptide groups.

Approved _____ Date _____

ENHANCED LC-MS/MS PROTEOMIC DIFFERENCE TESTING VIA INTEGRATION OF PEPTIDE ION INTENSITIES WITH SPECTRAL COUNTS

By

Peter Steven Straub

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

December, 2010

Nashville, Tennessee

Approved:

Assistant Professor David L. Tabb

Professor David L. Hachey

Assistant Professor Bing Zhang

ACKNOWLEDGEMENTS

I would like to thank my thesis committee for their patience and guidance throughout this work. I would like to thank Dr. Tabb for mentoring me. I would like to thank Dr.Hachey for providing me with valuable information on my initial project involving phospholipidosis.

I am also grateful to many others who assisted me throughout. This includes the members of the Caprioli and Tyska labs (Whitney Parson and Jessica Mazerik in particular). The valuable data sets and advice provided by Assistant Professor Matt Tyska and Assistant Professor Rob Flynn are greatly appreciated. The biostatistical advice provided by Lorenzo Vega Montoto, Ming Li, and Dean Billheimer is also greatly appreciated. Michael MacCoss and Michael Hoopman were also very helpful in providing assistance in implementing their tools.

I am indebted to the Department of Biomedical Informatics for their patience and cooperation over the last few years. I have benefited immensely from my interactions with my fellow students, professors and staff at Vanderbilt, and I am deeply grateful for this. Funding from the National Library of Medicine (Grant 5-T15-LM007450-08) made this work possible and I am extremely honored that they chose to reward me with access to an education of such a high quality. I'd also like to thank Vanderbilt University Medical Center's Digestive Disease Research Center, supported by NIH grant P30DK058404, for the Core Services performed for this project.

Finally, I would like to thank my friends and family for their support throughout this entire educational experience.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

LIST OF ABBREVIATIONS

MS ..... mass spectrometry

LC-MS/MS ..... liquid chromatography/tandem mass spectrometry

VUMC ..... Vanderbilt University Medical Center

ROC ..... Receiver Operating Characteristic

AUC ..... Area Under the Curve

VUC ..... Volume Under the Curve

PID ..... Peptide Isotopic Distribution

PPID ..... Persistent Peptide Isotopic Distribution

TP ..... True Positive

FP ..... False Positive

TN ..... True Negative

FN ..... False Negative

SCX ..... Strong Cation Exchange

IEF ..... Isoelectric Focusing

CHAPTER I

INTRODUCTION

Proteomic difference testing has emerged as a critical tool in biomedical research for elucidating the molecular mechanisms underlying complex biological pathways. Biomarker detection, drug discovery and studies of variations in biological proteomes provide instances of research areas which would benefit from an enhancement in the ability to detect changes in protein abundance. While liquid chromatography tandem mass spectrometry (LC-MS/MS) (1, 2) has successfully been utilized to achieve this, many challenges remain to be addressed.

Two separate experimental approaches can be used to achieve protein quantitation: the first one involves isotope labeling (3-6) and the second approach utilizes label-free shotgun LC-MS/MS (7-10). While isotope labeling can provide more reliable results, the tradeoff is higher experimental cost and labor. Label-free shotgun LC-MS/MS reduces the complexity of experimental data generation while being statistically less powerful and prone to higher error rates than isotopic labeling. Given that reduction of experimental complexity is highly desirable in many situations, improvements in label-free LC-MS/MS difference testing are needed.

Proteomic difference testing can be approached qualitatively (i.e. *which* proteins vary between different test cases) as well as quantitatively (i.e. *how much* does a specific protein vary under different conditions). Label-free LC-MS/MS produces two fundamental types of data that can answer these questions: peptide precursor ion peak intensities and spectral counts of identified MS/MS spectra. Both types of data can be used separately for qualitative and quantitative difference testing. A myriad of different experimental and statistical techniques have

been developed in recent years to harness these two data types (3-10). However, spectral counts and precursor intensities follow fundamentally different statistical distributions that have their own inherent strengths and weaknesses when applied to difference testing, and neither data type can be considered exclusively superior to the other for the purpose of difference testing. Ideally, the information provided by both spectral counts and precursor intensities can be utilized in concert to balance their respective strengths and drawbacks.

We hypothesized that a statistical test integrating spectral count data with precursor intensity data could result in improved overall test results. To test this hypothesis, we developed a new tool called IDPQuantify. IDPQuantify was developed with the following specific aims in mind:

Specific Aim I. Evaluate the performance of statistical difference testing using precursor intensities or spectral counts. In particular, we were interested in establishing the degree of independence (orthogonality vs. mutual information) between statistical difference tests using either count data or intensity data alone.

Specific Aim II. Evaluate a new statistical model for label-free qualitative difference testing that combines precursor ion intensities with spectral count data. The model uses Fisher's Method for combining p-values into a single p-value using the p-values from a precursor intensities-based test and a spectral count-based test.

Specific Aim III. Establish the validity of conducting difference testing at the peptide group-level. Protein difference testing is typically performed at the protein-level. Peptide group difference testing segregates peptides into peptide groups that are associated with a single protein group or shared across multiple protein groups. We wanted to show that difference testing at the peptide group-level is effective at reducing the noise associated with shared peptides.

This study is structured as follows: In Chapter II, we review the published literature on label-free LC-MS/MS quantitation and the relationship between spectral count data and precursor ion intensities. In Chapter III, IDPQuantify and our combined model are introduced and defined. IDPQuantify's features are explained along with our combined statistical model. The existing statistical tests are used to evaluate and compare the performance of our combined model as well as assess the viability of peptide group-level difference testing. Chapter IV introduces our evaluation data set. The results of the statistical test comparative analysis are described and discussed. Chapter V contains the final conclusions of this study and future directions.

CHAPTER II

BACKGROUND AND SIGNIFICANCE

Proteomic difference testing falls into two general categories: qualitative difference testing (*which* proteins changed in abundance) and quantitative difference testing (*how much* did the proteins change in abundance). When LC-MS/MS data is used to conduct the difference testing, the data available for difference testing also falls into two general categories: MS/MS spectral counts and MS precursor ion intensities. When isotopic or mass-tag labeling is used, the internal standard's signal is also available. Label-free quantitation lacks an internal standard signal but remains appealing due to its lower cost and ease of implementation. As such, improving quantitative methods which eliminate the need for internal standards by relying exclusively on spectral counts or precursor ion intensities is highly desirable.

The choice of one of these metrics (spectral counts or precursor ion intensities) over the other for label-free quantitation depends on the type of difference testing question under investigation (which proteins changed vs. what was the fold change). But neither spectral counts nor precursor ions can be considered the clearly superior choice of data for either qualitative or quantitative difference testing. In 2004, Liu *et al.* established the viability of spectral count totals (per protein) as a statistically useful metric that was both randomly sampled and directly proportional to protein abundance over two orders of magnitude (10).

A 2005 study by Old *et al.* (11) investigated how precursor ion intensities and spectral counts differ in their abilities to detect the identities of the proteins which changed in abundance and what was the fold change across cohorts. Using isotopic labeling as their control, Old *et al.*

found that spectral counts and precursor intensities both perform comparably for either the qualitative or quantitative forms of difference testing. But they also observed that spectral count-based methods tended to be more reliable for answering the question of *which* proteins changed in abundance, whereas precursor intensities were generally better for determining the fold-change in abundance between cohorts. This is in keeping with a more recent study by Park that also found better agreement between the fold-changes found using LC-MS precursor peaks versus spectral counts (12).

The comparable, but still distinct, performance of spectral counts and precursor ion intensities for quantitation is also reflected in differences in their dynamic ranges and vulnerabilities to error. For instance, Old *et al.* found that both spectral count and precursor intensity methods suffered from erroneous fold change estimates for low abundance proteins(11). Spectral counts suffered from a lack of continuity as counts approached zero in one or both cohorts, leading to an overestimation of ratios. Precursor intensities, on the other hand, tended to underestimate ratios for low intensity precursors due to baseline intensity noise contributing to precursor signals.

On the high end of the abundance range, Old *et al.* found the most accurate fold-change estimates for proteins when measuring highly abundant proteins with large numbers of spectral counts in both cohorts(11). At the same time, Old *et al.* found an under-sampling of spectral counts for the most abundant proteins which they attributed to the data dependent acquisition exclusion lists. Precursor intensities may also become non-linear for the most abundant peptides when electrospray ionization is used, due, in part, to the impact of ion suppression (13).

The correlation between spectral counts and precursor ion intensities for relative abundance measurements has also been investigated (16). Using $^{14}N/^{15}N$ metabolic labeling as a

control, Zybailov *et al.* found a strong positive Pearson product moment correlation of 0.64 between abundance ratios calculated using spectral counts vs. precursor intensities. When minimum spectral count (per protein) filters and minimum precursor ion signal-to-noise filters were employed, the correlation was even stronger, highlighting the difficulties both spectral counts and precursor intensity methods face when dealing with less concentrated peptides. Also similar to the findings of Old and others (14, 15), Zaibalov *et al.* found that spectral count-based methods were more reproducible and had a larger dynamic range than precursor intensity methods. But the larger dynamic range of spectral counts was attributed, in part, to the respective zero-boundary issues (e.g. spectral count-based methods yielding higher ratios when one cohort has near-zero counts vs. baseline noise contributing to lower ratios when low abundant precursor ions are compared). More recently, Park *et al.* also found that both spectral counts and LC-MS precursor intensities could be used for abundance ratio approximation, but they found precursor peak areas under the curve to be a more reliable approach than spectral counts (12).

The larger dynamic range of protein abundance ratios when using spectral-count vs. precursor intensities observed by both Old *et al*. and Zybailov *et al*. also highlights the differences in the constraints required to conduct the different types of statistical tests available for count data vs. normally distributed area/volume data. For example, when assessing how spectral counts compared to precursor intensities for qualitative difference testing, Old *et al.* used minimum thresholds for determining whether a protein was viable for statistical difference testing using either spectral counts or precursor intensities. Proteins that had at least 4 spectra (found amongst all peptides across replicates) in either cohort were deemed viable for spectral count difference testing using a modified G-test. But when precursor intensities were used,

Student's t-test was employed, and their methodology required proteins to have at least three peptides shared between cohorts for difference testing. The threshold of four unpaired spectral counts per protein is much easier to meet than three shared peptides across cohorts, and thus nearly 10-fold more proteins were tested using spectral counts than were tested using precursor intensities.

Not surprisingly, the range of ratios seen in proteins found to be differentially abundant using spectral counts, alone, was ~4 fold greater than the range of ratios in proteins found to be differentially abundant using precursor intensities alone (11). So in addition to the tendency of spectral count to overestimate abundance ratios (due to zero-bound continuity issues), the larger range of ratios seen in proteins found to be differentially abundant using spectral counts is expected, given the much larger number of spectral count-based tests conducted.

The distributions of spectral counts and precursor peak intensities are fundamentally different. Precursor peak intensities are a measure of *how much* of a given peptide is observed. Intensities follow an exponential decay distribution within a given run but are, ideally, normally distributed across replicates for a given peptide (at a specific charge state). Appropriate statistical tests for normally distributed data (e.g. Student's t-test) depend on repeated measures per cohort (at least three) in order to approximate both the mean and variance. Spectral counts, on the other hand, are a measure of *how frequently* we observe a given protein's peptides. Count data follows a Poisson distribution and can therefore be applied to non-parametric statistical tests such as Fisher's exact test and the G-test. Non-parametric tests are potentially quite useful for proteomic LC-MS/MS data sets. They are more robust against outliers compared to their parametric counter parts and because Fisher's exact test and the G-test do not require more than one replicate per cohort and are more robust against missing data. Thus, the substantially larger

number of differentially abundant peptides identified by Old *et al.* when using spectral counts was indeed partially caused by the different thresholds employed for spectral counts and precursor intensities, but those different thresholds were reflective of the more relaxed constraints available to frequency-based statistical tests compared to the intensity-based parametric tests (16).

Spectral counts are not exclusively applicable to non-parametric tests. The Normalized Spectral Abundance Factor (NSAF), developed by Zybailov *et al.*, first normalizes each protein's spectral counts by the length of the protein and penalizes the total spectral count for larger proteins due to the greater number of peptides generated by longer proteins (16). The resulting Spectral Abundance Factor (SAF) for a given protein is further normalized to the entire data set by dividing the SAF by the sum of SAFs for all identified proteins to get the NSAF. The natural log of these NSAF values turns out to be normally distributed, making parametric tests like Student's t-test available for spectral count data (16, 17).

By using a normalization factor that takes into account the different sizes of proteins that could have generated the same peptides, the NSAF metric also begins to address another key challenge in quantitative proteomics: how to account for shared peptides. The "bottom up" approach of shotgun proteomics gathers data at the peptide level while researchers are interested in protein identification and quantitation. This introduces multiple ambiguities in the interpretation of LC-MS/MS data. Consider a peptide A that that could have come from two proteins of differing lengths: protein 1 (longer) and protein 2. If protein 1 and protein 2 were both at equal abundance we would expect the longer of the two to generate more peptides and contribute more spectral counts. As such, if the spectral counts for peptide A are contributing to protein 1's total count, each individual spectral count is being drawn from a larger pool of

theoretical counts and should be given less weight than if the counts were being attributed to the shorter protein 2.

The original NSAF approach helped to address how to weight peptide spectral counts when attributing the same peptide's counts to proteins of differing lengths, but it left unresolved the issue of attributing the same spectral counts to multiple proteins. By adding the spectral counts of shared peptides to each potential parent protein, we are assuming that each of those proteins contributed equal numbers of precursor ions to each identified MS/MS spectra, a clearly erroneous assumption. The greater the number of proteins that could have contributed to a given peptide, the greater the ambiguity in the identity of the protein which contributed the peptide and the greater the number of extra counts that will be erroneously applied if shared peptide counts are equally applied to all of the possible protein contributors. For complex proteomes, shared peptides can be particularly troublesome. The human genome, for instance, is estimated to have ~1,000,000 protein isoforms, with 5-7 isoforms on average per open reading frame (18). Distinguishing between post-translationally modified proteins and their unmodified forms presents another source of shared peptide ambiguity. And the lower the abundance of a given protein, the greater the relative error for each erroneously attributed spectral count.

The best way to deal with all of these complications arising from shared peptides remains an open area of research. Zhang *et al.* recently compared the original NSAF method to several alternative approaches that accounted for shared peptides in the normalization process (19). They found the normalized spectral counts from the original NSAF method to have the worst linearity due, in part, to the multiple counting of shared peptides causing a systematic overestimation of the concentrations of less concentrated proteins. The top performing method they observed for addressing shared peptides involved distributing all of the spectral counts from

shared peptides to their potential parent proteins. The fraction of the shared spectral count that each potential parent protein receives is equal to the fraction of spectral counts from unshared peptides that a given protein received amongst the total number unshared spectral counts for that set of potential parent proteins. Zhang *et al.* (19) also evaluated the approach by simply discarding shared peptides from consideration. While this method had a superior dynamic range to that of the original NSAF method, it also tended to underestimate less abundant proteins.

Usaite *et al.* (15) also found a loss of linearity for spectral counts from low abundance proteins and demonstrated that shared peptides give much more variable and unreliable abundance ratios if the parent proteins are expressed at different levels (15). Using the spectral count data from two highly homologous proteins expressed at different levels, they calculated the abundance ratios for each separate peptide from the two proteins (using the original NSAF method for normalization). The abundance ratios of the unshared peptides clustered around two distinct population levels, whereas the shared peptides gave highly variable ratios. When they statistically difference tested several thousand proteins from a yeast proteome between Wild Type (WT) and Knockout (KO) cohorts, they found that 12% of the proteins had statistically different results depending on whether or not shared peptides were used.

An early metric similar to the NSAF, the Protein Abundance Index (PAI), was based not on spectral counts, but on the percentage of a protein's theoretical sequence that was identified in the runs. The PAI normalization factor simply divided the total number of unique peptides measured for a given protein by the theoretical number of unique peptides that could be generated from the protein (20). This led to the creation of the Exponentially Modified Protein Abundance Index (emPAI) by Ishihama *et al.* for use with precursor intensity-based quantitation based on the observation of a linear relationship between the PAI values and the natural log of

protein abundances (21). As would be expected, they also observed a linear relationship between a protein's spectral counts and its abundance but also observed that the number of unique peptides identified for a protein was linearly correlated with the log of the protein's abundance, suggesting a linear increase in observed sequence coverage corresponds to an exponential increase in protein abundance. The emPAI index is the exponential of the PAI and was shown to be strongly correlated with protein abundance (2.1):

$$\text{emPAI} = 10^{\text{PAI}} - 1 \qquad (2.1)$$

Interestingly, while sequence coverage may be the highest for the most abundant proteins, Ishihara *et al.* also found a reduced correlation between protein abundance and the emPAI metric for the most abundant proteins, highlighting how the most abundant proteins may be the easiest to identify but not necessarily the easiest to quantify (21).

A third normalization technique, the Protein Abundance Factor (PAF) developed by Link *et al.,* constitutes a mix of the NSAF and emPAI normalization techniques by factoring in both the extent of sequence coverage by non-redundant MS/MS spectra and the molecular weight of the candidate protein when normalizing protein concentrations (22). This approach can be used in the freely available tool BigCat (23).

NSAF, emPAI, and the PAF are examples of normalization and analytical approaches that are conducted at the protein level. An alternative approach to dealing with shared peptides is to first segregate peptides and proteins into peptide groups and protein groups, and conduct quantitative analysis at the group level instead of at the individual peptide level. Peptide groups consist of all peptides that can be explained by the same set of proteins, whereas protein groups

are defined as all proteins that can be explained by the same set of peptide groups (Figure 1). Protein groups and peptide groups are already used for assembling identified proteins from observed peptides via parsimonious analysis (i.e. restrict identified proteins to the smallest set of proteins required to explain the presence of all identified peptides) (24, 25). Instead of addressing shared peptides between unique proteins at the protein-level and peptide-level, group-level analysis segregates between unique peptide *groups* and shared peptide *groups*. All peptides in a given peptide group are considered to be shared between all of the proteins in each protein group associated with the peptide group.

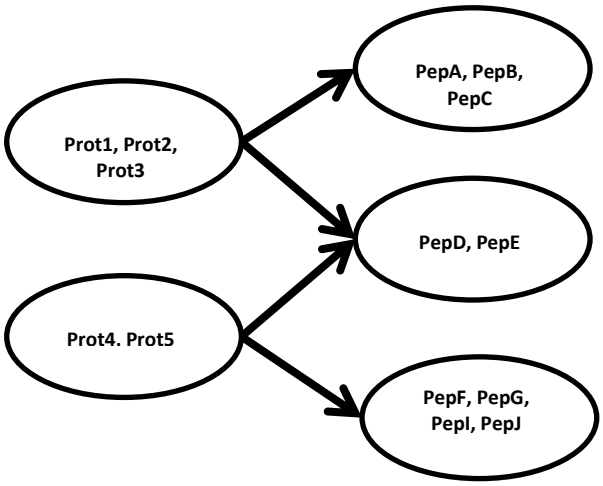


Figure 1. A schematic of the relationship between protein groups and peptide groups as defined in this study. "Shared Peptides" are considered to be peptides in a peptide group shared between two protein groups.

Quantitative peptide group-level analysis instead of protein-level analysis assumes a high degree of similarity in the relative expression levels of different proteins in the same protein group. This is not an entirely unreasonable assumption since proteins in the same protein group presumably share a high degree of sequence homology and are likely in the same protein family. Increased sequence homology (i.e. the number of peptides in a peptide group) confers greater likelihood of shared transcription factors and other regulatory mechanisms (26). Jin *et al.* tested this assumption and found that proteins in the same protein group do, indeed, appear to be functionally related and experience similar changes in relative abundance when the biological system is perturbed (27). They also found that the addition of shared peptides (within a given

protein group) did not negatively impact abundance calculations. This observation can be reconciled with previous findings by noting that Usaite *et al.* segregated peptides based on whether or not they were shared or unique to the candidate protein(15). Jin *et al.* took a different approach by segregating peptides based on whether or not they were shared between different sets of *functionally related* proteins or shared between different sets of *functionally unrelated* proteins. Overall, their findings suggest that shared peptides within an unshared peptide group should not be discarded, whereas the peptides in a shared peptide group may be more problematic for quantitation.

If the peptides found within a given unshared peptide group are exclusively used to conduct quantitative analysis, NSAF or some other normalization methods could presumably be applied separately for each protein in the associated protein group. Alternatively, protein-specific normalization could be skipped and each protein in the protein group could be treated equally. There are tradeoffs to both approaches. Normalizing each protein separately has the obvious advantage of factoring in differences in protein length and/or sequence coverage. But testing each protein separately comes at the cost of conducting a separate statistical test for each individual protein whereas group-level analysis requires only a single test per protein in a group. The fold-change reduction in the number of statistical tests would be equal to the average number of proteins in each protein group tested. Thus, type I and type II errors may be simultaneously reduced by protein-specific normalization but also increased due to additional testing vs. protein-group testing.

Whether or not proteins are tested at the protein or protein group level has the potential to impact the application of multiple testing correction methods. These methods skew results towards fewer false positives at the cost of more false negatives. Independence between separate

tests that generate a p-value is a requirement for many multiple testing correction methods. Proteins in the same protein group that are normalized separately have a clear violation of independence, complicating the application of multiple testing corrections. Similar problems arise if difference testing is restricted to the protein group-level using, all peptides from shared or unshared peptide groups associated with a given protein group.

Difference testing at the peptide group-level is a third alternative. It has the obvious disadvantage of reducing the overall amount of data used for a given protein's or protein group's testing. But by splitting the signal between shared and unshared peptide groups, the unshared peptide groups provide units of peptide signal grouped for greater sequence homology. This is because the proteins in a given protein group tend to be more homologous to each other than with randomly selected identified proteins. And as already mentioned, greater sequence homology is associated with a greater likelihood of shared transcription factors, and greater similarity change in expression in response to stimulus (26).

The issue of multiple testing correction is also a challenge in the field of mRNA microarray analysis, where thousands of genes are tested simultaneously, often on few replicates. Not surprisingly, LC-MS/MS proteomic analysis has borrowed a number of the data normalization and statistical testing techniques developed for microarray techniques (17, 28). One existing tool, QSpec, uses the hierarchical Bayes methodology for approximating false discovery rates following statistical analysis on spectral counts (29). Another technique, the empirical Bayes t-test, was designed for addressing the problem of type I and type II errors randomly arising from multiple tests on too few replicates. First developed by Lonnstedt and Speed (30), it uses Student's t-test with a modified t-statistic calculated using pooled variances empirically derived from the standard deviations observed over the entire data set.

A variant of the empirical Bayes t-test was made available in the "limma" package for R by Smyth (31, 32) and has already been employed in the label free LC-MS proteomic difference testing software package Corra (33). Corra uses LC-MS precursor intensity data to conduct its difference testing. In addition to using the empirical Bayes t-test to minimize false positives, Corra also addresses a problem not faced with spectral-count-based quantitation: peak alignment. Determining which MS1 peaks correspond to which peptide can be complicated because not all peptides are isolated and identified in each run. Elution profiles can change from one run to the next and therefore peak area normalization or inference of an unidentified precursor peak's peptide can require complex, computationally expensive algorithms.

Rosetta's "Elucidator," a commercially available software tool, is another software platform designed for label free or labeled quantitation using LC-MS or LC-MALDI precursor intensities and peak alignment algorithms (34). Elucidator groups MS1 precursor peaks based on their chromatographic parameters (retention time, m/z, inferred charge state, etc). Following peak alignment (using PeakTeller) and normalization, Elucidator conducts analysis of variance (ANOVA) on grouped peaks in order to identify candidates of differentially abundant peptides for subsequent MS/MS identification. This stands in contrast to spectral count-based label-free methods that group spectral counts based on identified peptides (and associated proteins). This highlights a fundamentally different aspect between MS and MS/MS data. MS data for *all* peptides are collected in profiling mode, but the specific peptides that contributed to each peak's intensity are ambiguous without MS/MS data. MS/MS spectral count data, on the other hand, is only collected for *some* peptides, but it also doesn't suffer from the ambiguity of precursor peaks because only identified MS/MS spectra are used for spectral counting. As such, de-noising steps, such as deconvolution of peak intensities of precursor peptide isotopic distributions (PIDs)

from their neighbors, are requirements for MS1 quantitation, but not for spectral count-based methods.

Another way in which the grouping of data points for quantitation can differ between precursor intensity-based methods and spectral count-based methods is the need for the exact same peptide (often at the same charge state) to be identified in multiple runs across cohorts. When precursor intensity methods are used, whether isotopic or label free methods are chosen, each individual peptide's precursor peaks are paired across runs and analyzed separately. Label free, spectral count-based methods group all spectral counts for a given protein or peptide group together within a given cohort or technical replicate, regardless of whether the same peptides were found across runs. Consequently, missing data is potentially a much larger issue when conducting precursor intensity-based methods because it may limit the number of peptides (and proteins, indirectly) that are viable candidates for a given statistical test available for precursor intensity analysis (i.e. Student's t-test requires three data points per cohort).

QuasiTel, a tool recently developed for spectral count-based analysis by Li *et al.*, was designed to address complications that can arise during quantitation of less abundant peptides using spectral counts (35). As discussed above, count-based based methods can be problematic for low abundance proteins, partly due to the lack of continuity as counts approach zero. Like precursor intensities, Poisson distributions are exponential and equal variances across the range of values cannot be assumed based on empirical evidence, limiting the statistical methods available for analysis. Like the NSAF normalization method, which was shown to exhibit equal variance when the natural log of the NSAF was used (16), QuasiTel's developers proposed the use of a quasi-Poisson likelihood maximization function to be used to fit spectral count data to linear models that assume equal variance (e.g. a Poisson regression). They spiked proteins at

different ratios to compare the performance of their quasi-likelihood model to detect differentially abundant proteins with the performance of classical difference tests (e.g. Fisher's exact test). They found their new model to perform comparably to the classical tests but not identically. In particular, when the spiked protein ratios were high (ratio 27) their quasi-likelihood performed quite similarly to the other tests, but when the ratios were at the lowest levels tested (ratio 3), the quasi-likelihood model demonstrated a clearly superior ability to detect differentially abundant proteins, albeit at the cost of more false positives. As the authors pointed out, an advantage of their model over classical non-parametric tests is the flexibility to adapt spectral counts to the fitted models to include additional parameters, such as the type of instrument used.

The observation of the QuasiTel authors that their model performed better at detecting differentially abundant proteins that were close in abundance, but at the cost of more false positives, is a reminder that different statistical methodologies present a different set of tradeoffs between sensitivity and specificity. Which test a researcher finds appropriate depends, in part, in the researcher's goals. If candidate proteins based on differential abundance are desired for subsequent analysis where the cost of false positives is not high (e.g. biomarker detection), a test like the QuasiTel model that yields more true positives at the expense of additional false positives may be the optimal choice. On the other hand, if there is a high cost for false positives, a more conservative test like Fisher's exact test may be desired. Such tradeoffs apply to the choice of precursor intensities over spectral count data for quantitation. Based on published studies, if a researcher is most interested in calculating abundance ratios between cohorts, both spectral counts and precursor intensities can accomplish this, but precursor intensities would probably provide the more accurate results. If the researcher desires to know which proteins

changed in abundance regardless of the fold-change, a spectral count-based approach would be recommended.

The published research on LC-MS quantitation provides a vast array of methods available for quantitation, but these are proposed as "either/or" options from which a researcher should choose just one to conduct their analysis. Some of the numerous tools available, such as Census, are capable of quantitative analysis using either spectral counts or precursor intensities (12). A fascinating question raised by the wide variety of data types and methods available for quantitation is whether or not there is a method for combining the information from both precursor intensities and spectral counts into a single model for quantitation.

Spectral count and precursor intensity quantitation represent two different methods of viewing the same underlying system. This suggests the possible use of meta-analysis techniques for combining the separate quantitative statistical test results on the same protein (or peptide group) from separate spectral count and precursor intensity tests. A number of different meta-analysis techniques exist for combining the results of multiple statistical tests. One method, the Winer method (36), involves the combination of t-statistics and degrees of freedom to generate a single new Z-statistic. This method however, approximates the variance of the Z-statistic using the degrees of freedom from the underlying combined tests and suffers if the tests were conducted on small numbers of samples (e.g. ≤ 10). Similarly, Stouffer's method generates new Z-statistics from the separate Z-statistics of the combined tests but it also suffers from reduced degrees of freedom, which is commonplace in LC-MS/MS data with few replicates (36).

The original meta-analysis method proposed by R. A. Fisher (referred to as "Fisher's Method") is rooted in the observation that the p-values from continuously distributed test statistics will have a uniform distribution under the null hypothesis for any underlying data

distribution (37). Fisher's Method is sometimes called the Inverse Chi-squared method because it uses a relationship between the uniform distribution and chi-squared distribution, namely that -2 * log of a uniformly distributed value will follow a chi-squared distribution on 2-degrees of freedom. Additionally, the sum of chi-squared distributed values will also follow a chi-squared distribution (37). These observations allow for the combination of the product of multiple p-values into a single test statistic, X^2 that follows a chi-squared distribution with 2k degrees of freedom, where k is the number of separate tests (2.2):

$$X^2 = -2\sum_{i=1}^{k} \ln(p_i) \qquad (2.2)$$

We can then generate a one-tailed p-value for X^2 given the 2k degrees of freedom. Fisher's Method of meta-analysis is generally considered the most powerful and widely recommended meta-analysis technique(37). The application of Fisher's Method to LC-MS/MS quantitation is an intriguing possible approach to building a statistical model that incorporates both spectral counts and precursor intensities.

Significance

The goal of this study was to test the hypothesis that proteomic difference testing could be improved by combining spectral count and precursor intensity data into a single test. We also wanted to compare the performance of the combined model to an array of classical statistical tests as well as the newer empirical Bayes t-test. Our combined model consists of using Fisher's Method of combining the p-values of two of our candidate tests. In Chapter III, we describe the

new combined statistical model and candidate tests in greater detail. We also describe IDPQuantify, a new module of the IDPicker protein identification software suite used to generate the spectral count and precursor intensity data to conduct the statistical testing. IDPQuantify reads protein and peptide identification results from IDPicker, a tool designed to confidently assemble identified proteins from peptide searches on shotgun proteomic data. IDPicker can take advantage of multiple scores per peptide and decoy FDR techniques for confident peptide identification coupled (38). Protein assembly with IDPicker includes bipartite analysis at the peptide group and protein group level for parsimonious scoring of candidate protein identification (24).

IDPQuantify extends the IDPicker pipeline by providing quantitative data for the peptides and proteins from IDPicker's protein identification results. IPDQuantify writes spectral count and precursor intensity data at the peptide and peptide group level (where "peptide group" is defined as all peptides shared by the same protein group). Unlike precursor intensity quantitation tools which conduct precursor intensity analysis on peaks prior to peptide identification, IDPQuantify restricts precursor ion intensities to those peaks that generate an identified spectrum.

IDPQuantify can also be configured to feed its quantitative output files to user-written R scripts for automated statistical analysis. In Chapter IV we describe our use of the 2009 Proteome Informatics Research Group (iPRG) *E. coli* data set, along with IDPQuantify and R, to compare and contrast multiple statistical difference tests using either spectral count data or precursor intensity data. These results were then compared to our combined model. Using the F_1-measure and ROC curves as our metrics, we show a significant improvement is possible by

combining both spectral count data and precursor intensity into a single statistical model conducted at the peptide group level.

In this study we also introduce difference testing at the peptide group level instead of at the protein level as is normally done. By grouping data (spectral counts or precursor intensities) at the peptide group level we are using observations made by Jin *et al.* that proteins in protein groups tend to be functionally related and also tend to change in abundance in a positively correlated manner (27). Thus, we are addressing the issue of spectral counts from shared peptides differently than the various shared peptide normalization techniques described above. Unlike approaches like NSAF or emPAI, our approach does not attempt to distribute the signal of shared peptides between potential parent proteins or normalize that signal separately for each protein as long as those proteins are in the same protein group and therefore assumed to be functionally related and likely to change in abundance similarly. In doing so, we lack the added accuracy potentially obtained by normalizing each protein's spectral count to the protein length, but we gain several important advantages. First, by not distributing the signal of spectral counts from shared peptides between the potential protein parents, we add to the power of our tests by using more information (in the form of spectral counts) for each statistical test. Combining counts at the peptide group level also helps address the problems discussed above that spectral-count based quantitative methods tend to have when counts are low. We also avoid complications associated with multiple testing corrections and the requirement for independence between tests. Proteins within the same protein group that are statistically tested separately are clearly not independent given their shared peptide signals. Additionally, significantly more statistical tests will have to be conducted if done at the protein level instead of the protein group level. By quantitating at the peptide group level we are reducing the overall number of tests, and

thus reducing the likelihood of type I errors, while increasing the independence between tests which makes the results amendable to multiple testing correction procedures.

CHAPTER III

IDPQUANTIFY AND STATISTICAL DIFFERENCE TESTING

Introduction

We introduce IDPQuantify, a new component of the IDPicker pipeline that generates output of precursor ion intensities from identified spectra at the peptide group level. IDPQuantify calculates precursor ion intensities using input files generated by two third-party tools built by the MacCoss group, Hardklör and Bullseye (39, 40). IDPQuantify is also capable of running user-written scripts in R for automated difference testing at the peptide group-level with precursor ion intensities and/or spectral counts (41).

Utility of IDPQuantify

IDPQuantify is a new open source module written in C# for use in the IDPicker proteomics pipeline (38). IDPQuantify reads idpXML files created by IDPicker, along with spectral source files in the .mzXML or .Raw formats, and generates output files containing spectral count data and precursor ion intensities for each peptide group found in each replicate. The output files are designed for easy extraction of peptide group-level for use in statistical difference testing in third party tools such the statistical package R. In addition, IDPQuantify produces a variety of normalized precursor intensity files (described below) that can alternatively be used for difference testing.

Users can also set up IDPQuantify to automate any number of difference tests written in R via the "analysis_config" text file. Within the "analysis_config" file, the idpXML input files used to generate the spectral count and PPID files are then assigned to cohort 1 or cohort 2 for difference testing. Users can create custom-made R scripts written to accept IDPQuantify's PPID and spectral count files and output the test results in a specified format (see Supplemental Materials File 4). Which PPID and/or spectral count files are to be processed by which R script is defined in the "analysis_config" file.

Hardklör, Bullseye, and Persistent Peptide Isotopic Distributions for Precursor Intensity Information Extraction

To generate ms1 precursor intensity data, IDPQuantify uses input files generated by two third-party tools, Hardklör (39) and Bullseye (40), both developed by the Hoopman *et al.*. Hardklör reads spectral data files (.RAW, .ms1, or .mzXML) and uses a novel algorithm to rapidly identify peptide isotopic distributions (PIDs) within ms1 scans. For each ms2 scan, Bullseye attempts to identify persistent PIDs found across sequential ms1 scans for the selected precursor peptide (Figure 2).

Hardklör for Isotopic Deconvolution and Precursor Peak Intensity Estimation

Hardklör uses a novel approach to deconvoluting neighboring isotopic distributions by using an efficient single-pass charge state inference algorithm. As the Hardklör authors point out, efficient charge state inference is useful in quantitative applications because it allows for a combinatorial approach to deisotoping overlapping PIDs that can be otherwise computationally prohibitively expensive (39). The number of theoretical combinations of overlapping PIDs that

must be considered when using the combinatorial approach grows exponentially with the number of charge states to be considered for the overlapping PIDs. Hardklör's efficient elimination of charge states from consideration for a given set of PIDs can significantly reduce the computation costs of the combinatorial approach, making it available when less computationally expensive methods might otherwise be used. Those include choosing a single PID to start the analysis with and iteratively subtracting computed PID signals from the remaining signal to be analyzed. These alternative approaches are vulnerable to the cascading effects of errors introduced early in the chain of analysis, which can be particularly harmful to quantitative studies. The ability to use the combinatorial approach for more accurate approximation of PID peak intensities is especially useful when analyzing complex samples that have had limited separation, such as one dimension LC separation, as opposed to multidimensional separation involving strong cation exchange (SCX) or isoelectric focusing (IEF).

Hardklör writes .hk files that contain summary information for each PID found in each MS1 scan in a given spectral data file. The intensity of the base peak of each PID is included in that output. As part of Hardklör's peak-picking algorithm, base peak signal is de-noised using the THRASH algorithm (39, 42). The THRASH algorithm defines baseline noise in a given m/z window as the mode of the distribution of peaks observed at that window and subtracts that from the measure base peak's area under the curve (AUC).

Bullseye for Precursor Peak Quantitation Using PPIDs

Peptides eluting from the column over a long enough chromatographic window can be captured by multiple sequential MS1 scans. If precursor intensities are to be used for

quantitation, we need to estimate the precursor peak intensities across the entire chromatographic window of a given peptide. Bullseye, a companion tool built to work with Hardklör, accomplishes this task. Bullseye was written for the purpose of more accurately estimating precursor ion m/z values than is typically done by instrument hardware/software. A given PID eluting over time that is measured in multiple MS1 scans may have slightly different m/z values in each MS1 scan, reflecting the noise in precursor m/z values used in peptide identification. Bullseye assumes PIDs that appear in sequential MS1 scans with approximately the same monoisotopic mass all come from the same peptide. As Hoopman *et al.* demonstrated, the average m/z of the precursor PID monoisotopic peaks observed over the entire peptide elution profile can yield a more accurate estimation of precursor monoisotopic m/z values than from a single scan alone (39). Bullseye's ability to accurately estimate precursor monoisotopic m/z values resulted in more accurate peptide identifications even when compared to the m/z values provided by a high-resolution Orbitrap (39).

Bullseye reads Hardklör's output files along with an ms2 spectral data file (.RAW or .mzXML). For each identified MS2 spectrum, Bullseye locates ms1 persistent PIDs (PPIDs) for the precursor ions corresponding to each MS2 spectra. Bullseye defines a PPID as the same PID in three out of four subsequent MS1 scans. PPID intensity values consist of the summation of the de-noised and baselined peak AUCs of the PID's base peak for each ms1 scan in the PPID. Thus, each PPID is a Volume Under Curve (VUC) calculation of the PPID's base peak (Figure 2). Users can generate Hardklör/Bullseye files on their own or let IDPQuantify automate the

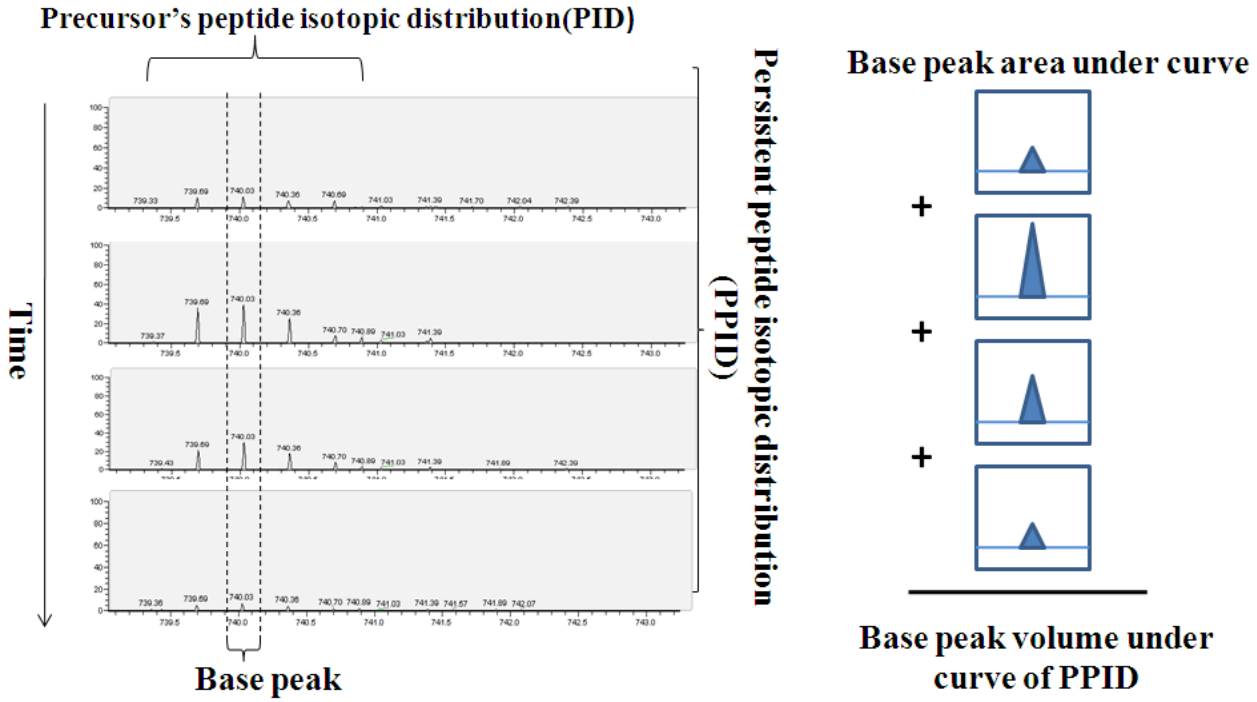


Figure 2. Precursor intensity information extracted by IDPQuantify consists of the volume under the curve (VUC) of the base peak of a persistent peptide isotopic distribution. IDPQuantify uses two third party tools to obtain PPIDs. First, peptide isotopic distributions (PID) in each MS scan are deconvoluted and identified by Hardklör. Next, for each MS/MS scan, Bullseye examines the triggering ion's PID and attempts to identify a corresponding persistent PID (PPID). The sum of the areas under the curve of the PIDs base peak is reported by Bullseye and used by IDPQuantify.

generation of the Hardklör and Bullseye files. IDPQuantify's basic unit of quantitation at the peptide level is the corresponding PPID generated by Hardklör and Bullseye.

Low intensity PIDs can result in atypical isotopic packets that make PID identification difficult and result in missing data points in our PPID analysis. When an identified MS/MS spectrum does not have a corresponding PPID, IDPQuantify attempts to construct a PPID using the mass and charge of the identified peptide and the precursor triggering ms1 scan peaks. If an incomplete PID is detected within a user-set retention time window (e.g. +/- 10 seconds around triggering ms1 scan), IDPQuantify reattempts to construct a PPID for that peptide using the incomplete PID as a starting point.

It is important to note that Bullseye identifies a PPID in a given replicate only if that PPID was selected for MS/MS identification. The tradeoff between this approach and other

methods that use peak alignment to identify unidentified precursor ions is that IDPQuantify's PPID data will include more missing data points, but also a reduced likelihood of misaligned/misidentified precursor peaks.

Output Files Generated by IDPQuantify

IDPQuantify generates a variety of output files containing spectral count and PPID data at the peptide and peptide group levels. The generation and use of these files is handled by IDPQuantify. The "*Distinct peptides per peptide group per replicate"* file contains the number of distinct peptides observed in each replicate in each peptide group. The cohort-level totals and all-replicate-level total counts are simple summations of the replicate-level unique peptides. The "*Unique peptides per peptide group per cohort/run"* file contains the number of unique peptides observed in each replicate in each peptide group and the unique peptides across each cohort and all replicates.

The "*Spectral count per peptide group"* file contains the number of MS/MS spectra that identified a peptide in a given peptide group in each replicate and across each cohort and all replicates. Similarly, the "*Spectral count per peptide*" file simply contains the number of spectral counts per peptide for each replicate. A related file is the "*Num peptide hits per peptide group with a PPID observed"* file contains the count of peptides in a peptide group that had at least one PPID observed in a given replicate. If multiple spectral counts are observed for a given peptide in a given replicate, that peptide will add only one to the count as long as at least one of those spectral hits had a corresponding PPID found.

The "*PPID intensity per peptide file"* file contains the total PPID intensity per individual peptide per replicate. If multiple MS/MS spectral hits are observed for a given peptide in a given replicate, the intensity values of observed PPIDs are summed. If no PPID is observed but at least one spectral hit is seen for a peptide in a given replicate and error code of -2 is listed. If no PPID is observed and no spectral hit is seen, an error code of -4 is listed. The *"Total PPID intensity per peptide group"* file contains the total PPID intensity for a given replicate for all peptides in a peptide group.

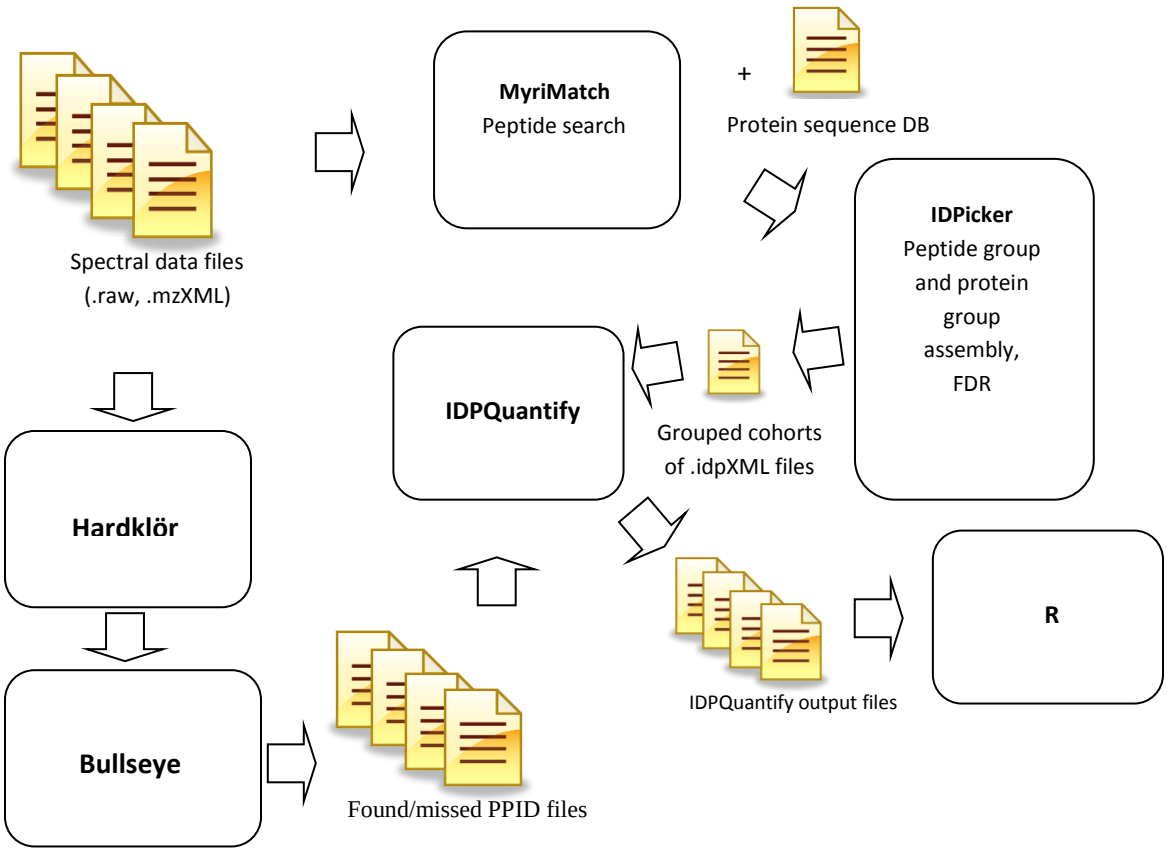


Figure 3: An overview of the PPID data file generation workflow utilized by IDPQuantify.

Statistical Difference Testing of LC-MS/MS Data

The *E. coli* data set used for evaluating test performance is described in greater detail in Chapter IV. The data set was difference tested using a variety of candidate statistical tests. The tests fall under general categories of classic statistical tests, the empirical Bayes tests, and hybrid tests using the geometric mean of spectral count and PPID-derived p-values combined using Fisher's Method. We refer to the use of Fisher's Method in this manner as the "Combined Model" throughout the rest of this study.

Two methods for weighting PPID values were tried for each PPID-based test, along with no weighting. PPID or spectral count-based tests also had two optional filters applied to eliminate low-information peptide groups from consideration. Each candidate test was run for all available combinations of weighting and filtering methods (and at different filtering levels). See below for details on weighting and filtering methods.

Classic Statistical Tests

We used three spectral count-based tests: Fisher's exact test, the Mann-Whitney U test, and a normalized Mann-Whitney U test (see below). Student's t-test and Welch's t-test were run using PPID data, while Fisher's exact test used spectral counts**.** The modified Mann-Whitney U test was chosen in order to minimize the occurrence of tied spectral counts. It was conducted as follows: First, spectral counts were summed for each replicate. The first replicate in cohort 1 (as specified in the "analysis_config" file) is chosen as a reference replicate for normalization. Next,

for each peptide group tested, the spectral count for each replicate is multiplied by the normalization factor:

$$Normalization\ Factor = {}^{(\text{Total spectral count of this replicate})}\Big/_{(\text{Total specral count of reference replicate})} \quad (3.1)$$

The modified Mann-Whitney U test is run on these normalized spectral counts. We also tested the Mann-Whitney U test using PPID values normalized on a 0-101 scale (see "Normalization Approaches" below).

Limma and the Empirical Bayes t-test

The "Linear Models for Microarray" (Limma) package for R was designed for microarray data sets (43). We used the empirical Bayes test in Limma using a variety of data normalization and transformations described below. Limma's empirical Bayes test uses a moderated t-statistic based on estimated sample variances derived using the entire microarray probe set data. The moderated t-statistic is calculated using posterior residual standard deviations in place of the normally derived standard deviations, shrinking variances towards a pooled estimated variance. The pooled variance, in turn, reduces the probability of under-estimated sample variances for a given gene (or peptide group) that will result in false positives from artificially high t-statistics.

The Combined Model: Fisher's Method for Combining p-values from Spectral Count and PPID Data

The final statistical technique evaluated in this study was the generation of a single p-value from the p-values of two separate tests using Fisher's Method for combining extreme probabilities. Any number of p-values could be combined using this method, and given the number of different tests evaluated in this study using a variety of weighting and filtering methods, the possible combinations of p-values that could be exhaustively combined is enormous. In order to explore this novel approach in a controlled manner, we limited our combined model to two p-values and always chose Fisher's exact test on spectral counts as one of the two inputs. As discussed in Chapter IV, Fisher's exact test was chosen as our "benchmark" test due to its ease of implementation, robustness against missing data, and published track record as a high-performance statistical test for spectral count data.

In order to derive a p-value using Fisher's Method, we first generate an X^2 statistic from our two (or more) p-values:

$$X^2 = -2\sum_{i=1}^{k} \ln(p_i) \qquad (3.2)$$

The X^2 statistic follows a χ^2distribution with 2k degrees of freedom, where k is the number of combined p-values. We can then generate a one-tailed p-value for X^2 given the 2k degrees of freedom using a built-function in R.

For each candidate test, its p-value (for a given peptide group) was paired with the p-value generated by Fisher's exact test. For candidate tests conducted on spectral count data, this resulted in two p-values based on spectral counts being combined using Fisher's Method, which raises an important caveat in the application of Fisher's Method in this context: Fisher's Method

assumes that the combined p-values are independent of each other. This assumption is clearly invalid when both p-values are derived from the same set of spectral counts using different statistical tests. This assumption is also violated, though to a lesser extent, when combining p-values from PPIDs and p-values from spectral counts if the PPIDs are generated only from identified spectra (e.g. PPIDs generated using Hardklör/Bullseye).

If precursor intensity values were derived independently of spectral counts (as is the case for many precursor intensity-based quantitation approaches), the degree to which the two data types would be independent is an intriguing question. For instance, more intense precursors are correlated with higher probability of peptide identification from MS/MS spectra. To partially address the question of independence between spectral count and precursor intensity data, we explored the differences in performance biases between the PPID and spectral count-based statistical tests evaluated in this study in Chapter IV.

Multiple Testing Correction

Due to the large number of tests conducted on each set of replicates, multiple testing correction is needed to minimize false positives. In this study, we applied the Benjamini-Hochberg correction for all statistical tests when using the iPRG *E. coli* training data set. Adjusted p-values of 0.05 were treated as significant (41). The Benjamini-Hochberg correction is designed to maximize the per-comparison error rate (PCER) rather than at controlling the family-wise error rate (FWER). Controlling for the PCER instead of the FWER makes this procedure appropriate for this study because we are applying the same procedure to a variety of different statistical tests and cannot select the optimal test to control each.

The Benjamini-Hochberg correction is an FDR controlling procedure that relaxes the Bonferroni procedure iteratively by setting the following constraint:

For hypothesis tests $H_1 \ldots H_m$, order all the respective p-values from smallest to largest. Find the largest p-value that satisfies the following condition (3.3):

$$p(i) \leq (i/m)q^* \qquad (3.3)$$

Where q^* is the desired FDR rate. Once the condition is satisfied, reject the null-hypothesis for $H_1 \ldots H_i$.

This method was shown by Benjamini and Hochberg to grow in power as the number of tests increased. This was also the case as the number of null hypotheses to be discarded increases (44). These qualities make the Benjamini-Hochberg procedure useful for a data set with a potentially large number of proteins tested (or peptide groups tested) that should truly have the null hypothesis rejected. As is discussed in Chapter IV, the data set "answer key" generated for use in this study contains ~10% (115) peptide groups for which the null hypothesis can be truly rejected, so we are not to be concerned about imposing this procedure with too few truly rejectable null hypotheses.

As discussed in Chapter II, we are conducting our difference testing at the peptide group-level instead of the protein level, which impacts the use of the Benjamini-Hochberg procedure in multiple ways. First, we are increasing the degree of independence between p-values that is clearly not upheld when conducting difference testing at the protein-level using the same

underlying quantitative data (e.g. spectral counts) that is renormalized for each protein. Independence between tests is a requirement of the Benjamini-Hochberg procedure in its attempt to control for the PCER instead of the FWER. A second point regarding the reduction in the number of tests conducted is that the Benjamini-Hochberg procedure's FDR controlling power increases with the number of tests conducted (44), so the benefit of reducing the number of tests is somewhat negated when using the Benjamini-Hochberg procedure to control the FDR.

Missing/Observed PPID Weighting

Not all identified spectra are matched to a PPID, resulting in errors that may arise from missing data. In order to maximize the potential value of Bayesian estimated variance, we investigated the use of weighting peptides based on the number of observed PPIDs or missing PPIDs for each peptide within a given peptide group.

Table 1. The contingency table used to calculate the probabilities of observing the number of matched and unmatched PPIDs across all replicates. Numbers in table assume five replicates per cohort and 32,174 out of 127,500 peptide/replicate pairs have at least one PPID.

	This peptide	All other peptides	Total
Matched PPIDs	2	32,174-2 = 32,172	32,174
Unmatched PPIDs	8	95,326-8= 95,318	95,326
Total	10	127,490	127,500

For each peptide group in a replicate, the reported intensity consists of the sum of the PPIDs for each individual peptide in the peptide group. The first optional filter we tested involved the weighting of PPID intensities for each peptide based on the number of missing or observed PPIDs for each individual peptide seen across all replicates. First, we counted the number of replicates that included a PPID for that peptide vs. the number of replicates with no PPID for that given peptide. If a PPID is observed for a given peptide/replicate pair we assign a

+1 to the "matched" count and assign a +1 to the "unmatched" count if no PPID is seen. Using these "matched" and "unmatched" counts across all replicates, we then calculated the hypergeometric distribution of PPID observations across all replicates for a given peptide (e.g. 0-10 PPIDs assuming 5 replicates per cohort) (Table 1). Next, for each peptide, IPDQuantify computes the p-value associated with observing fewer PPIDs across all replicates based on the hypergeometric distribution defined above. For instance, if a given peptide has three PPIDs observed across all 10 replicates, we calculate its p-value by summing the probabilities in the hypergeometric distribution for observing 0, 1, and 2 replicates (Table 2).

Table 2. The probabilities and corresponding p-values from the hypergeometric distribution of 32,174 "matched" and 95,326 "unmatched" PPIDs

Matched PPIDs	probability	p-value
0	0.105372895	0.000
1	0.265922106	0.105
2	0.301978191	0.371
3	0.203205768	0.673
4	0.089732319	0.876
5	0.0271699	0.966
6	0.005712784	0.993
7	0.000823633	0.999
8	7.79242E-05	1.000
9	4.36868E-06	1.000
10	1.10211E-07	1.000

These p-values are then used to weight the contribution of each peptide group. Thus, when a given peptide group's total PPID intensity is calculated for a given replicate by summing the PPIDs of each of the individual peptide PPIDs, the peptides with only one or two PPIDs observed across all replicates will contribute less than peptides with six or 7 observed PPIDs. This results in stronger weighting of those peptides for which we have observed the most evidence.

We also employed a similar filter using missing PPIDs per peptide. In this case, we counted the number of peptides that have at least one spectral count (in any charged state) that could not be matched to a PPID. This represents an instance where Bullseye was unable to find a PPID for a given identified peptide in a given replicate. The total number of "matched" and

"unmatched" PPIDs across all peptides and all replicates are calculated as above. Using these "unmatched" and "matched" counts we calculate the hypergeometric distribution of unmatched PPID observations for a given peptide (across all replicates) and subsequent p-values for the possibility of observing fewer missing but expected PPIDs across all replicates. We then use (1 - p-value) to weight the contribution of each peptide's PPID when calculating the peptide group PPID.

PPID Count/Spectral Count Filtering

A second category of filter evaluated in this paper involved simply removing from consideration those peptide groups that had less than the threshold value of PPIDs (or spectral counts) observed across all replicates. The "minimum PPID per peptide group" filter counts the total number of PPIDs observed across all replicates (in both cohorts) for a given peptide group and removes that peptide group from further consideration if the total number PPIDs is than the minimum threshold. This filter is applied before subsequent normalization steps (e.g. percentile normalization, see below), thus possibly altering the distribution of normalized values. The second filter we tried is a "minimum PPID in at least one cohort per peptide group" filter. This filter counts the total number of PPIDs observed in both cohorts for a given peptide group. If neither one of the cohorts meets the minimum PPID threshold, that peptide group is removed from further consideration. The minimum PPID filters used in this study were 0, 5, 10, 15, 20, and 25 for both filters. For statistical tests based on spectral counts, we employed peptide group spectral count filters analogous to the PPID filters described above.

Normalization Approaches

The distribution of PPID intensities within a given replicate closely matches the typical exponential decay distribution of spectral intensity data. While the distributions of different replicates tend to be similar, the raw intensity values can vary significantly from replicate to replicate, necessitating the need for normalization. Following peptide group-level PPID calculations, and optional missing PPID weighting and/or filtering, two different normalization approaches were tested: replicate-level percentile normalization, and maximum value per replicate normalization.

When percentile normalization was used, each replicate's peptide group-level intensities are percentile normalized separately. First, non-zero PPIDs are separated from the zero-values. Each non-zero PPID is ranked highest to lowest and assigned a percentile value, with 0 being the most intense and 99 being the least non-zero intensity. Next, the peptide groups which have PPID intensities of 0 but spectral counts greater than 0 are assigned a value of 100. Finally, peptide groups where no PPIDs were observed and none were expected are assigned a value of 101. This has the result of creating a flat, uniform distribution of non-zero PPID values for subsequent analysis. For maximum value normalization, we simply divided each peptide group PPID by the maximum PPID value in its replicate. This results in a distribution between 0 and 1 for subsequent analysis that retains its exponential decay distribution.

PPID Data Transformations

Following normalization and/or minimum PPID filtering, two different transformations were tested: the Tukey-Freeman arcsine transformation (45) and a log transformation. The Tukey-Freeman arcsine transformation is simply the arcsine of the square root of data normalized from 0-1, requiring division by 101 for percentile normalized PPID data. For uniformly distributed data, the Tukey-Freeman transformation results in a roughly Gaussian distribution with values between 0 and $\pi/2$, centered around $\pi/4$. When the arcsine is performed on exponential decay distributions (i.e. when replicates have been normalized by the maximum PPID value/replicate), we observe a modest shifting and broadening of the distribution away from the extremely low values. The second data transformation optionally tested was the log transformation, which has the effect of converting an exponential decay to an approximation of a normal distribution. PPIDs with a value of 0 are assigned a post-log transformed value of 0.

CHAPTER IV

IPRG *E. COLI* TRAINING SET

Introduction

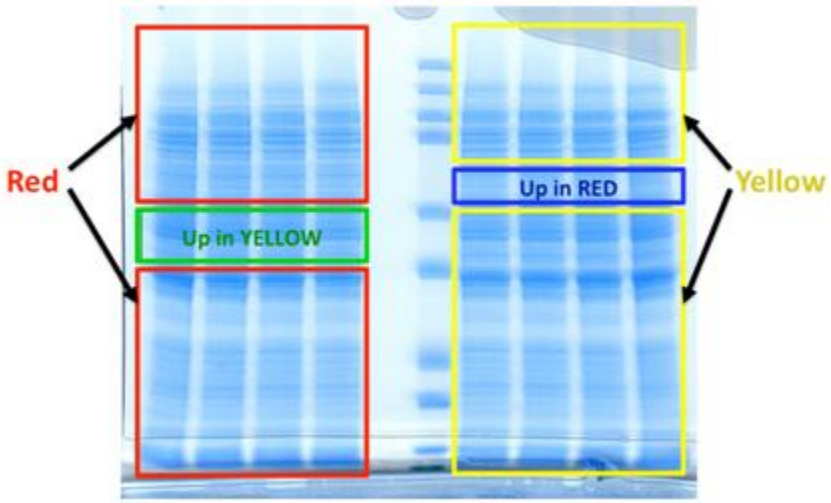
In order to evaluate and compare the performance of the candidate statistical tests and our combined model, we used a pair of LC-MS/MS data sets consisting of two *E. coli* proteomes, each with a different segment of their proteomes removed to create sets of differentially abundant proteins. The data set, the iPRG2009 Study data set, was generated by the Proteome Informatics Research Group iPRG, a part of the Proteomics Standards Research Group (sPRG) under the Association of Biomolecular Research Facilities (ABRF). It was designed to provide researchers with LC-MS/MS data known to contain a significant number of differentially abundant proteins that can be separately verified, yielding an "answer key" for use in difference testing method evaluation.

Materials and Methods

Experimental Data Sources

This data set includes two samples of LC-MS/MS data derived from *E. coli* lysates (labeled "Red" and "Yellow"), with five technical replicates for each sample. In addition, the Red/Yellow samples both had complementary "Blue" and "Green" answer key samples consisting of three LC-MS/MS technical replicates (Figure 4). The "Red" and "Yellow"

replicates were derived from the same *E. coli* lysate sample run on two halves of one gel with a single region excised from each half (The "Blue" and "Green" proteomic data sets). The Red, Yellow, Blue, and Green gels were all analyzed on an LTQ-Orbitrap, and the resulting data sets are freely available at proteomecommons.org.



source: http://www.abrf.org/ResearchGroups/ProteomicsInformaticsResearchGroup/Studies/iPRG2009_presentation.pdf
Figure 4: The "Red/Yellow" iPRG 2009 LC-MS/MS data set with "Blue/Green" LC-MS/MS answer keys. Differential proteomic analysis of the Blue/Green samples yields "true positives" of peptides up or down.

Peptide Identification

Raw spectral files were converted to the mzML format (46, 47) using msConvert (48). Searches were conducted on the 20080728-Sprot-ECOLI-BSA-Cntms-reverse.fasta protein database, which contained forward and reverse sequences for FDR estimations. Searches were configured to allow tryptic and semitryptic peptides and a static mass shift of 57.0125 Da for alkylated cysteines. Optional allowable modifications were oxidation of methionine (+15.996 Da), N-terminal acetylation (+42.013 Da), N-terminal pyroglutamate formation (−17.0265 Da). Search configuration parameters are shown in Supplemental Materials File 1. Peptides were filtered at a 5% FDR using IDPicker.

Results and Discussion

LC-MS/MS Data Set Summary Statistics

The iPRG "Red/Yellow" *E. coli* spectral files were first searched for peptides using MyriMatch (versions 1.3.2) (49) followed by peptide group and protein group assembly using IDPicker (version 2.6.129). Each cohort consisted of five technical replicates. A summary of the search results at the peptide, peptide-group, and protein-group-level of spectral counts and PPID measurements for the iPRG *E. coli* data sets are shown using in Table 3.

Peptide hits were determined using MyriMatch. Measured PPIDs and missing PPIDs were roughly equal between the two cohorts. The "Protein Groups/Peptide groups" ratio was 1.1, indicating that the only ~10% of peptide groups were shared between two or more protein groups. Bullseye (40) was able to find a PPID for 74% of identified spectra.

	Protein groups	**Unique peptides**	**Peptide groups**	**Spectral count**	**PPID measured**	**Missing PPID**	**Protein database**
***E. coli* R/Y**	1,403	9,124	1,275	43,687 (21,987R / 21,700 Y)	32,174 (16,399 R / 15,737 Y)	11,551 (5,588 R / 5,963 Y)	20080728-Sprot-ECOLI-BSA-Cntms-reverse.fasta

Table 3: MS Search Summary for *E. coli* Red/Yellow data set. The cohort labels are: R=Red, Y=Yellow

"Blue/Green" Answer Key

The iPRG *E. coli* "Blue/Green" data set was used as the answer key for which peptides were differentially present in the "Red/Yellow" samples. Peptide groups found to be differentially present in "Blue/Green" samples are also assumed to be differentially present in the

"Red/Yellow" samples. As above, peptides were searched for using MyriMatch (v. 1.3.2) using the same configuration as described above for the "Red/Yellow" search, followed by peptide group and protein group assembly using IDPicker (v 2.6.129). Summary results are shown in Table 4.

	Protein groups	Unique peptides	Peptide groups	Spectral count	Protein database
***E. coli* B/G**	492	3,870	522	14,740 (7,374 B / 7,366 G)	20080728-Sprot-ECOLI-BSA-Cntms-reverse.fasta

Table 4. MS Search Summary for *E. coli* Blue/Green data set. The cohort labels are: B=Blue, G=Green.

The "Blue/Green" data set had significantly fewer peptides, proteins, and spectral hits compared to its "Red/Yellow" counterpart. This was expected because of the fewer technical replicates per cohort (3 vs. 5) and the fact that the excised gel regions used to generate this data contained protein mixtures with lower diversity than the remaining Red/Yellow gels (See figure 4).

"Blue/Green" Answer Key Creation

Following peptide identification and peptide/protein group assembling, the answer key needed to be generated from the data sets. The "Blue" and "Green" data sets each consisted of different, but overlapping, segments of the *E. coli* proteome. A given protein found in the Blue or Green gel bands is presumably up or down in the Red and Yellow gel bands unless the same amount of protein is removed from both the Red and Yellow bands. As such, we defined our answer key as those proteins found to be differentially abundant between the Blue and Green gel

bands using an established, safe difference test. We chose to use spectral counts for Blue/Green difference testing. Spectral count-based methods tend to be better at determining which proteins changed in abundance than precursor intensity-based tests (see Chapter II), and our precursor intensities consist of a novel, untested metric (i.e. PPID intensity values from Bullseye for identified spectra only). With three replicates per cohort, the "Blue/Green" data set is potentially a candidate for non-parametric analogues to Student's t-test, such as the Mann-Whitney U test. But three replicates are the minimum allowable for such tests, making such a choice somewhat less sensitive. Fisher's exact test, on the other hand, combines the data from all replicates for each cohort, making it a safer choice when a few replicates are available.

After selecting the difference test to populate our answer key, we needed to determine whether or not multiple testing corrections were needed when defining the final set of "true positive" differentially present proteins. In this comparative study, protein quantitation is done at the peptide group level (See Chapter III). 522 peptide groups were tested on the "Blue/Green" data across three replicates per cohort. This is clearly a case of many tests on few replicates, necessitating proactive FDR control. On the other hand, the Blue and Green gel bands were excised at different locations in the "Red/Yellow" gels, so we should expect a large of number differentially present peptide groups and we may risk biasing the measured performance of the various statistical tests run on the "Red/Yellow" training data in favor of more conservative tests if we erroneously retain the null hypothesis for too many truly differentially present peptide groups in the "answer key."

Running Fisher's exact test without multiple testing correction (for a significance level of $p=0.05$), results in the identification of 170 out of the 522 peptide groups as significantly different. When the Benjamini-Hochberg correction is used, we lose approximately a third of the

peptide groups deemed significantly different. However, in spite of this correction, we have a list of 115 "true positives," i.e. fewer than 10% of the 1,275 peptide groups found in the "Red/Yellow" data set compared to 13% without correction. Because of the small number of replicates in our Blue/Green data sets, we chose the more conservative approach of using Benjamini-Hochberg correction for the answer key generation. In addition, we observed that many of the peptide groups removed by the Benjamini-Hochberg correction were among the least abundant of the pre-corrected set of "true positives" (TPs). Additionally, the well-established difficulties of spectral count-based tests on low abundance proteins (See Chapter II) suggests that these less abundant peptide groups are more likely to be false positives. All peptide groups in the Red/Yellow data set identified by IDPicker that were not in the set of "true positives" were assigned as "true negatives" in the answer key.

Benchmark Test vs. Classic Statistical Test

Using the "Blue/Green" answer key to evaluate statistical test performances and the "Red/Yellow" *E. coli* data as our data set, we compared Fisher's exact test on spectral counts to the following classical difference tests using either spectral counts or PPIDs: Student's t-test on PPIDs (with and without log transformation), Welch's t-test on PPIDs (with and without log transformation, the Mann-Whitney U test on spectral counts (modified and unmodified), and the Mann-Whitney U test on PPIDs (See Chapter III for details on candidate tests). Fisher's exact test was chosen as our benchmark due to its robustness in the face of noise/missing data and because it has already been evaluated and deemed effective for LC-MS/MS difference testing (50). But it should also have an inherent advantage over the other selected tests because we

used Fisher's exact test to define the "answer key," and it will therefore share the same hypothesis testing biases (e.g. the same family-wise error rate). Because our multiple testing correction procedure, the Benjamini-Hochberg correction (44), does not attempt to control for the family-wise error rate (FWER) but instead focuses on the per-comparison error rate (PCER), this may also give our benchmark test an advantage over the candidate tests if a larger amount of the error for our benchmark test will arise from the PCER compared to our candidate tests.

Due to the large number of tests conducted over the 10 replicates (1,275 peptide group tests), we started off using the Benjamini-Hochberg correction to reduce false positives. Due to a significant loss in sensitivity by both the Mann-Whitney U test and the modified Mann-

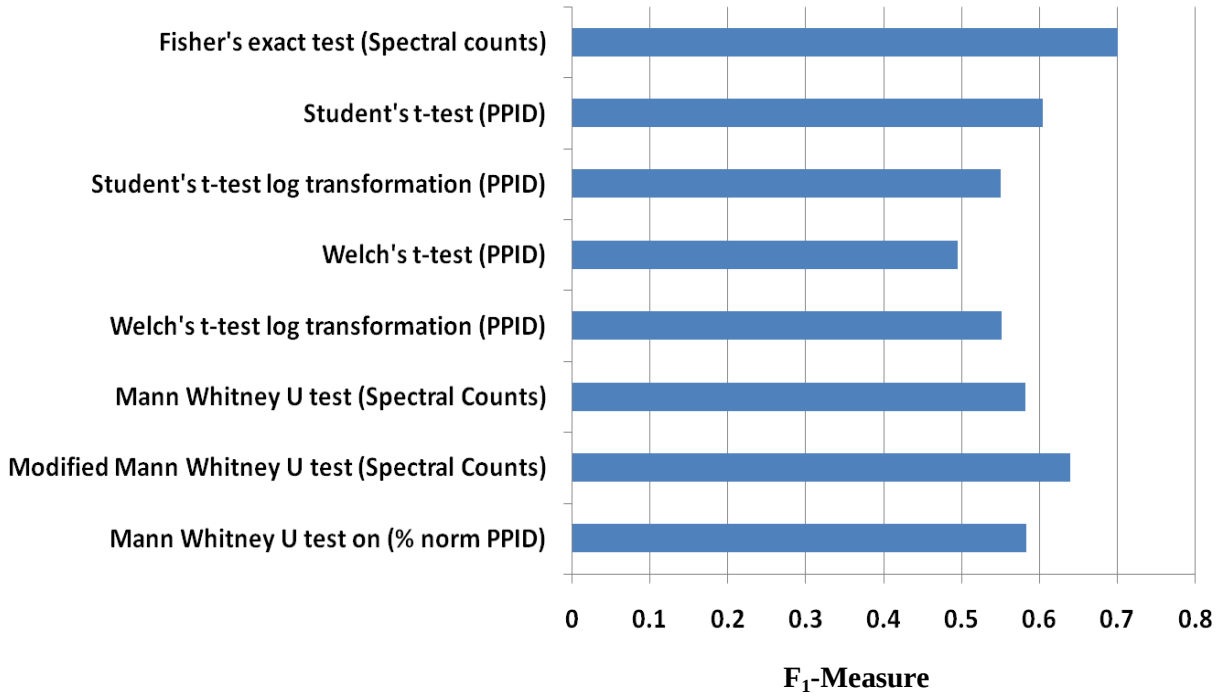


Figure 5. The F_1-measures for the classical difference tests vs. the Benchmark test on the Red/Yellow data sets. Results for Mann Whitney U tests did not include the Benjamini-Hochberg correction. No PPID/Spectral Count filtering or weighting methods were used for the results shown.

Whitney U tests, when used in conjunction with the Benjamini-Hochberg correction, multiple

testing correction was not applied for those tests in future analyses (data not shown). This is not unexpected based on sharply reduced statistical power than the Mann Whitney U test suffers as the number of replicates falls below 10 (51).

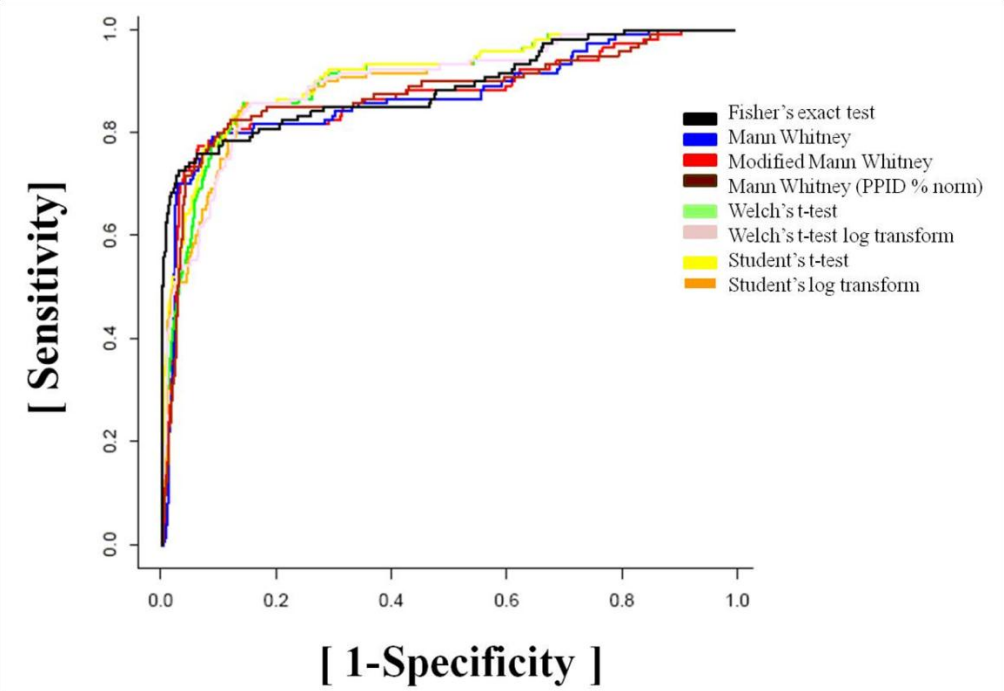


Figure 6. The ROC curves for the classical difference tests vs. the Benchmark test on the Red/Yellow data sets. Results for Mann Whitney U tests did not include the Benjamini-Hochberg correction. No PPID/Spectral Count filtering or weighting methods were used for the results shown. Our benchmark test, Fisher's exact test (Black), showed the greatest specificity but reduced sensitivity if higher false discovery rates were accepted.

The ROC curves and F_1-measure for each classical test were initially calculated using the answer key without any PPID/Spectral count filtering or weighting (see Chapter III). Results are shown in Figures 5 and 6. The ROC curve of Fisher's exact test clearly outperformed all other tests in terms of accuracy. The PPID-based tests (Student's t-test and Welch's t-test) showed the lowest accuracy, but the highest potential sensitivity if more FPs are to be tolerated. The F_1-

measure, which equally balances the value of a test's precision and recall, showed a clear advantage to using Fisher's exact test on spectral counts based compared to the other tests.

Next, we examined how PPID and spectral count filtering and weighting impacted the performance of these tests. Each test was run separately using a "minimum PPID/Spectral Count in at least one cohort" and a "minimum PPID/Spectral Count across all cohorts" filter. Filter values of 5, 10, 15, 20, and 25 were used for both filters applied. F_1-measures and AUC values are available in the Supplemental Materials Files 3 and 4. A PPID or spectral count filter of 5 slightly improved AUC values, but filter cutoffs above 5 did not help and began to hurt performance as the cutoff was increased. Similarly, F_1-measures improved with "minimum PPID/Spectral Count in at least one cohort" filter thresholds set at 5 or 10 PPID/Spectral counts before showing a negative impact with higher thresholds. The "minimum PPID/Spectral Count across all cohorts" filter did not appear to help for these tests. For the tests run on PPIDs, neither weighting method tried ("observed PPID" or "missing PPID" weighting) appeared to have consistently positive or negative impact.

Benchmark Test vs. Empirical Bayes t-test

Next, Fisher's exact test (on spectral counts) was used as a benchmark for comparing different empirical Bayes t-test methodologies. The empirical Bayes t-test was run using either PPIDs or spectral counts (See Chapter III). Results for the un-weighted/unfiltered tests are shown in Figures 7 and 8.

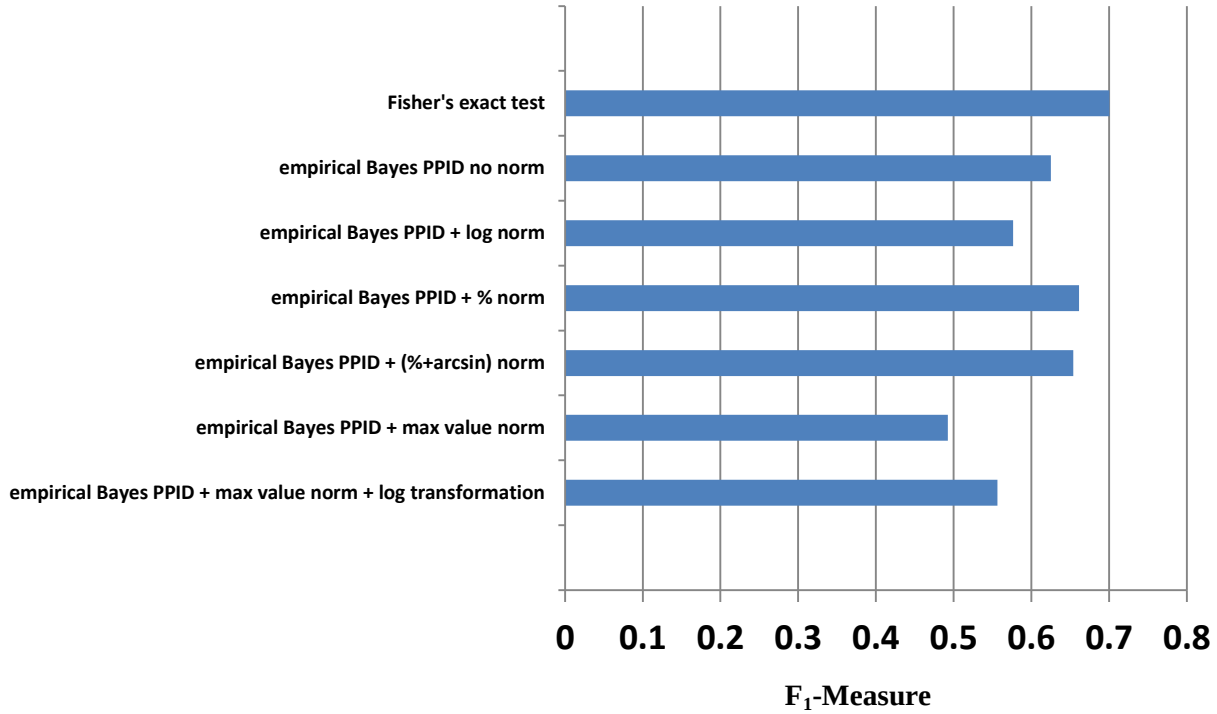


Figure 7. The F1-measures for the empirical Bayes t-tests vs. the Benchmark test (Fisher's exact test on spectral counts) on the Red/Yellow data sets.

Fisher's exact test outperformed the empirical Bayes t-tests when either F_1-measures or ROC analysis was used. As with the classical test comparisons, the ROC plots showed a much higher accuracy for the benchmark test, but a greater potential sensitivity amongst the empirical Bayes t-tests if more false positives are to be tolerated. The top performing tests based on the F_1-measure were: a) the empirical Bayes t-test using un-normalized spectral counts and b) the empirical Bayes t-test on PPIDs with percentile normalization and the empirical Bayes t-test on PPIDs with the arcsine(sqrt) transformation in addition to percentile normalization. When spectral count/PPID filtering was applied, only the minimum PPID/spectral count filtering appeared to have any sort of consistent positive impact, with a mild improvement when the "minimum PPID/Spectral Count in at least one cohort" filter is set to 5 or 10, and otherwise the impact was detrimental to performance (See AUC/F1-Measure tables in Supplemental Materials

Files 3 and 4). Neither PPID weighting methods showed a consistent positive or negative impact.

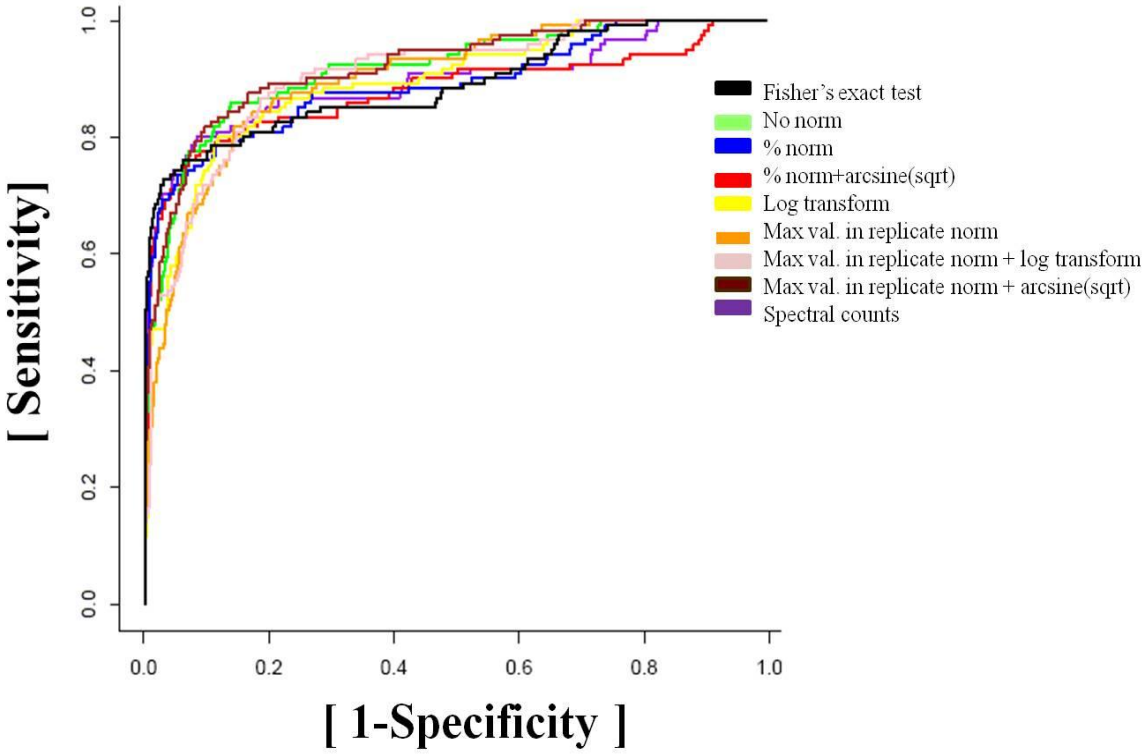


Figure 8. The ROC curves for the empirical Bayes t-test vs. the Benchmark test (Fisher's exact test on spectral counts) on the Red/Yellow data sets. All empirical Bayes tests were run using PPIDs, except the test coded in purple which was run using un-normalized spectral counts. Our benchmark test, Fisher's exact test (Black), showed the greatest specificity but reduced sensitivity if higher false discovery rates were accepted.

Benchmark Test vs. Combined Model

Our combined model that incorporates information from both precursor intensity and spectral count information for improved quantitation consists of using Fisher's Method to combine the p-values of a single test evaluated in this study (classical or empirical Bayes) with

the p-values of the benchmark test (Fisher's exact test on spectral counts at the peptide group level). This model was applied for all tests evaluated in this study. While the intent of this study was to combine precursor and spectral count data into a single statistical test for LC-MS/MS quantitation, we included candidate tests using spectral counts so PPID/Spectral Count and Spectral Count/Spectral Count combinations were examined. The combination of two tests on spectral counts is a reminder that independence is being violated to some extent in the application of Fisher's Method, although, as will be discussed below, the PPID and spectral count-based approaches appear to select for overlapping, but distinct sets of peptides.

Benchmark Test vs. Combined Model Using Classical Tests

The combined model on the classical tests showed a sharp rise in the F_1-measures that virtually matched that of the benchmark test in all cases (Figure 9). The ROC curves clustered around the benchmark test's ROC (Figure 10). The Student's t-test and Welch's t-test compared best to the benchmark test, both showing a higher sensitivity than the benchmark if more false positives are accepted. Table 5 shows the F_1-Measures of the candidate tests before and after the application of the combined model. In general, the worst performing tests showed the largest improvement. For example, the Welch's t-test on PPIDs, which performed the worst alone, showed the greatest rise after the combined model was applied (47%).

Table 5. Percentage Change in F_1-Measures of Classical Candidate Tests Following Application of Combined Model

Candidate Test	Candidate Test Alone	Combined Model (Candidate Test + Benchmark)	% Change
Student's t-test on PPIDs	0.60	0.71	18%
Student's t-test log transformation on PPIDs	0.55	0.70	27%
Welch's t-test on PPIDs	0.49	0.72	47%
Welch's t-test log transformation on PPIDs	0.55	0.70	27%
Mann Whitney U test on Spectral Counts	0.58	0.70	21%
Modified Mann Whitney U test on Spectral Counts	0.64	0.71	11%
Mann Whitney U test on PPIDs	0.58	0.72	24%

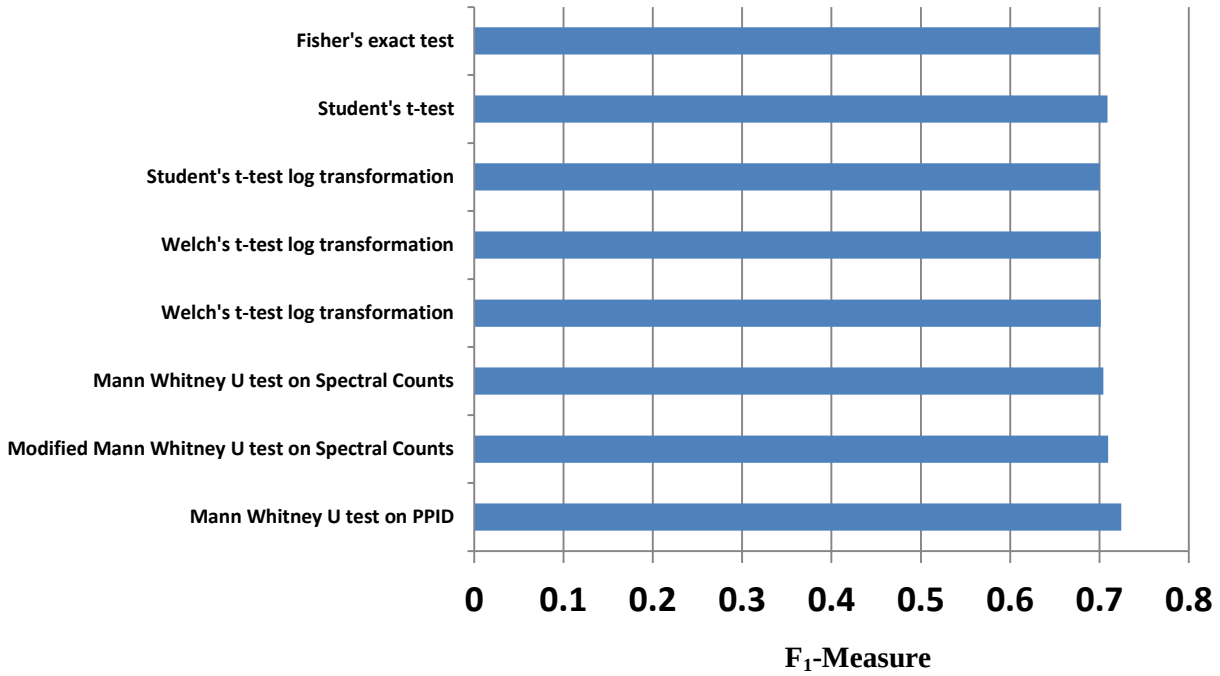


Figure 9. The F_1-measures for the combined model using the classical tests combined with the Benchmark test (Fisher's exact test on spectral counts) on the Red/Yellow data sets.

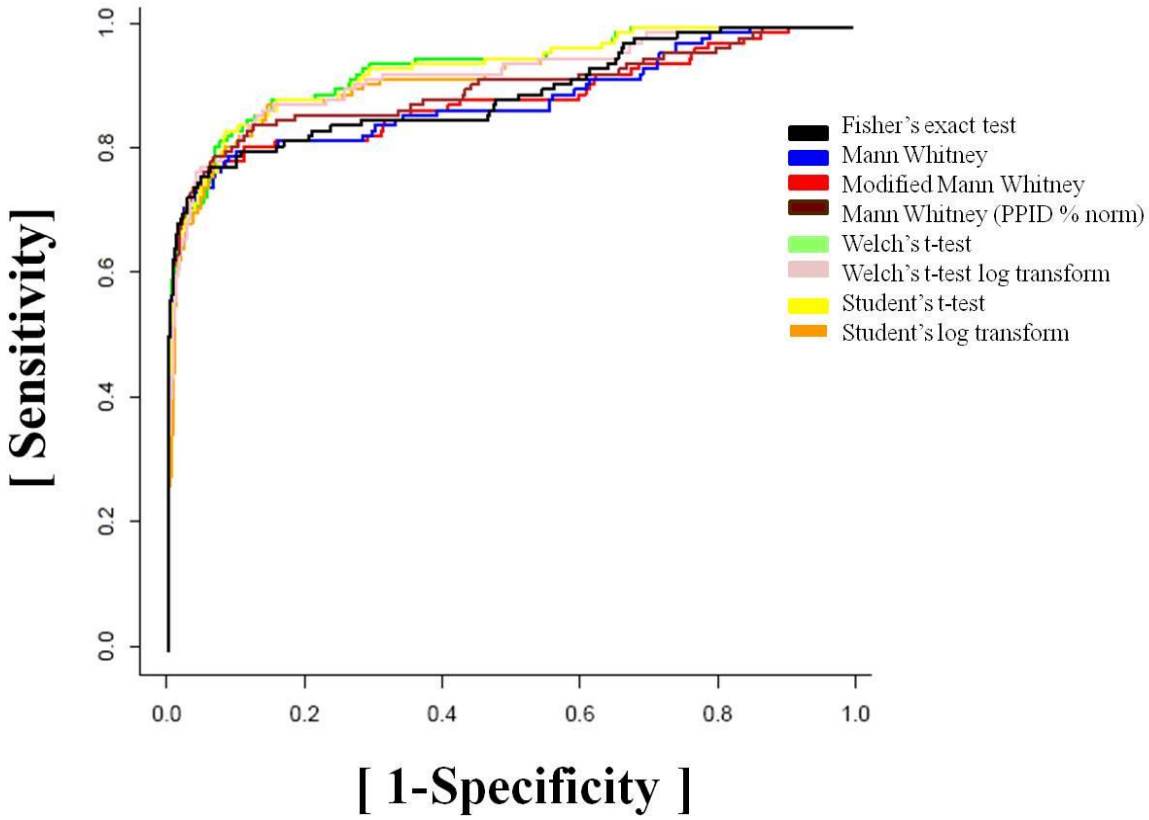


Figure 10. The ROC Curves for the classical difference test methods after combination with benchmark test (Fisher's exact test with spectral counts). The ROC curve for our benchmark test did not have the combined model applied. Our benchmark test, Fisher's exact test (Black), showed the greatest specificity but reduced sensitivity if higher false discovery rates were accepted.

Benchmark Test vs. Combined Model Using Empirical Bayes t-tests

The combined model on the empirical Bayes t-tests showed a sharp rise in the F_1-measures that matched or surpassed that of the benchmark test in all but two cases (Figure 11). The ROC curves clustered around the benchmark test's ROC (Figure 12). Similar to the combined model on the classical test, the combined model on the empirical Bayes ROC curves indicated a mild tradeoff of more false positives for fewer false negatives relative to the benchmark test. Table 6 shows the percentage change in F_1-Measures following application of the combined model for example tests (no weighting or filtering applied). Like the classical

tests, the empirical Bayes t-tests with the worst initial F_1-Measures (the tests employing "max value normalization") showed the greatest percentage increase.

Table 6. Percentage Change in F_1-Measures of Empirical Bayes t-test Candidate Tests Following Application of Combined Model

Candidate Test	Candidate Test Alone	Combined Model (Candidate Test + Benchmark)	% Change
empirical Bayes on PPIDs	0.63	0.73	16%
empirical Bayes on PPIDs + log normalization	0.58	0.7	21%
empirical Bayes on PPIDs + percentile normalization	0.66	0.74	12%
empirical Bayes on PPIDs + percentile+arcsine(sqrt) normalization	0.65	0.72	11%
empirical Bayes on PPIDs + max value normalization	0.49	0.61	24%
empirical Bayes on PPIDs + max value normalization + log transformation	0.56	0.73	30%
empirical Bayes on PPIDs + max value normalization+arcsine(sqrt)	0.5	0.6	20%
empirical Bayes on Spectral Counts	0.67	0.73	9%

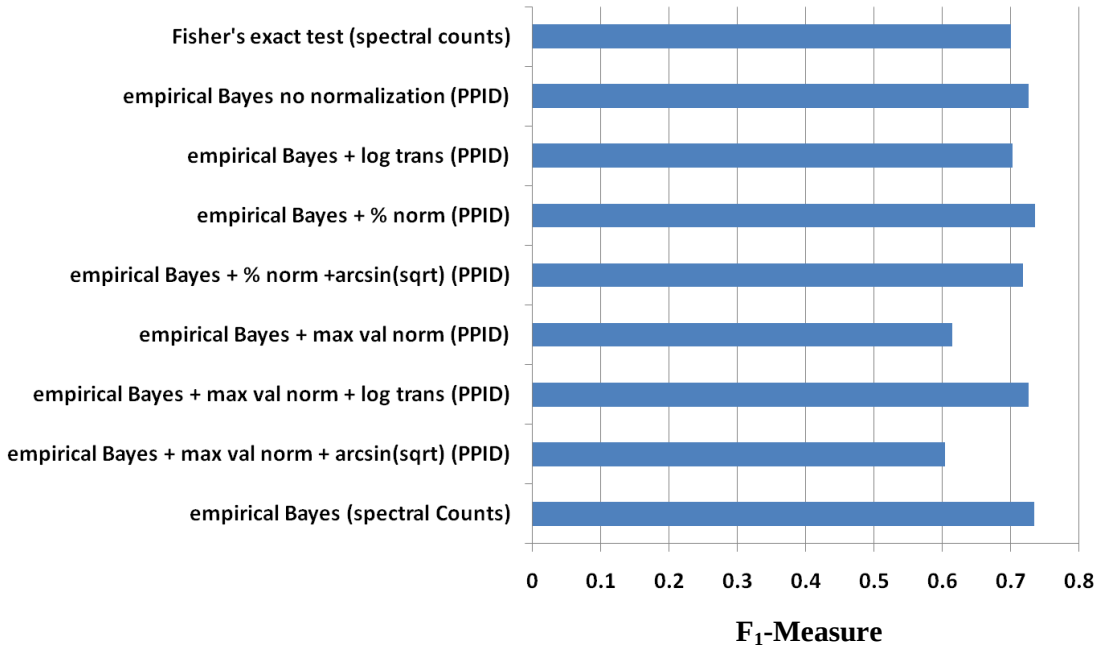


Figure 11. The F1-measures for the empirical Bayes difference test methods after combination with Fisher's exact test

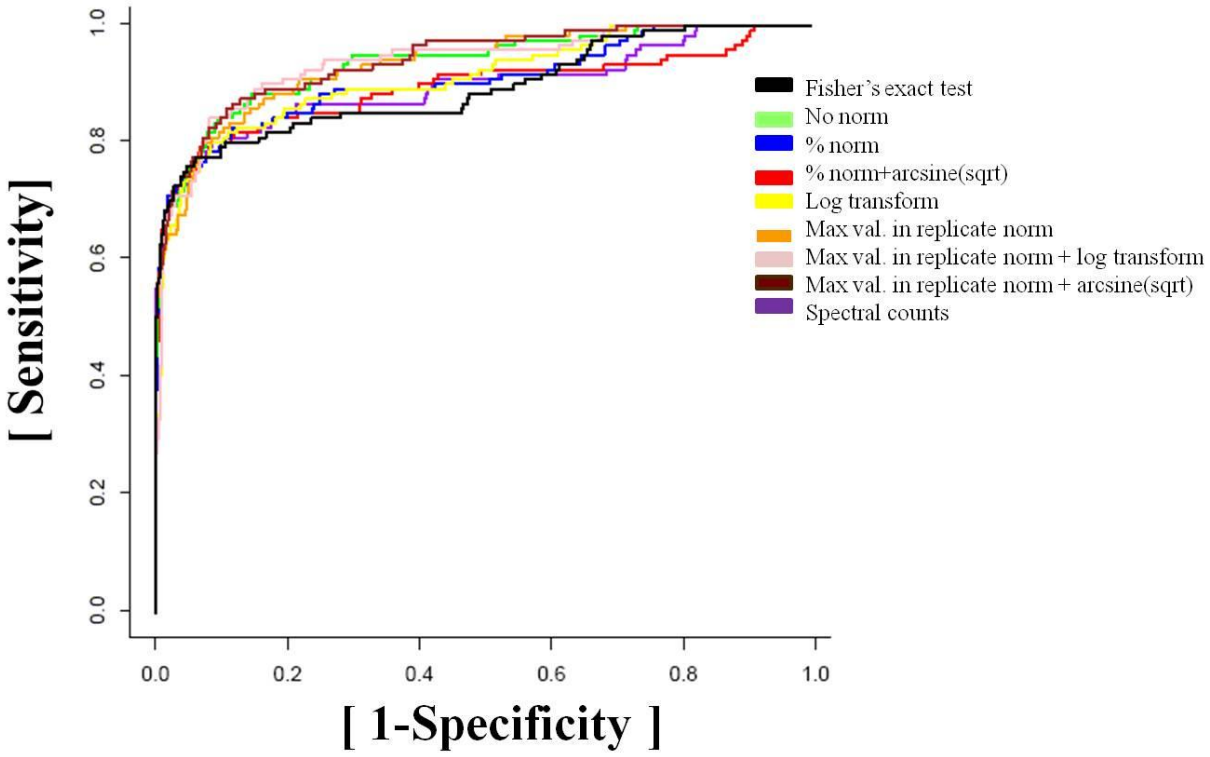


Figure 12. The AUC-measures for the combined model using the empirical Bayes t-test combined with the benchmark test (Fisher's exact test on spectral counts) test methods. The ROC curve for our benchmark test did not have the combined model applied. The benchmark test (Black) showed the specificity, with reduced sensitivity if higher false discovery rates were allowed.

Venn Analysis

The improved performance for virtually every test using Fisher's Method to combine Fisher's exact test with the candidate test can be partially explained by examining the overlap and differences in the difference test's ability to find true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN). In order to analyze the similarity in performance and discriminatory bias amongst the statistical tests examined in this study, Venn charts were created comparing the sets of TP, FP results for the benchmark test (Fisher's exact test on

spectral counts). The FN and TN plots can be inferred from the TP and FP plots and are not shown.

Venn Analysis for Classical tests vs. Benchmark Test

Venn charts of TPs and FPs generated by the classical tests vs. the benchmark test are shown in Figure 13. The Student's t-test had a strong overlap with the benchmark test, with each test sharing the same group of 60 TPs and only 11 and 9 unique TPs found by each test, respectively. The Welch's t-test, which performed the worst amongst the classical tests evaluated, found only 46 TPs overall (vs. 69 for the benchmark) and only 6 of those TPs were not also found by the benchmark test.

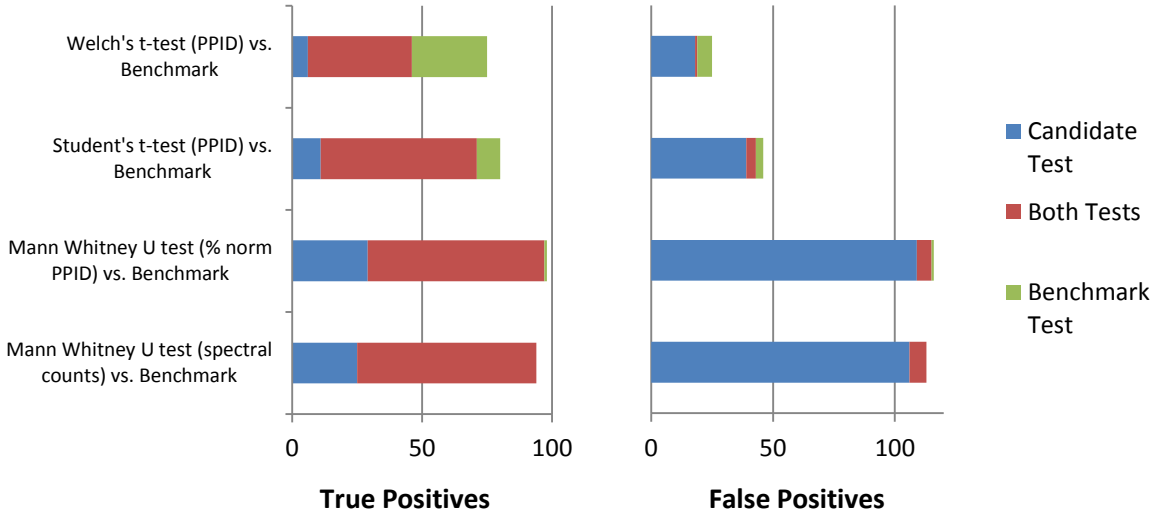


Figure 13. Venn charts of the true positives and false positives generated by the classical candidate tests vs. the benchmark test. A total of 115 true positives existed in the blue/green "answer key". The Mann Whitney U tests had no Benjamini-Hochberg procedure applied.

The Mann Whitney U tests (on spectral counts or PPIDs) found significantly more TPs than the benchmark test, including virtually all of the same TPs found by the benchmark test.

The high sensitivity of the Mann Whitney U tests was not unexpected. Likewise, the Mann Whitney U tests showed 10-fold (or greater) more FPs compared to the benchmark test. This was not unexpected as the Mann Whitney U tests were the only candidate tests to have the Benjamini-Hochberg correction removed in this study, highlighting the tradeoff between sensitivity and accuracy that multiple testing corrections offer.

The Student's t-test and Welch's t-test had ~6 and 3-fold more FPs respectively than the benchmark test (43 vs. 7 and 19 vs. 7, respectively). Interestingly, while the benchmark test had a very small number of false positives (7), the Student's t-test and Welch's t-test shared only 4 and 3 FPs, respectively, of the 7 FPs found by the benchmark test.

Venn Charts for Empirical Bayes t-tests vs. Benchmark Test

Example Venn charts of TPs and FPs generated by the empirical Bayes t-tests vs. the benchmark test are shown in Figure 14. Overall, the empirical Bayes t-test appeared to find most or all of the same TPs found by the benchmark test with 12-18 TPs not found by the benchmark test. Using the log transformation on PPID data worsened performance relative to un-normalized PPIDs, percentile normalized PPIDs, or even un-normalized spectral counts. This is in keeping with the relative performance of the classical tests, where the log transformation of PPID data for both the Student's t-test and Welch's t-test underperformed tests conducted on the raw PPID data (based on F_1-measure and ROC analysis). The negative impact of the log transformation on these candidate tests suggest the need for variance analysis on PPID data grouped at the peptide group level in order to explore the distribution and the most appropriate data normalization and transformations. The lower sensitivity of the empirical Bayes t-test on log transformed PPIDs

was coupled with an overall lower level of FPs relative to the other empirical Bayes t-tests, similar to the Welch's t-test.

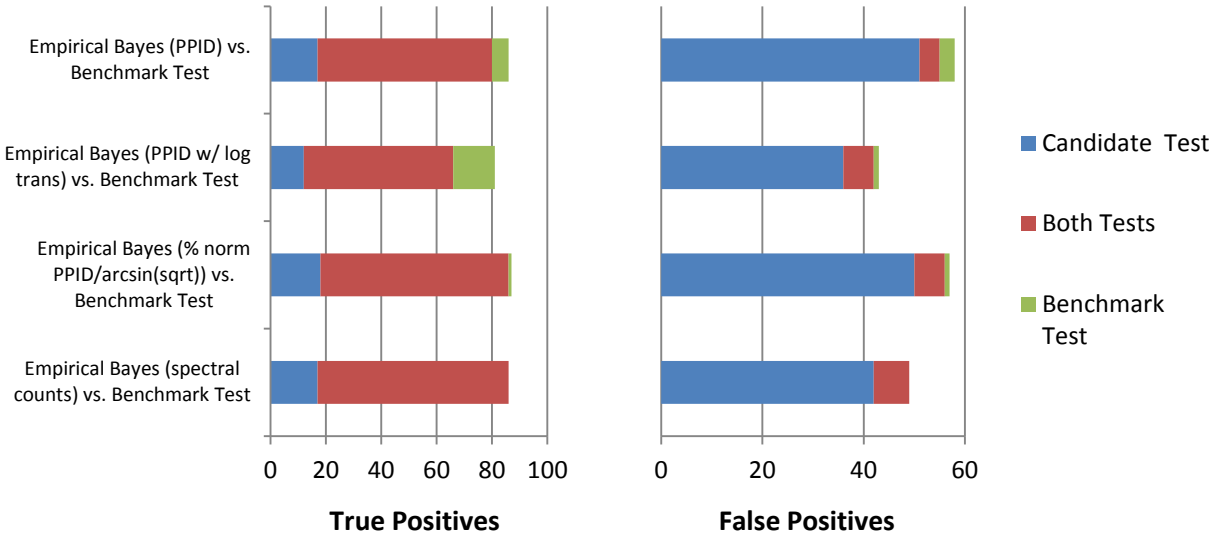


Figure 14. Venn charts of the true positives and false positives generated by the empirical Bayes candidate tests vs. the benchmark test. A total of 115 true positives existed in the blue/green "answer key"

It is worth noting that the empirical Bayes t-test on spectral counts appeared to perform much more similarly to its PPID counter-parts than the benchmark test on spectral counts. Likewise, the Mann Whitney U test on spectral counts performed much more similarly to the Mann Whitney U test on PPIDs than to Fisher's exact test on spectral counts. This suggests that the degree of independence between the performances of different candidate tests depend both on difference in the data type chosen (e.g. spectral counts vs. PPIDs) and also the inherent systematic biases specific to each test.

Venn Charts for Combined Model vs. Benchmark Test

Next, Venn charts of TPs and FPs generated by the combined model vs. the benchmark were analyzed. Figure 15 shows the impact of using the combined model with each of the example classical tests. In every case, both the TPs and FPs generated by the combined model showed a strong shift towards the TPs and FPs of the benchmark test. The combined models on the Mann Whitney U tests (with either spectral or PPIDs) showed a mild drop in TPs coupled with an even sharper drop in FPs relative to the Mann Whitney U tests alone. The combined model using Student's t-test on PPIDs showed a slight rise in TPs and a significant drop in FPs. Welch's t-test, which had fewer TPs than the benchmark test, showed the greatest increase in TPs (46 to 68) while also seeing a drop in FPs when used in the combined model from 19 to 6 FPs. This was the only instance where the combined model had fewer FPs than the benchmark test alone (which had 7 FPs).

Figure 15. Venn charts of true positives and false positives. For both true positives and false positives, the Venn Chart of the candidate test vs. the benchmark followed by a Venn chart comparing the benchmark test with the combine model (candidate test + benchmark test).

Figure 16 shows the impact of using the combined model with each of the example empirical Bayes tests. As with the classical tests, the combined model shifted both TPs and FPs strongly towards that of the benchmark test, resulting in a moderate drop in TPs and a much sharper drop in FPs. The impact of the combined model on TPs and FPs was similar whether PPIDs or spectral counts were used for the empirical Bayes test.

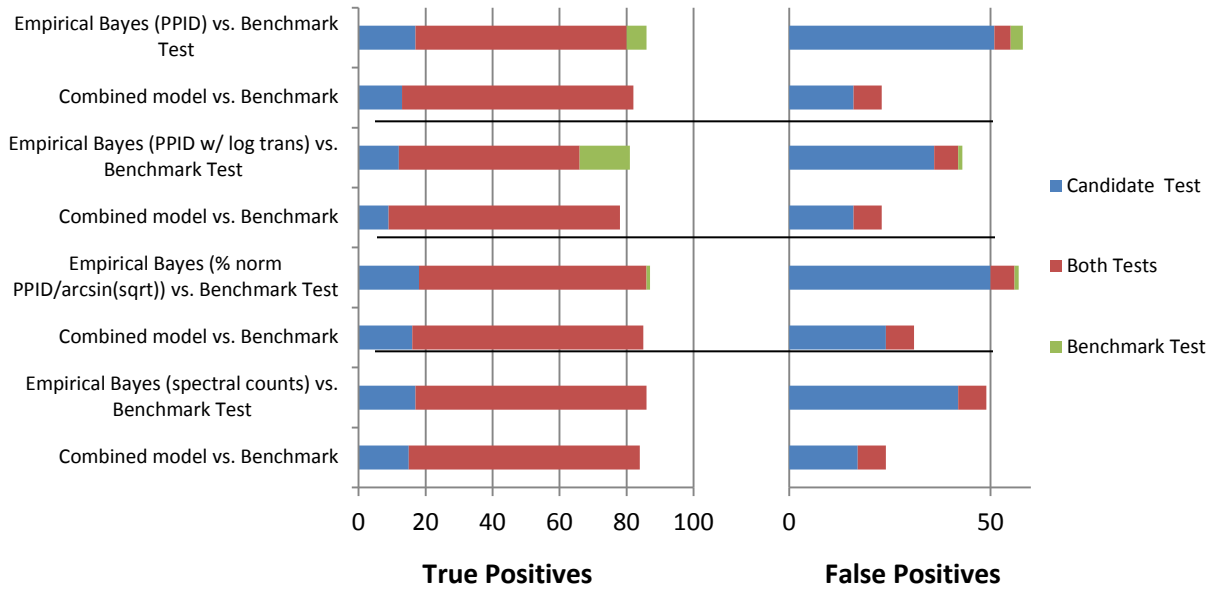


Figure 16. Venn charts of the true positives and false positives generated by the classical candidate tests vs. the benchmark test. A total of 115 true positives existed in the blue/green "answer key"

Venn Chart Analysis Summary

Overall, the Venn chart analysis was consistent with the F_1-measure and ROC analysis in terms of overall test performance while highlighting an apparent partial orthogonality between the behavior candidate tests and the benchmark test. On the basis of TPs, FPs, TNs and FNs, the benchmark test (Fisher's exact test on spectral counts) demonstrated a generally reduced

sensitivity from our candidate tests, but with a dramatically higher precision (e.g. 7 FPs for the benchmark test vs. 55 for empirical Bayes t-test vs. 115 maximum FPs seen for the Mann Whitney U test on PPID). The benchmark also showed a moderate tendency to select for different TPs than the candidate tests, although the most sensitive (and least accurate) tests tended to identify all of the same TPs as the benchmark test.

The Venn charts of the combined model demonstrated an ability to dramatically reduce the number of FPs at a minimal cost to sensitivity when the combined model is used to combine the candidate test with our highly precise benchmark test. In every case, the combined model had more TPs than the benchmark test alone (ranging from 4 -19 more TPs than the benchmark). The combined model still had more FPs than the benchmark test (which only had 7 FPs) in all but one case (when combined with Welch's t-test). Thus, the extra TPs acquired vs. the benchmark test by using the combined model still came at the cost of reduced precision relative to the benchmark test alone.

The single case where the combined model had fewer FPs than the benchmark test alone helps highlight the importance of selecting tests with different performance biases when using a meta-analysis procedure like Fisher's Method. Welch's t-test on PPIDs had the lowest sensitivity, with only 46 TPs, 40 of which were shared with the benchmark test. Of the 19 FPs found by Welch's t-test, only one was shared with the benchmark test, leaving 6 FPs found by the benchmark test not found by Welch's t-test. Overall, Welch's t-test appeared to have the least overlap with the benchmark test amongst all the candidate tests used in the Venn chart analysis. When the two tests are combined using Fisher's Method, the combined model yielded 72 TPs and only 6 FPs. That we were able to obtain more TPs and fewer FPs by combining

them using Fisher's Method than achieved with either test alone indicates the value of using tests with different performance biases when using Fisher's Method.

Available techniques for meta-analysis, including Fisher's Method, presume that the different classes of tests being combined are independent of each other. If changes in protein (or peptide group) abundance is the topic being studied using LC-MS/MS, this would require that the two tests combined using Fisher's Method are independent of each other. This raises the question of the degree of independence between the different statistical tests combined in this study using Fisher's Method. As discussed in Chapter II, published studies comparing the use of spectral counts with precursor intensities for LC-MS/MS quantitation observed that both spectral counts and precursor intensities were potentially useful but they have different strengths and weaknesses (e.g. spectral counts are better for asking which proteins changed in abundance, precursor intensities are better at asking how much they changed in abundance). Both methods suffered from erroneous abundance ratios for low-intensity proteins, albeit in different directions (spectral counts tended to overestimate abundance ratios while precursor intensities tended to overestimate abundance ratios).

Based on those observations, spectral count-based tests and precursor ion-based tests could be considered at least somewhat independent of each other although the degree of independence remains ambiguous. In this study, the precursor intensity data used (PPIDs) is directly dependent on peptide identifications, so our precursor intensity values are clearly less independent of the spectral count data than if we had relied exclusively on LC-MS data to conduct precursor intensity quantitation (as is done with tools like Corra (33)).

In some of the cases, our combined model used p-values from two tests using spectral counts (e.g. the Mann Whitney U test and empirical Bayes t-test on spectral counts), and yet we

still observed a distinct difference between the performance of those tests and the benchmark test. This suggests that some degree of independence between the tests used in our combined models may also arise from different biases in the statistical tests. For instance, robustness in the face of missing data will vary between tests. How the power of statistical tests changes as the number of replicates grows is another factor that will contribute to differences in systemic test biases. All of the candidate tests in this study were parametric or non-parametric variants of the t-test and would therefore suffer much more from missing data points and limited replicates than Fisher's exact test which consolidates all cohort replicates into a single sum. This could be particularly troublesome for the PPID-based tests in our study that would experience significantly more missing data points since ~25% of spectral counts were unable to be matched to a correlated PPID.

Independence Between Candidate Tests and the Benchmark Test

In order to further explore the independence and different biases of the candidate tests with our benchmark test, we created $\log_2$-scale plots of the ratio of spectral counts vs. the ratio of percentile normalized PPIDs for the Red and Yellow cohorts for each peptide group tested (counts and PPIDs summed for each cohort for ratio calculations). Separate plots were created for each candidate test, highlighting the TP and FP peptide groups found using that test. The TPs and FPs of the benchmark test are also highlighted, where the red circles indicate a TP/FP peptide group found only using the candidate test, a blue circle indicates a TP/FP peptide group found only using the benchmark test, and a purple circle shows peptide groups found using both the candidate and benchmark tests.

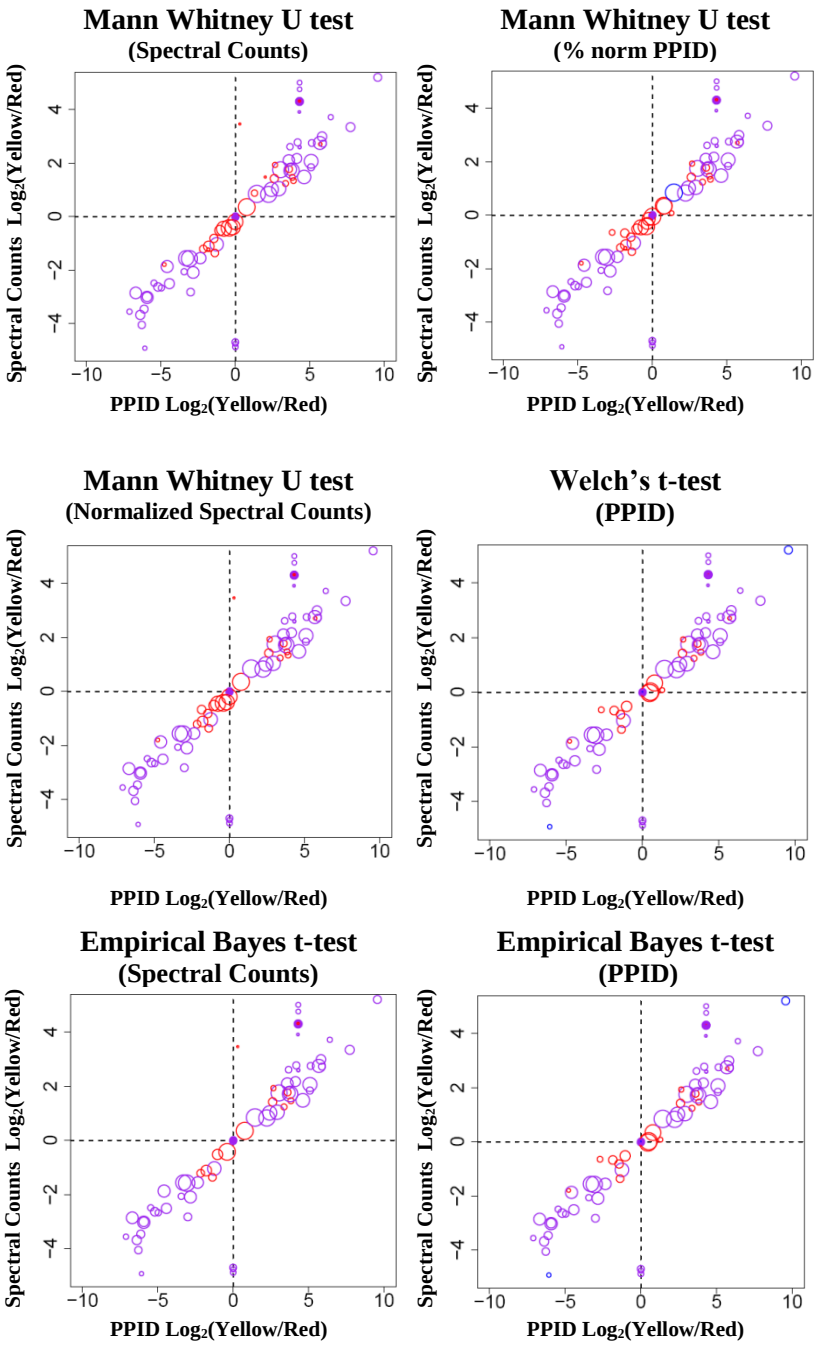


Figure 17. TPs for example candidate tests. For each peptide group, the $\log_2$ scaled ratios of the total spectral count in Red vs. Yellow cohorts (Y-axis) and the ratio of the total percentile normalized PPID values in Red vs. Yellow cohorts (X-axis) was plotted. Peptide groups accurately identified as differentially abundant (true positives) for both the candidate test and the benchmark test (Fisher's exact test on spectral counts) are color coded as follows: Red = TP found only by the candidate test. Blue = TP found only by the benchmark test. Purple = TP found by both tests. Relative sizes of highlighted peptide groups correspond to the average percentile normalized PPID values.

Example plots are shown in Figure 17 for Welch's t-test, the Mann Whitney U test (using percentile normalized PPIDs, unnormalized spectral counts, and normalized spectral counts), and

the empirical Bayes t-test (using PPIDs and spectral counts). All tests in the example plots below had no PPID weighting or PPID/spectral filtering applied. The plots are somewhat asymmetric due to our handling of instances when the Red cohort values (the denominator) were 0 vs. Yellow cohort (numerator) values of 0. We chose a default ratio of 20 when the Red cohort was zero,

corresponding to a $\log_2(20)$ ratio of ~4.3 (on either axis). A value of zero (on either axis) indicates the Yellow cohort had a value of zero (spectral counts or PPIDs).

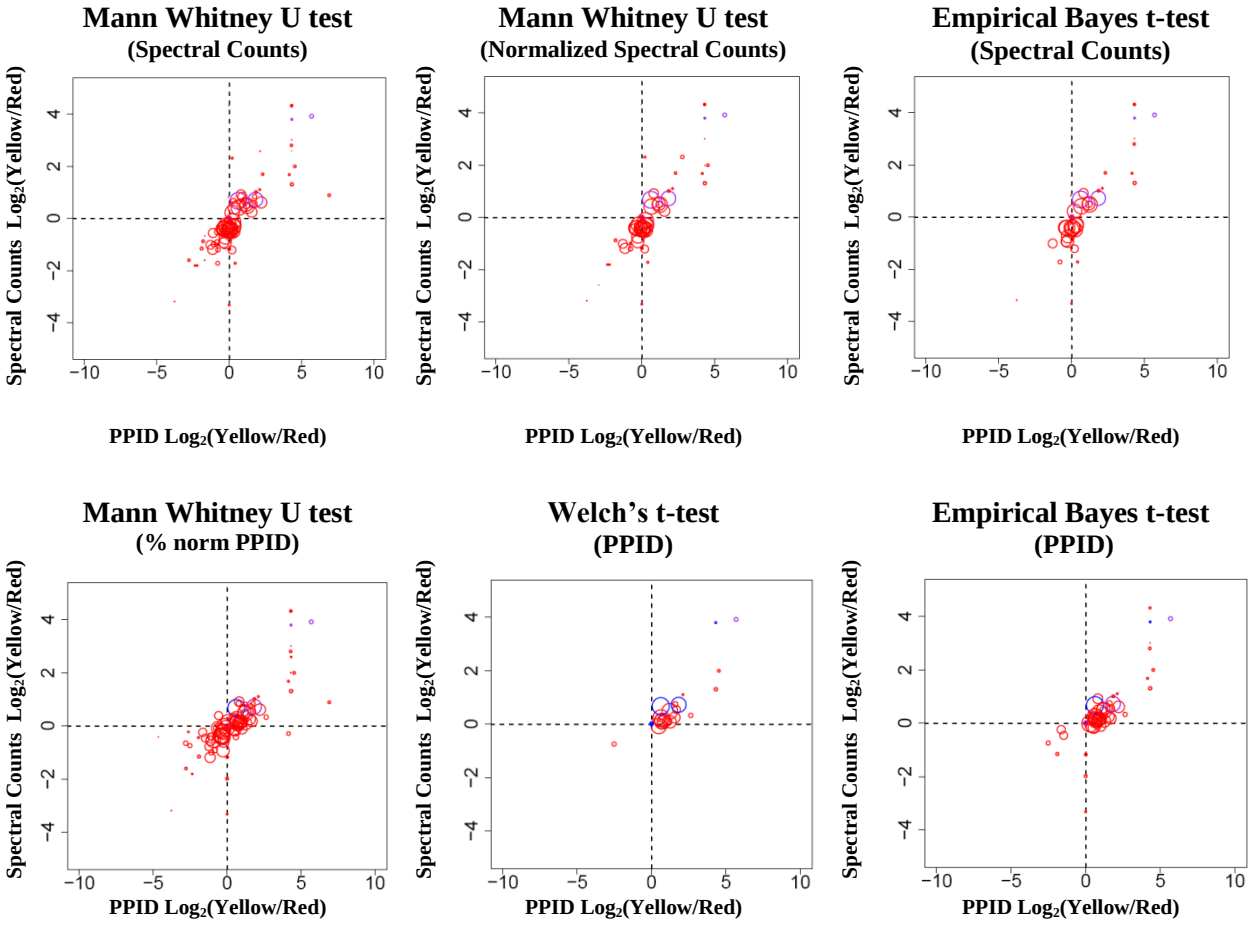


Figure 18. FPs for example candidate tests. For each peptide group FP, the $\log_2$ scaled ratios of the total spectral count in Red vs. Yellow cohorts (Y-axis) and the ratio of the total percentile normalized PPID values in Red vs. Yellow cohorts (X-axis) was plotted. Peptide groups inaccurately identified as differentially abundant (false positives) for both the candidate test and the benchmark test (Fisher's exact test on spectral counts) are color coded as follows: Red = FP found only by the candidate test. Blue = FP found only by the benchmark test. Purple = FP found by both tests. Relative sizes of highlighted peptide groups correspond to the average percentile normalized PPID.

The distribution of the TPs tended fell into two general categories: 1. Peptide groups that had a value of zero in either the Red or Yellow cohorts (those points at the [0,0] or laying at [0, $-\log_2(20)$] or [$\log_2(20)$, $\log_2(20)$]), and 2. Peptide groups with a non-zero value in both cohorts but a relatively strong difference between cohorts (i.e. peptide groups not clustered around the

center). Because the set was generated by excising different gel bands from two *E. coli* proteomes, we would expect many of the proteins in the answer key to have sharply different abundances. In addition, the construction of our "answer key" using the "Blue/Green" data set used a fairly conservative test (Fisher's exact test using the Benjamini-Hochberg correction).

The FPs (Figure 18) also tended to fall into two general categories: 1. Low abundance peptide groups with a 0 value for either the Red or Yellow, and 2. High abundant peptide groups with abundance ratios close to 1 (close to 0 in $\log_2$ scale). It is not surprising that we would find false positives from low abundant peptides with a cohort containing a 0 value from spectral counts given the well-established issues spectral counts face as counts approach zero. In our study, the PPIDs will have additional zero-boundary complications due to the fact that IDPQuantify searches for PPIDs based on spectral counts and could not find a PPID ~25% of the time.

The FPs from moderate to high abundance peptide groups with $\log_2$ abundance ratios close to 0 seem to represent a qualitatively different type of FP than the low abundance peptide groups. Peptide groups with $\log_2$ abundance ratios close to 0 are likely to be at equal or near equal abundance. Such peptide groups would be expected to yield FPs because they lie at the boundary separating the "no difference" peptide groups from the "mildly different" peptide groups.

Interestingly, the FPs exhibited a consistent pattern between the spectral count-based tests and the PPID-based tests. The PPID-based tests tended to show FPs with relatively higher PPID ratios than spectral count ratios (Figure 18, bottom row), whereas the spectral count-based tests tended to show the opposite clustering with FPs that had somewhat higher spectral count ratios than PPID ratios (Figure 18 top row). This pattern was apparent when comparing the same test

run on either PPIDs or spectral counts (e.g. the empirical Bayes t-test and the Mann Whitney U test), indicating that different biases in the candidate tests were not the sole cause of this pattern. This observation suggests a partial degree of independence between the information provided by spectral counts and precursor intensities.

It is worth noting that sets of TPs and FPs both had a large number of peptide groups with zero spectral counts or PPIDs in either the Red or Yellow cohorts, but the TPs with a zero-cohort tended to be moderately more abundant than the FPs (larger circles in the TP plots at the center or ~4.3 on either axis vs. tiny circles in the FP plots). This is consistent with the observation above that minimum PPID/spectral filtering tended to mildly improve test performance when a filter of 5 or 10 PPID/spectral counts was applied but higher filters began to detrimentally impact the candidate tests' performance.

Evaluation of Peptide Group-Level Difference Testing

For Specific Aim III, we sought to establish the validity of difference testing at the peptide group-level. Because difference testing at the peptide group-level avoids the step of normalizing spectral counts (or PPIDs) for each individual protein in a protein group (which may vary in size or sequence coverage), we would like to show that peptide group-level difference testing will be able to appropriately handle the issue of peptides in a given peptide group being shared amongst the proteins in their respective associated protein groups.

Bipartite graphs of the peptide groups and protein groups from four example candidate tests are shown below (Figure 19). Each graph is shown for the candidate test alone, without the combined model applied. Each candidate test's graph shows the peptide groups and protein

groups from all protein group clusters that contain a single peptide group found to be differentially abundant (TPs and FPs). Unshared peptide groups and associated protein groups are not shown. Circles represent protein groups, triangles represent peptide groups (TP, FN, TN), and squares represent FP peptide groups. White triangles accepted the null hypothesis. Triangles and squares that reject the null hypothesis (i.e. TPs and FPs) are colored yellow-to-black, with yellow the least significant and black most significant of the declared TPs. As shown in the figures, the differentially abundant peptide groups were almost exclusively amongst unshared peptide groups.

Of the four bipartite graphs shown in Figure 19, only four shared peptide groups rejected the null hypothesis. Three out of those four shared peptide groups were FPs, with the sole TP shared peptide group found by Fisher's exact test. If a shared peptide group is found to be differentially abundant, than we would expect at least one of associated protein groups to have an unshared peptide group that also rejects the null hypothesis. This was not the case for the either of the unshared peptide groups connected to the protein groups associated with the shared peptide group identified by Fisher's exact test. These unshared peptide groups were not FNs either (FN data not shown), raising the possibility that this peptide group is in the "answer key" erroneously. Regardless, that three out of four of the shared peptide groups were FPs suggests that unshared peptide groups are a far less error prone than shared peptide groups, as would be expected.

An additional observation is that the FPs tended to be among the least significant of those peptide groups that rejected the null hypothesis (i.e. more yellow than black). This may

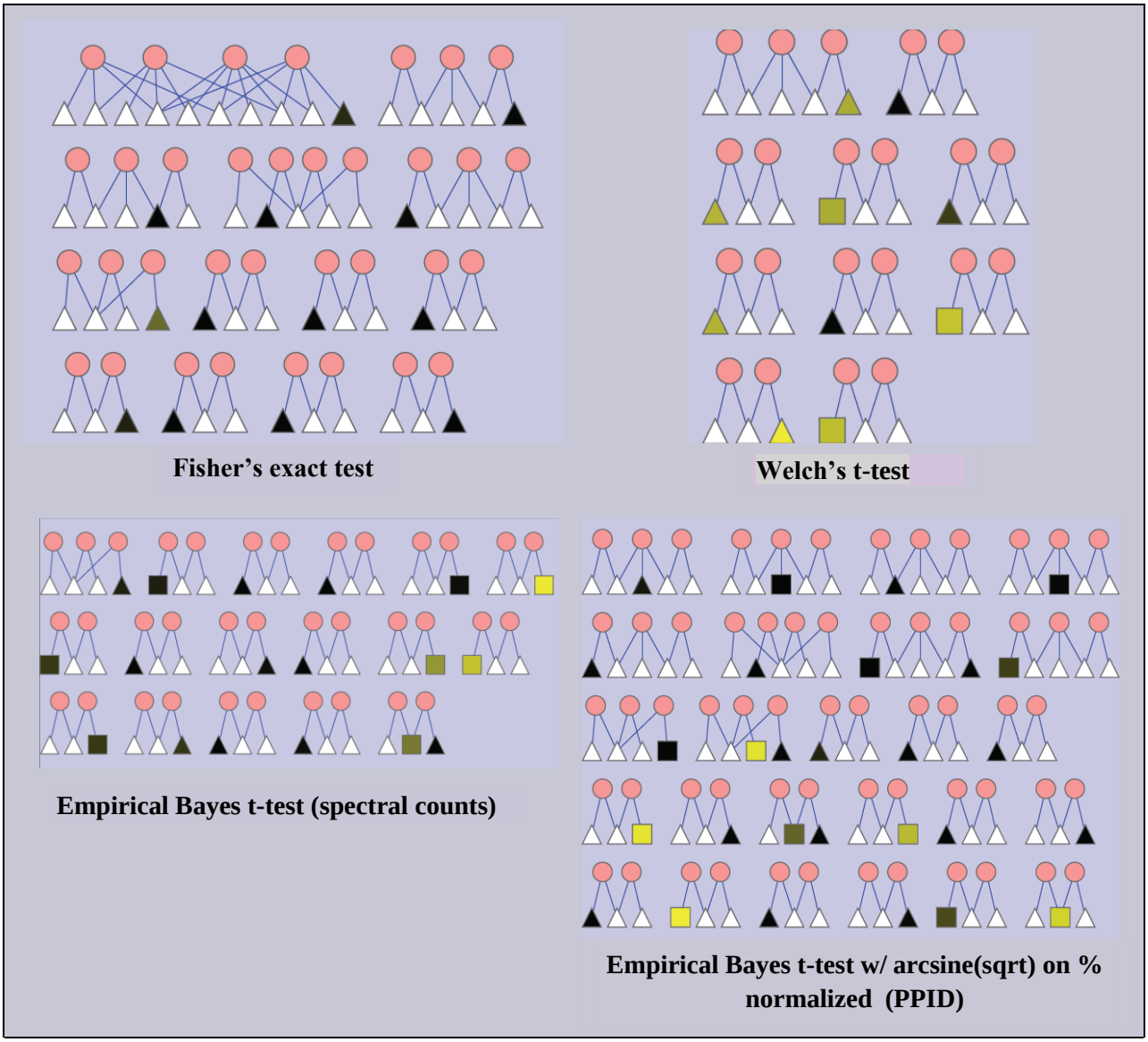


Figure 19. The peptide groups and associated protein groups from example candidate tests showing p-values significance levels and FPs for each protein group cluster contain at least one peptide group yielding a significant test result. Circles represent protein groups, triangles represent peptide groups (TP, FN, TN), and squared represent FP peptide groups. White triangles accepted the null hypothesis. TPs and FPs are colored yellow-to-black, with yellow the least significant and black most significant of the declared TPs.

partially explain the apparent gains in performance that came from applying Fisher's Method in our combined model. If FPs tend to have less significant p-values, they are more prone to being eliminated by a technique like Fisher's Method.

Summary and Further Discussion

In this chapter, we used the iPRG 2009 Study *E. coli* LC-MS/MS data set to compare and contrast the performance of statistical difference testing at the peptide group level using either spectral counts or precursor ion intensities separately. We also examine the use of two PPID-weighting techniques to account for missing PPIDs and two spectral count/PPID filtering techniques to eliminate peptide groups with few data points. Using the F_1-measure and ROC analysis as our performance metrics and Fisher's exact test (on spectral counts) as our benchmark, we first observed that no test could surpass the benchmark test's ability to identify 69 out of 115 TPs with only 7 FPs. Most of the candidate tests were able to find more TPs, but at the cost of significantly more FPs. Both weighting techniques showed moderate increases or decreases to test performance inconsistently. The filtering methods both showed mildly positive results when filters were set to exclude only those peptide groups with fewer than 5 or 10 PPIDs or spectral counts per cohort (or across both cohorts)

When candidate tests were combined with the benchmark test using Fisher's Method (our combined model), we observed a sharp drop in FPs with only a moderate drop in TPs relative to the candidate tests alone. These promising results required further investigation because Fisher's Method, or other meta-analysis techniques, requires the separate p-values to be independent of each other. The PPID values used in this study were generated using identified spectra as a starting point, so the PPIDs and spectral count values could not be fully independent in our study. At the same time, the extremely high precision of the benchmark test compared to the candidate tests suggests the benchmark test had a distinctly different bias towards not accepting

FPs exhibited by the other candidate tests when either spectral counts or PPIDs were used for the candidate tests. Venn chart analysis comparing the TPs and FPs of candidate tests with our benchmark test also indicate that the benchmark test was also selecting for an overlapping, but still distinctly different TPs than then other tests. These observations indicate that spectral count and PPID data contain both mutual information but also orthogonal information since there are some TPs and FPs found exclusively using spectral counts and some found only using PPIDs.

The partial independence between PPIDs and spectral counts may results in moderately different abundance ratios, but that moderate difference can result in qualitatively different statistical test results. This is highlighted by the plots of the $\log_2$ ratio between the Red and Yellow cohorts of spectral counts (y-axis) and PPIDs (x-axis). The $\log_2$ ratio plots of FPs demonstrated a tendency of PPID-based statistical tests to yield FPs with low PPID ratios, but even lower spectral count ratios. Similarly, spectral count-based candidate tests yielded FPs with low spectral count ratios, but even lower PPID ratios. For these peptide groups with equal or nearly equal abundances, a modest independence in the spectral count data and PPID data may result in a rejection of the null hypothesis with one data type and acceptance of the null hypothesis with the other.

Old *et al.* observed a similar pattern when they statistically tested for differentially abundant proteins using either precursor intensities or spectral counts. They found most of proteins that were found to be differentially abundant when using precursor intensities but not when using spectral counts were proteins with low abundance ratios. They also found that a 2.3 fold change was the smallest fold change they could identify with 95% confidence using precursor peak abundance(11). Their findings underscore the greater difficulty in difference testing faced as the abundance ratio approaches 1.

Our results strongly suggest that our combined model can give researchers a powerful new approach of combining the results from two statistical tests with different biases into a single hybrid test that exhibits properties of both constituent tests. In particular, we demonstrated that combining the results of a conservative test (high precision, lower sensitivity) with more sensitive but less precise tests resulted in hybrid tests that exhibited enhanced sensitivities with a moderate cost to precision over the conservative test. This approach could be useful in experiments where researchers are willing to accept a moderate level of FPs in order to identify more TPs but still want to avoid the large numbers of TPs yielded by many quantitative methods.

It is also worth noting that the performance of our candidate tests was similar to the results of iPRG 2009 Study. In the study, the iPRG 2009 *E. coli* data set was distributed to multiple proteomics groups (52). Each group was asked to identify the differentially abundant proteins using any method of their choice and return their results the study authors to compare the performances of the different approaches. Each group's test performance was evaluated at multiple error rate thresholds (1%, 5%, 10%, and 25%). The authors found that groups using spectral counts consistently correctly identified more differentially abundant proteins at all thresholds due to the high error rates seen for groups using precursor intensities. The one exception was at the 25% error rate threshold where a group using precursor intensities had the top performance. As the authors pointed out, the use of precursor intensities also requires more experience than spectral count methods and many of the protein differences in the iPRG 2009 Study data set were much stronger than would be expected for biological samples. Both of these points suggest that the number of FPs generated by precursor intensity-based methods in biomarker studies will be an ongoing challenge going forward. Consistent with the iPRG 2009

Study, our evaluations of the separate candidate tests found our benchmark test, Fisher's exact test on spectral counts, to be more conservative than any of the PPID-based tests. Fisher's exact test is thus an excellent statistical test to pair with precursor intensity-based tests for researchers interested in a statistical test less conservative than Fisher's exact test but more conservative than available the is cat precursor intensity-based methods.

We also addressed the feasibility of difference testing at the peptide group-level as opposed to the protein level as is normally done, as specified in Specific Aim III. We found few shared peptides groups that yielded rejected null hypotheses, and those that did reject the null hypothesis were largely false positives. We also observed that the false positives tended to have less significant p-values than that the true positives. This demonstrates the utility of peptide group analysis when using the combined model if false positives tend to have the less significant p-values amongst the peptide groups with rejected null hypothesis. Overall, it appears that peptides belonging to shared peptide groups should be ignored in quantitative analysis, whereas quantitative data from peptides belonging to an unshared peptide group should be retained and used for quantitative analysis. The peptides in an unshared peptide group are still shared peptides, but shared only between the proteins in a single protein-group (unless there is only a single peptide or single protein in the peptide group and protein group). This method of rejecting quantitative peptide data from shared peptide groups and only using unshared peptide group data represents a novel form of grouping and filtering quantitative data from shared peptides.

CHAPTER V

CONCLUSION AND FUTURE WORK

Advances in mass spectrometry have resulted in enormous improvements in both the quality and quantity of LC-MS/MS data. But unlocking the full biological knowledge that can be learned from that data will require new methodologies of quantitative analysis and statistical techniques. We hypothesized that enhanced proteomic difference testing can be achieved by combining spectral count data with precursor intensity data (labeled our "combined model" in this study). We also hypothesized that difference testing at the peptide group-level provided a viable means of group peptides and filtering out noise from shared peptides. To achieve this, we had three specific aims: I. To compare performance and independence of difference testing using spectral count and precursor intensities alone; II. To compare the performance of difference testing using Fisher's Method to combine p-values of PPID and spectral count-based statistical tests vs. difference testing using each data type alone; and III. To establish the validity of difference testing at the peptide group-level vs. the protein level. All three specific aims were accomplished in this study and both of our hypotheses were validated.

Specific Aim I was enabled by the development of IDPQuantify (Chapter III), which provided peptide group-level spectral count and precursor intensity data in the form of PPIDs. The iPRG 2009 *E. coli* data set with a known "answer key" was then used to evaluate and compared the statistical test alone (Chapter IV). A variety of candidate statistical tests using either PPIDs or spectral counts were chosen for this analysis. We found that spectral count-based testing using our benchmark test (Fisher's exact test on spectral counts) provided superior

specificity, but reduced sensitivity compared to PPID-based tests. The PPID-based test, conversely, had greater potential sensitivity, but at the cost of far more FPs than Fisher's exact test.

The orthogonality and mutual information contained in spectral count and PPID data was further characterized using Venn chart analysis and cohort abundance ratio analysis. Venn chart analysis showed distinctly different sets of TPs and FPs that were likely to be yielded using spectral count or PPID-based tests. Using the abundance ratio plots, we observed that the set FPs for all tests contained a large number of low-abundant peptide groups, regardless of whether spectral counts or PPIDs were used. Among the more abundant FPs, however, a pattern emerged distinguishing the spectral count tests and PPID tests. The more abundant FPs generated by spectral count-based tests tended to have a low spectral count ratio, but an even lower PPID ratio. For PPID-based tests, the opposite pattern was seen. This suggested that a given LC-MS/MS data set is going to contain a mix of peptide groups that are likely to trigger FPs from spectral count methods, but not precursor intensity methods, and vice versa.

For Specific Aim II, we then evaluated our combined model for difference testing to the candidate tests alone. This model was shown to create hybrid statistical results that tended to eliminate weakly rejected null hypotheses from either test, while retaining the most strongly rejected null hypotheses. When candidate statistical tests were combined with the more conservative Fisher's exact tests, the results were a modest drop in sensitivity but much sharper drop in the error rate (relative to the candidate test alone). This appears to be due, in part, to the erroneously rejected null hypothesis (FPs) tending to have less significant p-values (e.g. closer to 0.05 than 0) (Figure 19). Thus, FPs generated by the candidate tests were more likely to be discarded when using Fisher's method than true positives (TPs).

For Specific Aim III, the analysis of peptide group-level difference testing in place of protein-level difference testing, consisted of viewing the tendency of shared peptide groups and unshared peptide groups to yield FPs. In the four candidate test examples shown in Figure 31, there were a total of four shared peptide groups total with rejected null hypotheses, three of which were FPs. This indicates that peptides that are shared across multiple protein groups are a greater source of error than peptides shared only across proteins in a single protein group. and calls for relying primarily on peptides from unshared peptide groups for difference testing. While we conducted our difference testing without normalizing each protein individually (such as using NSAF or emPAI), peptide group level analysis may still be useful to researchers that desire protein-specific normalization. Following the grouping of peptides into peptide groups and the discarding of quantitative data from shared peptide groups, researchers are still free to further normalize each protein individually. As discussed previously, however, conducting a separate test for each protein in a protein group raises issues of independence and expands the total number of tests conducted. A greater number of tests can lead to additional type I errors and conflicts with multiple testing correction procedures, such as the Benjamini-Hochberg procedure.

Independence of tests is also a requirement for the application of Fisher's Method. Our spectral count and precursor intensity data were clearly not independent since precursor intensities, in the form of PPIDs, were derived from spectral counts. But while the spectral count and PPID data in this study may not have been fully independent, the behavior of spectral count and PPID-based tests explored in Specific Aim I suggests that spectral counts and PPIDs were independent enough for our combined model to yield useful results.

Overall, we established that the use of spectral count or precursor intensities for difference testing involves a tradeoff between the more sensitive precursor-based approach and the more accurate spectral count-based approach. These differences in test performance were due, in part, to differences the underlying data but also the types of tests available for each data type (e.g. parametric vs. non-parametric). We also established that these differences in test performance included overlapping, but still distinct sets of TPs and FPs obtained when using either metric. We then demonstrated that our combined model gives researchers a tool for creating hybrid statistical result sets that tended to have a slight drop in sensitivity from the PPID-based tests along with a sharp drop in FPs. The FPs, in turn, tended to have the least significant p-values among the rejected null hypotheses, making them more likely to be discarded upon application of our combined model. We also demonstrated that peptide group-level analysis is effective at segregating peptides into noisier shared peptide groups and less noisy unshared peptide groups.

An obvious area for future work involves using precursor intensities that were truly generated independently from spectral counts. This could involve the use of peak alignment strategies on LC-MS data to extract precursor intensities without any knowledge of peptide identification. In addition to making the two types of data more independent of each other, peak alignment strategies for precursor intensity generation would likely improve the performance of tests relying on precursor intensities given the difficulty many parametric tests face with missing data points. Any method that improves either precursor intensity-based testing or spectral count-based testing alone will also improve the results from our combined model. Thus, the performance of the combined model using improved de-noising, normalization, or data imputation strategies for missing data is something to be explored.

Another issue needing further analysis is the use of our combined model on more complicated data sets with less striking differences in abundance. As discussed in Chapter IV, the method for creating the iPRG 2009 Study data set resulted in more pronounced differences in abundance for the TPs than would be normally expected. In data sets involving fewer TPs with less prominent changes, many TPs found by a single test may be only weakly significant (i.e. a p-value close to 0.05) and Fisher's Method may result in the erroneous rejections. In such a case, the pairing of two sensitive, but less conservative, tests using Fisher's Method may be advised.

Alternatively, more than two tests could be combined to provide a larger number of p-value "votes" for each peptide group. Three or more tests will result at least two tests sharing either precursor intensities or spectral counts, so this approach would come with the added caveat of a lack of independence. Two tests using the same data type, however, may still be partially independent if the tests follow very different statistical assumptions or biases. For instance, while the Mann Whitney U test and Fisher's exact test can both be run on spectral counts, the Mann Whitney U test suffers much more from few replicates than Fisher's exact test. The extent to which different statistical tests behave independently when run on the same data is an important question for applying this combined model to proteomic quantitation.

The application of the combined model to protein-level difference testing rather than peptide group difference testing could also be evaluated. One of the disadvantages of peptide group-level difference testing over protein-level difference testing is that peptide group-level difference test will generally use less information (i.e. less peptide-specific quantitative data points) than if the test was conducted at the level of individual proteins and shared peptides are accounted handled instead of discarded. The impact of the peptide group-level information

dilution needs to be addressed with for species with more complex proteomes. Humans, for instance, will have far more shared peptide groups than *E. coli* and more overall protein identifications. This dilutes the amount of sampling that can be conducted per-protein, resulting in fewer proteins per peptide group. At the same time, because human proteins tend to have more isoforms than *E. coli*, the protein groups will likely consist of more individual proteins. Thus, strategies that divide spectral count data between shared proteins may end up with little signal for each of the separate protein tests. In addition, the potential lack of independence if quantitative data is shared for multiple proteins (as in the original NSAF method) would need to be addressed. Applying the combined model to protein-level difference testing also raises the issue of extra non-independent tests compared to peptide group-level testing when employing multiple testing correction procedures.

The potential lack of independence between tests begs for more understanding of how to best control the FDR using our combined model. The Benjamini-Hochberg procedure used in this study assumed independence and focuses on controlling for the PCER instead of the FWER. Non-independent tests may end up causing more extreme p-values after Fisher's Method is applied, in particular if more than two non-independent p-values are to be combined and multiple p-values rely on the same data (spectral counts or precursor intensities). In such a case, using a more conservative FWER procedure to control the FDR may be in order.

Another approach may be to use a procedure that does not assume independence, such as the positive False Discovery Rate (pFDR) method using q-values developed by Storey *et. al* (53). As the authors pointed out, an advantage of the pFDR over the Benjamini-Hochberg correction is that the Benjamini-Hochberg correction attempts to control the FDR for all tests, whereas the pFDR procedure only attempts to control for FDR among those tests that resulted in a rejected

null hypothesis. This makes the pFDR much less conservative in instances when few null hypotheses are rejected. For more complex data sets with subtler shifts in abundance the pFDR procedure may be a better choice than Benjamini-Hochberg.

Biomarker detection is an area where our combined model could be especially useful. Striking a balance between sensitivity and accuracy is important for such studies and our model could enable researchers with greater flexibility in modulating the performance characteristics of their tests. Our combined model assumes two tests, one on spectral counts and one on precursor intensities, leaving a great deal of discretion to the researcher in determining which statistical tests to use and what type of normalization and FDR controlling procedures are employed. For each of the two tests, there is a tradeoff between the sensitivity and power of the tests coupled with the strength of the FDR controlling procedure. The results of these choices are further coupled when the test p-values are combined using Fisher's Method. Just this simple two-test scenario leaves many possible permutations of pairs of tests coupled with FDR methods that have yet to be evaluated.

Another possible application of our combined model is the improved detection of low concentration proteins that change in abundance. As described in Chapter II, published studies indicate that both spectral count and precursor intensity-based methods suffer higher error rates for low-abundant proteins. Our combined model may be able to reduce that error rate by providing a second piece of evidence for each protein testing. Furthermore, by conducting difference testing at the peptide-group level, we may be better suited to accurately test low-abundance proteins. Peptide group-level difference testing separates the less-reliable data from shared peptide groups from the more reliable data from unshared peptide groups. Subsequent

testing of the unshared peptide group data together, instead of diluting it further as shared peptides, could be particularly useful for low abundance proteins where the signal is limited.

Label-free proteomic quantitation using LC-MS/MS data has been a challenging area of research for many years. The data acquired are noisy and complex and are nowhere near as noisy or complex as the biological systems being studied. For LC-MS/MS to remain an invaluable tool to researchers, new methods will need to be developed to keep pace with the new investigative demands of the research community. In this study we presented two such new methods and a new tool enabling those methods. It is hoped that the methods introduced in this study will enable researchers to gain more real insights from their data with fewer errors.

REFERENCES

1. Becker, C. H., and Bern, M. (2010) Recent developments in quantitative proteomics. *Mutat Res* Epub ahead of print.
2. Saar, E., Gerostamoulos, D., Drummer, O. H., and Beyer, J. (2010) Identification and quantification of 30 antipsychotics in blood using LC-MS/MS. *J Mass Spectrom* Epub ahead of print.
3. Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol* 17, 994-999.
4. Ross, P. L., Huang, Y. N., Marchese, J. N., 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., and Pappin, D. J. (2004) Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics* 3, 1154-1169.
5. Yao, X., Freas, A., Ramirez, J., Demirev, P. A., and Fenselau, C. (2001) Proteolytic 18O labeling for comparative proteomics: model studies with two serotypes of adenovirus. *Anal Chem* 73, 2836-2842.
6. Ong, S. E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A., and 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, 376-386.
7. Wang, W., Zhou, H., Lin, H., Roy, S., Shaler, T. A., Hill, L. R., Norton, S., Kumar, P., Anderle, M., and Becker, C. H. (2003) Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Anal Chem* 75, 4818-4826.
8. Chelius, D., and Bondarenko, P. V. (2002) Quantitative profiling of proteins in complex mixtures using liquid chromatography and mass spectrometry. *J Proteome Res* 1, 317-323.
9. Bondarenko, P. V., Chelius, D., and Shaler, T. A. (2002) Identification and relative quantitation of protein mixtures by enzymatic digestion followed by capillary reversed-phase liquid chromatography-tandem mass spectrometry. *Anal Chem* 74, 4741-4749.
10. Liu, H., Sadygov, R. G., and Yates, J. R., 3rd (2004) A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem* 76, 4193-4201.
11. Old, W. M., Meyer-Arendt, K., Aveline-Wolf, L., Pierce, K. G., Mendoza, A., Sevinsky, J. R., Resing, K. A., and Ahn, N. G. (2005) Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol Cell Proteomics* 4, 1487-1502.
12. Park, S. K., Venable, J. D., Xu, T., and Yates, J. R., 3rd (2008) A quantitative analysis software tool for mass spectrometry-based proteomics. *Nat Methods* 5, 319-322.
13. Shen, Y., Zhao, R., Berger, S. J., Anderson, G. A., Rodriguez, N., and Smith, R. D. (2002) High-efficiency nanoscale liquid chromatography coupled on-line with mass spectrometry using nanoelectrospray ionization for proteomics. *Anal Chem* 74, 4235-4249.
14. Venable, J. D., Dong, M. Q., Wohlschlegel, J., Dillin, A., and Yates, J. R. (2004) Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat Methods* 1, 39-45.
15. Usaite, R., Wohlschlegel, J., Venable, J. D., Park, S. K., Nielsen, J., Olsson, L., and Yates Iii, J. R. (2008) Characterization of global yeast quantitative proteome data generated from the wild-type and glucose repression saccharomyces cerevisiae strains: the comparison of two quantitative methods. *J Proteome Res* 7, 266-275.
16. Zybailov, B., Mosley, A. L., Sardiu, M. E., Coleman, M. K., Florens, L., and Washburn, M. P. (2006) Statistical analysis of membrane proteome expression changes in Saccharomyces cerevisiae. *J Proteome Res* 5, 2339-2347.

17. Pavelka, N., Fournier, M. L., Swanson, S. K., Pelizzola, M., Ricciardi-Castagnoli, P., Florens, L., and Washburn, M. P. (2008) Statistical similarities between transcriptomics and quantitative shotgun proteomics data. *Mol Cell Proteomics* 7, 631-644.
18. Humphery-Smith, I. (2004) A human proteome project with a beginning and an end. *Proteomics* 4, 2519-2521.
19. Zhang, Y., Wen, Z., Washburn, M. P., and Florens, L. (2010) Refinements to label free proteome quantitation: how to deal with peptides shared by multiple proteins. *Anal Chem* 82, 2272-2281.
20. Rappsilber, J., Ryder, U., Lamond, A. I., and Mann, M. (2002) Large-scale proteomic analysis of the human spliceosome. *Genome Res* 12, 1231-1245.
21. Ishihama, Y., Oda, Y., Tabata, T., Sato, T., Nagasu, T., Rappsilber, J., and Mann, M. (2005) Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol Cell Proteomics* 4, 1265-1272.
22. Dong, J., Lai, R., Jennings, J. L., Link, A. J., and Hinnebusch, A. G. (2005) The novel ATP-binding cassette protein ARB1 is a shuttling factor that stimulates 40S and 60S ribosome biogenesis. *Mol Cell Biol* 25, 9859-9873.
23. McAfee, K. J., Duncan, D. T., Assink, M., and Link, A. J. (2006) Analyzing proteomes and protein function using graphical comparative analysis of tandem mass spectrometry results. *Mol Cell Proteomics* 5, 1497-1513.
24. Zhang, B., Chambers, M. C., and Tabb, D. L. (2007) Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. *J Proteome Res* 6, 3549-3557.
25. Nesvizhskii, A. I., Keller, A., Kolker, E., and Aebersold, R. (2003) A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem* 75, 4646-4658.
26. Maslov, S., Sneppen, K., Eriksen, K. A., and Yan, K. K. (2004) Upstream plasticity and downstream robustness in evolution of molecular networks. *BMC Evol Biol* 4:9.
27. Jin, S., Daly, D. S., Springer, D. L., and Miller, J. H. (2008) The effects of shared peptides on protein quantitation in label-free proteomics by LC/MS/MS. *J Proteome Res* 7, 164-169.
28. Quackenbush, J. (2002) Microarray data normalization and transformation. *Nat Genet* 32 Suppl, 496-501.
29. Choi, H., Fermin, D., and Nesvizhskii, A. I. (2008) Significance analysis of spectral count data in label-free shotgun proteomics. *Mol Cell Proteomics* 7, 2373-2385.
30. Lonnstedt, I., and Speed, T. (2002) Replicated Microarray Data. *Statistica Sinica* 12, 31-46.
31. Wettenhall, J. M., and Smyth, G. K. (2004) limmaGUI: a graphical user interface for linear modeling of microarray data. *Bioinformatics* 20, 3705-3706.
32. Smyth, G. K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 3:3, Epublication.
33. Brusniak, M. Y., Bodenmiller, B., Campbell, D., Cooke, K., Eddes, J., Garbutt, A., Lau, H., Letarte, S., Mueller, L. N., Sharma, V., Vitek, O., Zhang, N., Aebersold, R., and Watts, J. D. (2008) Corra: Computational framework and tools for LC-MS discovery and targeted mass spectrometry-based proteomics. *BMC Bioinformatics* 9:542.
34. Neubert, H., Bonnert, T. P., Rumpel, K., Hunt, B. T., Henle, E. S., and James, I. T. (2008) Label-free detection of differential protein expression by LC/MALDI mass spectrometry. *J Proteome Res* 7, 2270-2279.
35. Li, M., Gray, W., Zhang, H., Chung, C. H., Billheimer, D., Yarbrough, W. G., Liebler, D. C., Shyr, Y., and Slebos, R. J. (2010) Comparative shotgun proteomics using spectral count data and quasi-likelihood modeling. *J Proteome Res* 9, 4295-4305.
36. Wolf, F. M. (1986) *Meta-analysis:quantitative methods for research*, 1st Ed., pp. 39-44, Sage Publications, Newbury Park, CA.

37. Hedges, L. V. (1986) *Statistical Methods for Meta-analysis*, 1st Ed., pp. 27-28, 37-39, Academic Press., San Diego, CA.
38. Ma, Z. Q., Dasari, S., Chambers, M. C., Litton, M. D., Sobecki, S. M., Zimmerman, L. J., Halvey, P. J., Schilling, B., Drake, P. M., Gibson, B. W., and Tabb, D. L. (2009) IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. *J Proteome Res* 8, 3872-3881.
39. Hoopmann, M. R., Finney, G. L., and MacCoss, M. J. (2007) High-speed data reduction, feature detection, and MS/MS spectrum quality assessment of shotgun proteomics data sets using high-resolution mass spectrometry. *Anal Chem* 79, 5620-5632.
40. Hsieh, E. J., Hoopmann, M. R., MacLean, B., and MacCoss, M. J. (2010) Comparison of database search strategies for high precursor mass accuracy MS/MS data. *J Proteome Res* 9, 1138-1143.
41. Team, R. D. C. (2010) R: A Language and Environment for Statistical Computing.
42. Horn, D. M., Zubarev, R. A., and McLafferty, F. W. (2000) Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *J Am Soc Mass Spectrom* 11, 320-332.
43. Ritchie, M. E., Silver, J., Oshlack, A., Holmes, M., Diyagama, D., Holloway, A., and Smyth, G. K. (2007) A comparison of background correction methods for two-colour microarrays. *Bioinformatics* 23, 2700-2707.
44. Benjamini, Y., and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Statist Soc B* 57, 289-300.
45. Freeman, M. F., and Tukey, J. W. (1950) Transformations related to the angular and the square root. *Ann of Math Statistics* 21, 607-611.
46. Deutsch, E. (2008) mzML: a single, unifying data format for mass spectrometer output. *Proteomics* 8, 2776-2777.
47. Martens, L., Chambers, M., Sturm, M., Kessner, D., Levander, F., Shofstahl, J., Tang, W. H., Rompp, A., Neumann, S., Pizarro, A. D., Montecchi-Palazzi, L., Tasman, N., Coleman, M., Reisinger, F., Souda, P., Hermjakob, H., Binz, P. A., and Deutsch, E. W. (2010) mzML - a Community Standard for Mass Spectrometry Data. *Mol Cell Proteomics* Epub ahead of print.
48. Kessner, D., Chambers, M., Burke, R., Agus, D., and Mallick, P. (2008) ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* 24, 2534-2536.
49. Tabb, D. L., Fernando, C. G., and Chambers, M. C. (2007) MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J Proteome Res* 6, 654-661.
50. Zhang, B., VerBerkmoes, N. C., Langston, M. A., Uberbacher, E., Hettich, R. L., and Samatova, N. F. (2006) Detecting differential and correlated protein expression in label-free shotgun proteomics. *J Proteome Res* 5, 2909-2918.
51. Meyer, K. M., Mooij, W. M., Vos, M., Hol, W. H., and Putten, W. H. (2009) The power of simulating experiments. *Ecological Modeling* 220, 2594-2597.
52. Searle, S. C., Tabb, D. L., Falkner, J. A., Kowalak, J. A., Meyer-Arendt, K., Martens, L., Askenazi, M., Rudnick, P. A., Seymour, S. L., and Lane, W. S. (2009) iPRG2009 study: testing for differences between complex samples in proteomics datasets. *ABRF Poster*.
53. Storey, J. D., and Tibshirani, R. (2001) Estimating the positive false discovery rate under dependence, with applications to DNA microarrays. *Technical Report No. 2001-28, Department of Statistics, Stanford University*.

Supplemental File 1. MyriMatch Search Configuration Parameters

MyriMatch search parameters for iPRG 2009 *E. coli* LC-MS/MS data set were as follows:

CleavageRules = [|[M|K|R . .]
NumMinTerminiCleavages = 1
CalculateRelativeScores = 0
DynamicMods = M * 15.994915 (Q! % -17.026549 C & 57.021464
AdjustPrecursorMass = true
MinPrecursorAdjustment = -1.008665
MaxPrecursorAdjustment = 1.008665
PrecursorAdjustmentStep = 1.008665
NumSearchBestAdjustments = 3
PrecursorMzTolerance = 0.1
FragmentMzTolerance = 0.5
TicCutoffPercentage = 0.95
UseAvgMassOfSequences = 0
UseChargeStateFromMS = 1
NumChargeStates = 5

Supplemental File 2. F_1-Measures for Candidate Tests

Table S2.1 F_1-Measures For Classical Tests Using the "Minimum Spectral Count/PPID in at least one cohort" filter

	Minimum Spectral Count/PPID in at least one cohort					
Candidate Test	**0**	**5**	**10**	**15**	**20**	**25**
Fisher's exact test on Spectral Counts (Benchmark Test)	0.70	0.70	0.70	0.70	0.64	0.62
Student's t-test on PPIDs	0.60	0.60	0.61	0.58	0.52	0.46
Student's t-test log transformation on PPIDs	0.55	0.55	0.58	0.50	0.44	0.38
Welch's t-test on PPIDs	0.49	0.49	0.50	0.43	0.41	0.31
Welch's t-test log transformation on PPIDs	0.55	0.55	0.55	0.46	0.42	0.36
Mann Whitney U test on Spectral Counts	0.58	0.60	0.65	0.66	0.62	0.59
Modified Mann Whitney U test on Spectral Counts	0.64	0.65	0.70	0.72	0.67	0.63
Mann Whitney U test on PPIDs	0.58	0.59	0.65	0.67	0.65	0.61

Table S2.2 F_1-Measures For Classical Tests Using the "Minimum Spectral Count/PPID across both cohorts" filter

	Minimum Spectral Count/PPID across both cohorts					
Candidate Test	**0**	**5**	**10**	**15**	**20**	**25**
Fisher's exact test on Spectral Counts (Benchmark Test)	0.70	0.70	0.70	0.70	0.70	0.70
Student's t-test on PPIDs	0.60	0.60	0.61	0.59	0.56	0.46
Student's t-test log transformation on PPIDs	0.55	0.55	0.58	0.52	0.47	0.40
Welch's t-test on PPIDs	0.49	0.49	0.50	0.46	0.45	0.34
Welch's t-test log transformation on PPIDs	0.55	0.55	0.54	0.46	0.43	0.38
Mann Whitney U test on Spectral Counts	0.58	0.58	0.58	0.58	0.58	0.58
Modified Mann Whitney U test on Spectral Counts	0.64	0.64	0.64	0.64	0.64	0.64
Mann Whitney U test on PPIDs	0.58	0.60	0.61	0.62	0.64	0.61

Table S2.3 F_1-Measures For Combined Model (Classical Tests + Benchmark) Using the "Minimum Spectral Count/PPID in at least one cohort" filter

	Minimum Spectral Count/PPID in at least one cohort					
Candidate Test	**0**	**5**	**10**	**15**	**20**	**25**
Fisher's exact test on Spectral Counts (Benchmark)	0.70	0.70	0.70	0.70	0.64	0.62
Student's t-test on PPIDs	0.71	0.70	0.72	0.71	0.69	0.67
Student's t-test log transformation on PPIDs	0.70	0.70	0.71	0.69	0.68	0.66

Welch's t-test on PPIDs	0.72	0.72	0.72	0.71	0.69	0.67
Welch's t-test log transformation on PPIDs	0.70	0.70	0.70	0.68	0.68	0.66
Mann Whitney U test on Spectral Counts	0.70	0.70	0.73	0.73	0.71	0.68
Modified Mann Whitney U test on Spectral Counts	0.71	0.71	0.73	0.73	0.72	0.69
Mann Whitney U test on PPIDs	0.72	0.72	0.74	0.74	0.73	0.70

Table S2.4 F1-Measures For Combined Model (Classical Tests + Benchmark) Using the "Minimum Spectral Count/PPID across both cohorts" filter

	Minimum Spectral Count/PPID across both cohorts					
Candidate Test	**0**	**5**	**10**	**15**	**20**	**25**
Fisher's exact test on Spectral Counts (Benchmark)	0.70	0.70	0.70	0.70	0.70	0.70
Student's t-test on PPIDs	0.71	0.70	0.72	0.72	0.71	0.69
Student's t-test log transformation on PPIDs	0.70	0.70	0.72	0.70	0.70	0.68
Welch's t-test on PPIDs	0.72	0.72	0.72	0.71	0.71	0.69
Welch's t-test log transformation on PPIDs	0.70	0.70	0.70	0.69	0.69	0.68
Mann Whitney U test on Spectral Counts	0.70	0.70	0.70	0.70	0.70	0.70
Modified Mann Whitney U test on Spectral Counts	0.71	0.71	0.71	0.71	0.71	0.71
Mann Whitney U test on PPIDs	0.72	0.72	0.74	0.74	0.72	0.70

Table S2.5 F1-Measures For Empirical Bayes t-test Using the "Minimum Spectral Count/PPID in at least one cohort" filter

	Minimum Spectral Count/PPID in at last one cohort					
Candidate Test	**0**	**5**	**10**	**15**	**20**	**25**
Fisher's exact test on Spectral Counts (Benchmark)	0.70	0.70	0.70	0.70	0.64	0.62
empirical Bayes on PPIDs	0.63	0.65	0.66	0.63	0.59	0.49
empirical Bayes on PPIDs + missed PPID weighting	0.61	0.66	0.63	0.62	0.57	0.50
empirical Bayes on PPIDs + observed PPID weighting	0.62	0.66	0.65	0.62	0.58	0.49
empirical Bayes on PPIDs + log normalization	0.58	0.58	0.61	0.53	0.50	0.42
empirical Bayes on PPIDs + log normalization + missed PPID weighting	0.57	0.58	0.62	0.56	0.51	0.44
empirical Bayes on PPIDs + log normalization + observed PPID weighting	0.56	0.59	0.62	0.55	0.51	0.45

empirical Bayes on PPIDs + percentile normalization	0.66	0.65	0.69	0.64	0.59	0.49
empirical Bayes on PPIDs + percentile normalization + missed PPID weighting	0.69	0.65	0.67	0.64	0.58	0.49
empirical Bayes on PPIDs + percentile normalization + observed PPID weighting	0.64	0.65	0.71	0.67	0.62	0.51
empirical Bayes on PPIDs + percentile+arcsine(sqrt) normalization	0.65	0.66	0.70	0.66	0.59	0.49
empirical Bayes on PPIDs + percentile+arcsine normalization + missed PPID weighting	0.67	0.66	0.69	0.64	0.58	0.49
empirical Bayes on PPIDs + percentile+arcsine normalization + observed PPID weighting	0.66	0.66	0.72	0.67	0.61	0.51
empirical Bayes on PPIDs + max value normalization	0.49	0.49	0.49	0.49	0.48	0.43
empirical Bayes on PPIDs + max value normalization + missed PPID weighting	0.50	0.50	0.50	0.50	0.49	0.44
empirical Bayes on PPIDs + max value normalization + observed PPID weighting	0.45	0.44	0.47	0.48	0.48	0.43
empirical Bayes on PPIDs + max value normalization + log transformation	0.56	0.53	0.54	0.51	0.49	0.43
empirical Bayes on PPIDs + max value normalization+log transformation + missed PPID weighting	0.55	0.53	0.54	0.52	0.50	0.45
empirical Bayes on PPIDs + max value normalization+log transformation + observed PPID weighting	0.56	0.51	0.53	0.49	0.48	0.43
empirical Bayes on PPIDs + max value normalization+arcsine(sqrt)	0.50	0.48	0.50	0.49	0.47	0.42
empirical Bayes on PPIDs + max value normalization+arcsine(sqrt) + missed PPID weighting	0.50	0.48	0.52	0.50	0.49	0.43
empirical Bayes on PPIDs + max value normalization+arcsine(sqrt) + observed PPID weighting	0.44	0.44	0.48	0.48	0.48	0.43
empirical Bayes on Spectral Counts	0.67	0.69	0.72	0.70	0.65	0.60

Table S2.6 F1-Measures For Empirical Bayes t-test Using the "Minimum Spectral Count/PPID across both cohorts" filter

	Minimum Spectral Count/PPID across both cohorts					
Candidate Test	**0**	**5**	**10**	**15**	**20**	**25**
Fisher's exact test on Spectral Counts (Benchmark)	0.70	0.70	0.70	0.70	0.70	0.70
empirical Bayes on PPIDs	0.63	0.65	0.64	0.63	0.60	0.53
empirical Bayes on PPIDs + missed PPID weighting	0.61	0.63	0.64	0.62	0.59	0.52
empirical Bayes on PPIDs + observed PPID weighting	0.62	0.66	0.65	0.63	0.60	0.52
empirical Bayes on PPIDs + log normalization	0.58	0.59	0.61	0.55	0.52	0.44
empirical Bayes on PPIDs + log normalization + missed PPID weighting	0.57	0.59	0.61	0.55	0.51	0.44
empirical Bayes on PPIDs + log normalization + observed PPID weighting	0.56	0.56	0.60	0.57	0.53	0.45
empirical Bayes on PPIDs + percentile normalization	0.66	0.65	0.65	0.63	0.59	0.51
empirical Bayes on PPIDs + percentile normalization + missed PPID weighting	0.69	0.65	0.66	0.62	0.58	0.52
empirical Bayes on PPIDs + percentile normalization + observed PPID weighting	0.64	0.65	0.69	0.62	0.62	0.53
empirical Bayes on PPIDs + percentile+arcsine(sqrt) normalization	0.65	0.65	0.65	0.65	0.66	0.62
empirical Bayes on PPIDs + percentile+arcsine normalization + missed PPID weighting	0.67	0.64	0.65	0.65	0.66	0.62
empirical Bayes on PPIDs + percentile+arcsine normalization + observed PPID weighting	0.66	0.64	0.65	0.65	0.66	0.60
empirical Bayes on PPIDs + max value normalization	0.49	0.49	0.48	0.47	0.46	0.42
empirical Bayes on PPIDs + max value normalization + missed PPID weighting	0.50	0.49	0.49	0.48	0.46	0.42
empirical Bayes on PPIDs + max value normalization + observed PPID weighting	0.45	0.44	0.45	0.45	0.44	0.40
empirical Bayes on PPIDs + max value normalization + log transformation	0.56	0.53	0.54	0.50	0.47	0.44
empirical Bayes on PPIDs + max value normalization+log transformation + missed PPID weighting	0.55	0.53	0.54	0.50	0.47	0.45
empirical Bayes on PPIDs + max value normalization+log	0.56	0.54	0.52	0.49	0.47	0.42

transformation + observed PPID weighting						
empirical Bayes on PPIDs + max value normalization+arcsine(sqrt)	0.50	0.48	0.49	0.47	0.45	0.41
empirical Bayes on PPIDs + max value normalization+arcsine(sqrt) + missed PPID weighting	0.50	0.48	0.49	0.47	0.46	0.42
empirical Bayes on PPIDs + max value normalization+arcsine(sqrt) + observed PPID weighting	0.44	0.44	0.45	0.44	0.44	0.40
empirical Bayes on Spectral Counts	0.67	0.67	0.67	0.67	0.67	0.67

Table S2.7 F1-Measures For Combined Model (Empirical Bayes t-test + Benchmark) Using the "Minimum Spectral Count/PPID at least one cohort" Filter

	Minimum Spectral Count/PPID across at least one cohort					
Candidate Test	**0**	**5**	**10**	**15**	**20**	**25**
Fisher's exact test on Spectral Counts (Benchmark)	0.70	0.70	0.70	0.70	0.64	0.62
empirical Bayes on PPIDs	0.73	0.72	0.72	0.71	0.69	0.67
empirical Bayes on PPIDs + missed PPID weighting	0.71	0.72	0.72	0.71	0.68	0.67
empirical Bayes on PPIDs + observed PPID weighting	0.70	0.70	0.69	0.70	0.69	0.67
empirical Bayes on PPIDs + log normalization	0.70	0.69	0.73	0.71	0.70	0.67
empirical Bayes on PPIDs + log normalization + missed PPID weighting	0.69	0.68	0.71	0.71	0.70	0.67
empirical Bayes on PPIDs + log normalization + observed PPID weighting	0.68	0.69	0.72	0.71	0.69	0.68
empirical Bayes on PPIDs + percentile normalization	0.74	0.72	0.74	0.72	0.69	0.67
empirical Bayes on PPIDs + percentile normalization + missed PPID weighting	0.74	0.72	0.75	0.73	0.70	0.68
empirical Bayes on PPIDs + percentile normalization + observed PPID weighting	0.74	0.72	0.74	0.73	0.72	0.69
empirical Bayes on PPIDs + percentile+arcsine(sqrt) normalization	0.72	0.70	0.74	0.72	0.69	0.67
empirical Bayes on PPIDs + percentile+arcsine normalization + missed PPID weighting	0.72	0.71	0.75	0.73	0.71	0.68
empirical Bayes on PPIDs + percentile+arcsine normalization + observed PPID weighting	0.71	0.72	0.75	0.73	0.72	0.69
empirical Bayes on PPIDs + max value normalization	0.61	0.62	0.61	0.62	0.62	0.61

empirical Bayes on PPIDs + max value normalization + missed PPID weighting	0.61	0.61	0.61	0.61	0.62	0.61
empirical Bayes on PPIDs + max value normalization + observed PPID weighting	0.58	0.57	0.59	0.60	0.61	0.59
empirical Bayes on PPIDs + max value normalization + log transformation	0.73	0.69	0.69	0.68	0.66	0.64
empirical Bayes on PPIDs + max value normalization+log transformation + missed PPID weighting	0.72	0.68	0.67	0.65	0.65	0.64
empirical Bayes on PPIDs + max value normalization+log transformation + observed PPID weighting	0.72	0.67	0.69	0.66	0.66	0.65
empirical Bayes on PPIDs + max value normalization+arcsine(sqrt)	0.60	0.59	0.62	0.62	0.63	0.61
empirical Bayes on PPIDs + max value normalization+arcsine(sqrt) + missed PPID weighting	0.60	0.58	0.62	0.61	0.62	0.60
empirical Bayes on PPIDs + max value normalization+arcsine(sqrt) + observed PPID weighting	0.59	0.57	0.59	0.59	0.60	0.58
empirical Bayes on Spectral Counts	0.73	0.74	0.75	0.74	0.73	0.69

Table S2.8 F1-Measures For Combined Model (Empirical Bayes t-test + Benchmark) Using the "Minimum Spectral Count/PPID across both cohorts" Filter

	Minimum Spectral Count/PPID across both cohorts					
Candidate Test	**0**	**5**	**10**	**15**	**20**	**25**
Fisher's exact test on Spectral Counts (Benchmark)	0.70	0.70	0.70	0.70	0.70	0.70
empirical Bayes on PPIDs	0.73	0.71	0.71	0.72	0.71	0.68
empirical Bayes on PPIDs + missed PPID weighting	0.71	0.71	0.72	0.71	0.70	0.68
empirical Bayes on PPIDs + observed PPID weighting	0.70	0.70	0.71	0.69	0.68	0.67
empirical Bayes on PPIDs + log normalization	0.70	0.70	0.73	0.72	0.72	0.69
empirical Bayes on PPIDs + log normalization + missed PPID weighting	0.69	0.69	0.72	0.70	0.70	0.68
empirical Bayes on PPIDs + log normalization + observed PPID weighting	0.68	0.69	0.72	0.70	0.71	0.68
empirical Bayes on PPIDs + percentile normalization	0.74	0.72	0.74	0.74	0.71	0.68
empirical Bayes on PPIDs + percentile normalization + missed	0.74	0.73	0.74	0.74	0.71	0.70

PPID weighting						
empirical Bayes on PPIDs + percentile normalization + observed PPID weighting	0.74	0.73	0.73	0.73	0.71	0.70
empirical Bayes on PPIDs + percentile+arcsine(sqrt) normalization	0.72	0.72	0.73	0.73	0.72	0.69
empirical Bayes on PPIDs + percentile+arcsine normalization + missed PPID weighting	0.72	0.72	0.74	0.74	0.72	0.71
empirical Bayes on PPIDs + percentile+arcsine normalization + observed PPID weighting	0.71	0.73	0.72	0.71	0.71	0.70
empirical Bayes on PPIDs + max value normalization	0.61	0.62	0.61	0.61	0.61	0.60
empirical Bayes on PPIDs + max value normalization + missed PPID weighting	0.61	0.60	0.60	0.61	0.60	0.60
empirical Bayes on PPIDs + max value normalization + observed PPID weighting	0.58	0.57	0.58	0.58	0.59	0.58
empirical Bayes on PPIDs + max value normalization + log transformation	0.73	0.68	0.71	0.70	0.68	0.66
empirical Bayes on PPIDs + max value normalization+log transformation + missed PPID weighting	0.72	0.69	0.69	0.68	0.66	0.64
empirical Bayes on PPIDs + max value normalization+log transformation + observed PPID weighting	0.72	0.69	0.68	0.67	0.67	0.65
empirical Bayes on PPIDs + max value normalization+arcsine(sqrt)	0.60	0.60	0.61	0.61	0.61	0.60
empirical Bayes on PPIDs + max value normalization+arcsine(sqrt) + missed PPID weighting	0.60	0.59	0.61	0.61	0.60	0.59
empirical Bayes on PPIDs + max value normalization+arcsine(sqrt) + observed PPID weighting	0.59	0.58	0.59	0.58	0.58	0.57
empirical Bayes on Spectral Counts	0.73	0.73	0.73	0.73	0.73	0.73

Supplemental File 3. ROC curve AUCs for Candidate Tests

Table S3.1 AUCs For Classical Tests Using the "Minimum Spectral Count/PPID in at least one cohort" filter

	Minimum Spectral Count/PPID in at least one cohort					
Candidate Test	**0**	**5**	**10**	**15**	**20**	**25**
Fisher's exact test on Spectral Counts (Benchmark)	0.85	0.88	0.87	0.85	0.82	0.81
Student's t-test on PPIDs	0.87	0.88	0.87	0.83	0.80	0.77
Student's t-test log transformation on PPIDs	0.86	0.88	0.86	0.83	0.81	0.77
Welch's t-test on PPIDs	0.87	0.88	0.86	0.83	0.80	0.77
Welch's t-test log transformation on PPIDs	0.86	0.87	0.86	0.82	0.80	0.76
Mann Whitney U test on Spectral Counts	0.84	0.88	0.89	0.87	0.84	0.82
Modified Mann Whitney U test on Spectral Counts	0.84	0.88	0.89	0.87	0.83	0.82
Mann Whitney U test on PPIDs	0.84	0.88	0.88	0.86	0.84	0.82

Table S3.2 AUCs For Classical Tests Using the "Minimum Spectral Count/PPID across both cohorts" filter

	Minimum Spectral Count/PPID across both cohorts					
Candidate Test	**0**	**5**	**10**	**15**	**20**	**25**
Fisher's exact test on Spectral Counts (Benchmark)	0.85	0.85	0.85	0.85	0.85	0.85
Student's t-test on PPIDs	0.87	0.89	0.88	0.83	0.82	0.79
Student's t-test log transformation on PPIDs	0.86	0.88	0.87	0.83	0.81	0.79
Welch's t-test on PPIDs	0.87	0.88	0.86	0.83	0.81	0.84
Welch's t-test log transformation on PPIDs	0.86	0.87	0.86	0.83	0.81	0.79
Mann Whitney U test on Spectral Counts	0.84	0.84	0.84	0.84	0.84	0.84
Modified Mann Whitney U test on Spectral Counts	0.84	0.84	0.84	0.84	0.84	0.84
Mann Whitney U test on PPIDs	0.84	0.85	0.85	0.86	0.86	0.83

Table S3.3 AUCs For Combined Model (Classical Tests + Benchmark) Using the "Minimum Spectral Count/PPID at least one cohort" Filter

	Minimum Spectral Count/PPID in at least one cohort					
Candidate Test	**0**	**5**	**10**	**15**	**20**	**25**
Fisher's exact test on Spectral Counts (Benchmark)	0.85	0.88	0.87	0.85	0.82	0.81
Student's t-test on PPIDs	0.89	0.90	0.90	0.90	0.89	0.89
Student's t-test log transformation on PPIDs	0.88	0.90	0.90	0.90	0.90	0.89
Welch's t-test on PPIDs	0.89	0.90	0.90	0.89	0.90	0.89
Welch's t-test log transformation on PPIDs	0.88	0.90	0.90	0.90	0.90	0.89
Mann Whitney U test on Spectral Counts	0.85	0.89	0.90	0.90	0.90	0.89
Modified Mann Whitney U test on Spectral Counts	0.85	0.89	0.90	0.90	0.90	0.90
Mann Whitney U test on PPIDs	0.86	0.90	0.90	0.90	0.90	0.89

Table S3.4 AUCs For Combined Model (Classical Tests + Benchmark) Using the "Minimum Spectral Count/PPID at across both cohorts" Filter

	Minimum Spectral Count/PPID across both cohorts					
Candidate Test	**0**	**5**	**10**	**15**	**20**	**25**
Fisher's exact test on Spectral Counts (Benchmark)	0.85	0.88	0.87	0.85	0.82	0.81
Student's t-test on PPIDs	0.89	0.90	0.88	0.83	0.82	0.79
Student's t-test log transformation on PPIDs	0.88	0.90	0.90	0.90	0.90	0.89
Welch's t-test on PPIDs	0.89	0.90	0.88	0.83	0.82	0.79
Welch's t-test log transformation on PPIDs	0.88	0.90	0.90	0.90	0.90	0.90
Mann Whitney U test on Spectral Counts	0.85	0.85	0.85	0.85	0.85	0.85
Modified Mann Whitney U test on Spectral Counts	0.85	0.85	0.85	0.85	0.85	0.85
Mann Whitney U test on PPIDs	0.86	0.86	0.87	0.87	0.87	0.86

Table S3.5 AUCs For Empirical Bayes t-test Using the "Minimum Spectral Count/PPID in at least one cohort" filter

	Minimum Spectral Count/PPID in at last one cohort					
Candidate Test	**0**	**5**	**10**	**15**	**20**	**25**
Fisher's exact test on Spectral Counts (Benchmark)	0.85	0.88	0.87	0.85	0.82	0.81
empirical Bayes on PPIDs	0.88	0.89	0.87	0.84	0.81	0.78
empirical Bayes on PPIDs + missed PPID weighting	0.87	0.89	0.87	0.84	0.81	0.78

empirical Bayes on PPIDs + observed PPID weighting	0.88	0.89	0.87	0.84	0.81	0.78
empirical Bayes on PPIDs + log normalization	0.86	0.88	0.87	0.83	0.81	0.78
empirical Bayes on PPIDs + log normalization + missed PPID weighting	0.85	0.88	0.87	0.83	0.81	0.78
empirical Bayes on PPIDs + log normalization + observed PPID weighting	0.86	0.88	0.87	0.83	0.81	0.78
empirical Bayes on PPIDs + percentile normalization	0.85	0.88	0.87	0.83	0.81	0.78
empirical Bayes on PPIDs + percentile normalization + missed PPID weighting	0.85	0.88	0.87	0.84	0.81	0.78
empirical Bayes on PPIDs + percentile normalization + observed PPID weighting	0.85	0.87	0.86	0.83	0.81	0.78
empirical Bayes on PPIDs + percentile+arcsine(sqrt) normalization	0.84	0.88	0.87	0.83	0.81	0.78
empirical Bayes on PPIDs + percentile+arcsine normalization + missed PPID weighting	0.84	0.89	0.87	0.84	0.81	0.78
empirical Bayes on PPIDs + percentile+arcsine normalization + observed PPID weighting	0.83	0.88	0.87	0.83	0.81	0.78
empirical Bayes on PPIDs + max value normalization	0.86	0.87	0.86	0.83	0.81	0.78
empirical Bayes on PPIDs + max value normalization + missed PPID weighting	0.86	0.87	0.86	0.83	0.81	0.78
empirical Bayes on PPIDs + max value normalization + observed PPID weighting	0.87	0.87	0.86	0.83	0.81	0.78
empirical Bayes on PPIDs + max value normalization + log transformation	0.87	0.88	0.86	0.83	0.81	0.78
empirical Bayes on PPIDs + max value normalization+log transformation + missed PPID weighting	0.87	0.88	0.86	0.83	0.81	0.78
empirical Bayes on PPIDs + max value normalization+log transformation + observed PPID weighting	0.86	0.87	0.86	0.83	0.81	0.78
empirical Bayes on PPIDs + max value normalization+arcsine(sqrt)	0.88	0.89	0.87	0.84	0.81	0.78
empirical Bayes on PPIDs + max value normalization+arcsine(sqrt) + missed PPID weighting	0.88	0.89	0.87	0.84	0.81	0.78
empirical Bayes on PPIDs + max value normalization+arcsine(sqrt) + observed PPID weighting	0.88	0.89	0.87	0.84	0.81	0.78

empirical Bayes on Spectral Counts	0.86	0.89	0.87	0.84	0.81	0.78

Table S3.6 AUCs For Empirical Bayes t-test Using the "Minimum Spectral Count/PPID across both cohorts" filter

	Minimum Spectral Count/PPID across both cohorts					
candidate test	**0**	**5**	**10**	**15**	**20**	**25**
Fisher's exact test on Spectral Counts (Benchmark)	0.85	0.85	0.85	0.85	0.85	0.85
empirical Bayes on PPIDs	0.88	0.89	0.88	0.83	0.82	0.79
empirical Bayes on PPIDs + missed PPID weighting	0.87	0.89	0.88	0.83	0.81	0.79
empirical Bayes on PPIDs + observed PPID weighting	0.88	0.89	0.88	0.83	0.82	0.79
empirical Bayes on PPIDs + log normalization	0.86	0.88	0.87	0.83	0.81	0.78
empirical Bayes on PPIDs + log normalization + missed PPID weighting	0.85	0.88	0.87	0.83	0.81	0.78
empirical Bayes on PPIDs + log normalization + observed PPID weighting	0.86	0.88	0.87	0.83	0.81	0.78
empirical Bayes on PPIDs + percentile normalization	0.85	0.88	0.87	0.83	0.81	0.78
empirical Bayes on PPIDs + percentile normalization + missed PPID weighting	0.85	0.88	0.87	0.83	0.81	0.78
empirical Bayes on PPIDs + percentile normalization + observed PPID weighting	0.85	0.87	0.87	0.82	0.81	0.78
empirical Bayes on PPIDs + percentile+arcsine(sqrt) normalization	0.84	0.85	0.86	0.87	0.87	0.84
empirical Bayes on PPIDs + percentile+arcsine normalization + missed PPID weighting	0.84	0.84	0.85	0.86	0.86	0.83
empirical Bayes on PPIDs + percentile+arcsine normalization + observed PPID weighting	0.83	0.83	0.84	0.85	0.86	0.83
empirical Bayes on PPIDs + max value normalization	0.86	0.87	0.86	0.83	0.81	0.79
empirical Bayes on PPIDs + max value normalization + missed PPID weighting	0.86	0.87	0.86	0.82	0.81	0.78
empirical Bayes on PPIDs + max value normalization + observed PPID weighting	0.87	0.87	0.86	0.83	0.81	0.79

empirical Bayes on PPIDs + max value normalization + log transformation	0.87	0.88	0.87	0.82	0.81	0.78
empirical Bayes on PPIDs + max value normalization+log transformation + missed PPID weighting	0.87	0.88	0.87	0.82	0.81	0.78
empirical Bayes on PPIDs + max value normalization+log transformation + observed PPID weighting	0.86	0.87	0.86	0.82	0.81	0.78
empirical Bayes on PPIDs + max value normalization+arcsine(sqrt)	0.88	0.89	0.88	0.83	0.82	0.79
empirical Bayes on PPIDs + max value normalization+arcsine(sqrt) + missed PPID weighting	0.88	0.89	0.88	0.83	0.81	0.79
empirical Bayes on PPIDs + max value normalization+arcsine(sqrt) + observed PPID weighting	0.88	0.89	0.87	0.83	0.82	0.79
empirical Bayes on Spectral Counts	0.86	0.86	0.86	0.86	0.86	0.86

Table S3.7 AUCs For Combined Model (Empirical Bayes t-test + Benchmark) Using the "Minimum Spectral Count/PPID in at least cohorts" filter

	Minimum Spectral Count/PPID in at last one cohort					
candidate test	**0**	**5**	**10**	**15**	**20**	**25**
Fisher's exact test on Spectral Counts (Benchmark)	0.85	0.88	0.87	0.85	0.82	0.81
empirical Bayes on PPIDs	0.89	0.91	0.91	0.90	0.90	0.90
empirical Bayes on PPIDs + missed PPID weighting	0.89	0.91	0.90	0.90	0.90	0.90
empirical Bayes on PPIDs + observed PPID weighting	0.89	0.91	0.90	0.90	0.90	0.90
empirical Bayes on PPIDs + log normalization	0.87	0.90	0.90	0.90	0.90	0.90
empirical Bayes on PPIDs + log normalization + missed PPID weighting	0.87	0.90	0.90	0.90	0.90	0.90
empirical Bayes on PPIDs + log normalization + observed PPID weighting	0.87	0.90	0.90	0.90	0.90	0.90
empirical Bayes on PPIDs + percentile normalization	0.87	0.90	0.90	0.90	0.90	0.89
empirical Bayes on PPIDs + percentile normalization + missed PPID weighting	0.87	0.90	0.90	0.90	0.90	0.89
empirical Bayes on PPIDs + percentile normalization + observed PPID weighting	0.86	0.89	0.90	0.90	0.90	0.89
empirical Bayes on PPIDs + percentile+arcsine(sqrt) normalization	0.86	0.90	0.90	0.90	0.90	0.89

empirical Bayes on PPIDs + percentile+arcsine normalization + missed PPID weighting	0.85	0.90	0.90	0.90	0.90	0.89
empirical Bayes on PPIDs + percentile+arcsine normalization + observed PPID weighting	0.85	0.89	0.90	0.90	0.90	0.89
empirical Bayes on PPIDs + max value normalization	0.89	0.90	0.90	0.90	0.89	0.89
empirical Bayes on PPIDs + max value normalization + missed PPID weighting	0.89	0.90	0.90	0.90	0.89	0.89
empirical Bayes on PPIDs + max value normalization + observed PPID weighting	0.89	0.90	0.90	0.89	0.89	0.89
empirical Bayes on PPIDs + max value normalization + log transformation	0.89	0.91	0.90	0.90	0.90	0.89
empirical Bayes on PPIDs + max value normalization+log transformation + missed PPID weighting	0.90	0.91	0.90	0.90	0.90	0.89
empirical Bayes on PPIDs + max value normalization+log transformation + observed PPID weighting	0.89	0.90	0.90	0.90	0.90	0.89
empirical Bayes on PPIDs + max value normalization+arcsine(sqrt)	0.90	0.91	0.90	0.90	0.89	0.89
empirical Bayes on PPIDs + max value normalization+arcsine(sqrt) + missed PPID weighting	0.90	0.91	0.90	0.90	0.89	0.89
empirical Bayes on PPIDs + max value normalization+arcsine(sqrt) + observed PPID weighting	0.89	0.90	0.90	0.89	0.89	0.89
empirical Bayes on Spectral Counts	0.86	0.90	0.90	0.90	0.90	0.90

Table S3.8 AUCs For Combined Model (Empirical Bayes t-test + Benchmark) Using the "Minimum Spectral Count/PPID across both cohorts" filter

	Minimum Spectral Count/PPID across both cohorts					
candidate test	**0**	**5**	**10**	**15**	**20**	**25**
Fisher's exact test on Spectral Counts (Benchmark)	0.85	0.88	0.87	0.85	0.82	0.81
empirical Bayes on PPIDs	0.89	0.91	0.91	0.90	0.90	0.90
empirical Bayes on PPIDs + missed PPID weighting	0.89	0.91	0.90	0.90	0.90	0.90
empirical Bayes on PPIDs + observed PPID weighting	0.89	0.91	0.90	0.90	0.90	0.90
empirical Bayes on PPIDs + log normalization	0.87	0.90	0.90	0.90	0.90	0.90
empirical Bayes on PPIDs + log normalization + missed PPID weighting	0.87	0.90	0.90	0.90	0.90	0.90

empirical Bayes on PPIDs + log normalization + observed PPID weighting	0.87	0.90	0.90	0.90	0.90	0.90
empirical Bayes on PPIDs + percentile normalization	0.87	0.90	0.90	0.90	0.90	0.89
empirical Bayes on PPIDs + percentile normalization + missed PPID weighting	0.87	0.90	0.90	0.90	0.90	0.89
empirical Bayes on PPIDs + percentile normalization + observed PPID weighting	0.86	0.89	0.90	0.90	0.90	0.89
empirical Bayes on PPIDs + percentile+arcsine(sqrt) normalization	0.86	0.90	0.90	0.90	0.90	0.89
empirical Bayes on PPIDs + percentile+arcsine normalization + missed PPID weighting	0.85	0.90	0.90	0.90	0.90	0.89
empirical Bayes on PPIDs + percentile+arcsine normalization + observed PPID weighting	0.85	0.89	0.90	0.90	0.90	0.89
empirical Bayes on PPIDs + max value normalization	0.89	0.90	0.90	0.90	0.89	0.89
empirical Bayes on PPIDs + max value normalization + missed PPID weighting	0.89	0.90	0.90	0.90	0.89	0.89
empirical Bayes on PPIDs + max value normalization + observed PPID weighting	0.89	0.90	0.90	0.89	0.89	0.89
empirical Bayes on PPIDs + max value normalization + log transformation	0.89	0.91	0.90	0.90	0.90	0.89
empirical Bayes on PPIDs + max value normalization+log transformation + missed PPID weighting	0.90	0.91	0.90	0.90	0.90	0.89
empirical Bayes on PPIDs + max value normalization+log transformation + observed PPID weighting	0.89	0.90	0.90	0.90	0.90	0.89
empirical Bayes on PPIDs + max value normalization+arcsine(sqrt)	0.90	0.91	0.90	0.90	0.89	0.89
empirical Bayes on PPIDs + max value normalization+arcsine(sqrt) + missed PPID weighting	0.90	0.91	0.90	0.90	0.89	0.89
empirical Bayes on PPIDs + max value normalization+arcsine(sqrt) + observed PPID weighting	0.89	0.90	0.90	0.89	0.89	0.89
empirical Bayes on Spectral Counts	0.86	0.90	0.90	0.90	0.90	0.90

Supplemental File 4. Analysis Config File

When run from the command line, IDPQuantify takes a single argument, "-idpAnalysisConfigFile", that specifies the pathway to analysis config file. Users configure IDPQuantify using the "analysis_config" file. Each line in the file contains an attribute and a value in the following format:

-attribute_name "attribute_value"

There are two types of attributes, Global Attributes and Per Statistical Run Attributes. Global Attributes are only defined once. Per Statistical Run Attributes are specified for each separate R-script to be run. When Per Statistical Run Attributes are declared in the analysis_config file the attribute name should be followed by a unique number.

The attributes are as follows:

Table S4.1 Global attribute name and values for 'analysis_config' file

Attribute Name	Attribute Meaning
resultOutputFilesRootFolder	Pathway to folder where files will be created.
resultFilesConfigFileFullPath	Pathway to 'files_config' file defining the two cohorts to be analyzed (See Supplemental File 5).
resultSpcCntPerPepPerRep	File name of file written by IDPQuantify listing the spectral counts per peptide per replicate.
resultPPIDPerPepUnnorm	File name of file written by IDPQuantify listing the PPID per peptide per replicate.
resultPPIDObservedPerPepPerRep	File name of file written by IDPQuantify listing the presence or absence (1 or 0) of at least one PPID observed per peptide per replicate.
resultPPIDMissingPerPepPerRep	File name of file written by IDPQuantify listing a "1" if IDPQuantify expected a PPID but could not find one per peptide per replicate. Otherwise a "0" is listed.
resultPPIDPerPep_missingPPIDWeight	File name of file written by IDPQuantify containing the PPID per peptide weighted according to missing PPIDs.
resultPPIDPerPep_foundPPIDWeight	File name of file written by IDPQuantify containing the PPID per peptide weighted according to observed PPIDs.
resultPPIDPerPep_unweighted	File name of file written by IDPQuantify containing the PPID per peptide with no weighting.
resultPPIDPerPepGrp_missingPPIDWeight	File name of file written by IDPQuantify containing the PPID per peptide group weighted according to missing PPIDs.
resultPPIDPerPepGrp_foundPPIDWeight	File name of file written by IDPQuantify containing the PPID per peptide group weighted according to observed PPIDs.
resultPPIDPerPepGrp_unweighted	File name of file written by IDPQuantify containing the

	PPID per peptide group with no weighting.
resultSpcCntPerPepGrp	File name of file written by IDPQuantify containing the Spectral Count per peptide group.
resultPPIDCountPerPepGrp	File name of file written by IDPQuantify containing the number of PPIDs observed per peptide group.
resultPPIDPercentileNorm_foundPPIDWeight	File name of file written by IDPQuantify containing the percentile normalized PPID per peptide group weighted according to observed PPIDs.
resultPPIDPercentileNorm_missingPPIDWeight	File name of file written by IDPQuantify containing the percentile normalized PPID per peptide group weighted according to missing PPIDs.
resultPPIDPercentileNorm_noPPIDWeight	File name of file written by IDPQuantify containing the percentile normalized PPID per peptide group with no weighting.
resultSpcCntMinusPPIDPerPepGrp	File name of file written by IDPQuantify containing the number of observed spectral counts, observed PPIDs, and missing PPIDs per peptide group per cohort.
resultUniqPepSeqPerPepGrp	File name of file written by IDPQuantify containing the number of unique peptide sequence per peptide group.
resultNumPepSeqPerPepGrp	File name of file written by IDPQuantify containing the number of total peptide sequences, including duplicate sequences, per peptide group. Sequences are summed per replicate per peptide group.
resultPhenotypeDataFileForRCohort	The filename of the "phenotype.txt" data file used for R scripts that utilize the Limma package in R. The pathway of this file is fed to each auto-run R script as one of the parameters.
pathwayToRScriptExe	Pathway to the RScript.exe executable.
pathwayToIDPProtGrpToPepGrpFiles	Pathway to IDPicker result file containing the protein group-to-peptide group mapping.

Table S4.2 Per Statistical Run Attribute Name and Values for 'analysis_config' File

Attribute Name	Attribute Meaning
resultRpvalsfile#	File name of file name fed by IDPQuantify into R script for the resulting statistical test output file.
resultRpvalsPPIDFile#	Filename of quantitative data file to be fed into R script for statistical analysis. Can be PPID or spectral count file generated by IDPQuantify (see Table S4.1).
resultRpvalsfileTestLabel#	Label to be used to describe this statistical test run in summary files and constructing output file names.
resultPvalRScript#	Pathway to R script to be run for this statistical run

Supplemental File 5. Files Config file

Users specify the location of IDPicker result files and spectral files (e.g. .RAW files) using the "files_config" file. Each line in the file contains an attribute and a value in the following format:

-attribute_name "attribute_value"

There are two types of attributes, Global Attributes and Per Replicate Attributes. Global Attributes are only defined once. Per Replicate Attributes are specified for each separate MS run that is to be included in one of the two cohorts. When Per Replicate Attributes are declared in the analysis_config file the attribute name should be followed by a unique number.

Table S5.1 Global Attribute Name and Values for 'files_config' File

Attribute Name	Attribute Meaning
idpQuanCohort1	A comma-delimited list of numbers corresponding to the unique identifiers for each MS run in cohort 1. The unique identifiers correspond to the "#" following the "Per Replicate" attributes for a given replicate (See Table S5.2).
idpQuanCohort2	A comma-delimited list of numbers corresponding to the unique identifiers for each MS run in cohort 1. The unique identifiers correspond to the "#" following the "Per Replicate" attributes for a given replicate (See Table S5.2).
hardklorPathway	Pathway to Hardklör executable.
bullseyePathway	Pathway to Bullseye executable.

Table S5.2 Per Replicate Attribute Name and Values for 'files_config' File

Attribute Name	Attribute Meaning
idpXMLFile#	Pathway to .idpXML for this replicate.
hkFilePath#	Pathway to Hardklör output file for this replicate. If file does not exist, IDPQuantify will attempt to create it when run.
bullseyeFoundFile#	Pathway to Bullseye output file of MS/MS spectra with matching PPIDs for this replicate. If file does not exist, IDPQuantify will attempt to create it when run.
bullseyeMissedFile#	Pathway to Bullseye output file of MS/MS spectra with missing PPIDs for this replicate. If file does not exist, IDPQuantify will attempt to create it when run.
spectralInputFile#	Pathway to spectral file (e.g. .RAW) for this replicate.
bullseyeMS2DataFile#	Pathway to file containing MS/MS data for use when generating Bullseye files.
hardklorMS2DataFile#	Pathway to file containing MS data for use when generating Hardklör files.
idpQuan#	The pathway to the .idpQuan file generated by IDPQuantify containing PPIDs per peptide group.