Integrative analysis frameworks for improved peptide and protein identifications from tandem mass spectrometry data

by

Avinash Kumar Shanmugam

A dissertation submitted in partial fulfilment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2015

Doctoral Committee:

Associate Professor Alexey I. Nesvizhskii, Chair
Professor Philip C. Andrews
Assistant Professor Yuanfang Guan
Assistant Professor Hui Jiang
Associate Professor Jun Li

Table of Contents

List of Figures

List of Tables

CHAPTER 1
Introduction

BACKGROUND

Tandem Mass Spectrometry

Mass spectrometry refers to a range of analysis techniques used to identify the chemical composition of a sample by measuring the mass-to-charge (m/z) ratio of its charged components. Mass spectrometry applied to identification of proteins was originally performed using a technique called Peptide mass fingerprinting (PMF)[1]. PMF involves enzymatic digestion of the protein, typically using trypsin, into smaller peptides, followed by analysis of these in a mass spectrometer to measure their m/z values. The resulting ‘mass spectrum’ can be used to generate a list of candidate proteins or in some simple cases identify the protein directly. But, in cases where the list of potential proteins is large, PMF might not provide specific identifications of proteins.

Tandem mass spectrometry, also referred to as MS/MS, is an improvement on the PMF technique. In Tandem MS, after measurement of the m/z values of peptides, some of the peptides are isolated, subjected to fragmentation in a collision cell and the m/z values of the resulting fragments measured. As the process of fragmentation of a peptide in the collision cell is fairly well understood, measurement of the fragmentation mass spectrum of a peptide allows for identification of its sequence using techniques such as *de novo* peptide sequencing[2], Database searching[3] or Spectral library searching[4]. For ease of reference, the first MS analysis cycle that analyzes peptides is called the MS1 scan, and the resulting spectrum the MS1 spectrum, while the second analysis cycle analyzing fragments is called the MS2 or MS/MS scan with its resultant MS2 or MS/MS spectra.

Shotgun proteomics

Tandem mass spectrometry by itself can allow accurate identifications of simple samples of proteins. However, to achieve high throughput identification of proteins from complex mixtures of, it needs to be coupled with a workflow of separation techniques and informatics pipelines. Such a workflow is referred to as shotgun proteomics[5]. The initial preparatory steps of shotgun proteomics, involve enzymatic digestion of complex protein mixtures and separation of the resulting peptides, typically using liquid chromatography, resulting in the loss of connections between protein and their constituent peptides. The computational analysis portion of it focuses on rebuilding these connections, identifying peptides and mapping them to their parent proteins using knowledge of the expected protein sequences in the sample (often the proteome of the organism under study). In this way it is analogous to its namesake ‘shotgun sequencing’, used in genomics.

Informatics pipelines for shotgun proteomics

Informatics pipelines, as mentioned above, are critical in allowing shotgun proteomics workflows to identify proteins in a high throughput manner from complex mixtures. While there are multiple informatics pipelines in use by the proteomics community, almost all of them can be seen to have three main components (i) Spectrum identification (ii) PSM (peptide to spectrum match) validation and (iii) Protein inference. Each of these components have been implemented as different software tools, most of which are inter-operable with each other allowing for mixing and matching of the various components.

Spectrum identification

As mentioned before, there are several methods to perform peptide identification from MS/MS spectra. But database searching is the most widely used approach, by far, and was the method of choice for the analyses described in this dissertation. Database searching is a process of spectral matching by comparing the experimental spectra obtained from tandem mass spectrometry to theoretical spectra generated from the sequence of all the peptides present in a protein FASTA file (often referred to as a protein database). Scoring functions, typically some form of dot product between the experimental and theoretical spectra, are used to compute a match score for each peptide to spectrum match (PSM) and identify the top scoring peptide match for each spectrum. Apart from the match score, most database search engines also include some form of significance score, such as an expect value or delta score, which captures how much better the score of the top match is compared to other peptide matches to the same spectrum. Commonly used database search engines include X!Tandem[6], MS-GF+[7], Comet[8], Sequest[9] and Mascot[10], of which X!Tandem and MS-GF+ were used for the analyses described in this dissertation.

PSM validation

The peptide to spectrum matches (PSMs) reported by the database search engines provide the best matching peptide to each spectrum in the MS/MS data. However, the best match might not always be a correct match. For instance, if the correct peptide for a particular spectrum is not present in our protein database, even the best matching peptide to that spectrum would only be a false positive match. The PSM validation step helps us to distinguish between significant, high confidence PSMs and spurious, lower confidence PSMs. PSM validation methods use a variety of methods, such as mixture modelling or support vector machines, to re-scale and convert the spectrum match scores and significance scores from the database search into identification probabilities, i.e. probability of the peptide to spectrum match being a true positive identification. Popular PSM validation tools include PeptideProphet[11] and Percolator[12], with PeptideProphet being used for the analyses in this dissertation.

Protein inference

During protein inference[13], peptides that were previously identified in the proteome informatics pipeline are mapped to protein sequences to infer the proteins present in the sample. Since the same peptides often map to more than one protein, the results of protein inference are typically presented as protein groups, sets of proteins that share peptides. Protein groups may be distinguishable, with the proteins in the group sharing peptides but also having their individual unique peptides, or indistinguishable, when all proteins in the group are only identified by the shared peptides. A protein identification probability based on the probabilities of peptides mapped to each protein is also typically calculated during protein inference. ProteinProphet[14] and

IDpicker[15] are examples of software tools that perform protein inference, with ProteinProphet the software of choice for the analyses described in this dissertation.

False Discovery Rate estimation: Target-Decoy strategy

While the peptide and protein probabilities estimated in the proteome informatics pipelines can provide a reasonable measure of the accuracy of the identifications, they are not perfect. False discovery rates are a more commonly used metric in the proteomics community to measure accuracy of identifications. An FDR value of 1% or 5% is typically used as the threshold above which peptide or protein identifications are considered confident and acceptable. The most widely used method for estimating the false discovery rate is the target-decoy strategy[16].

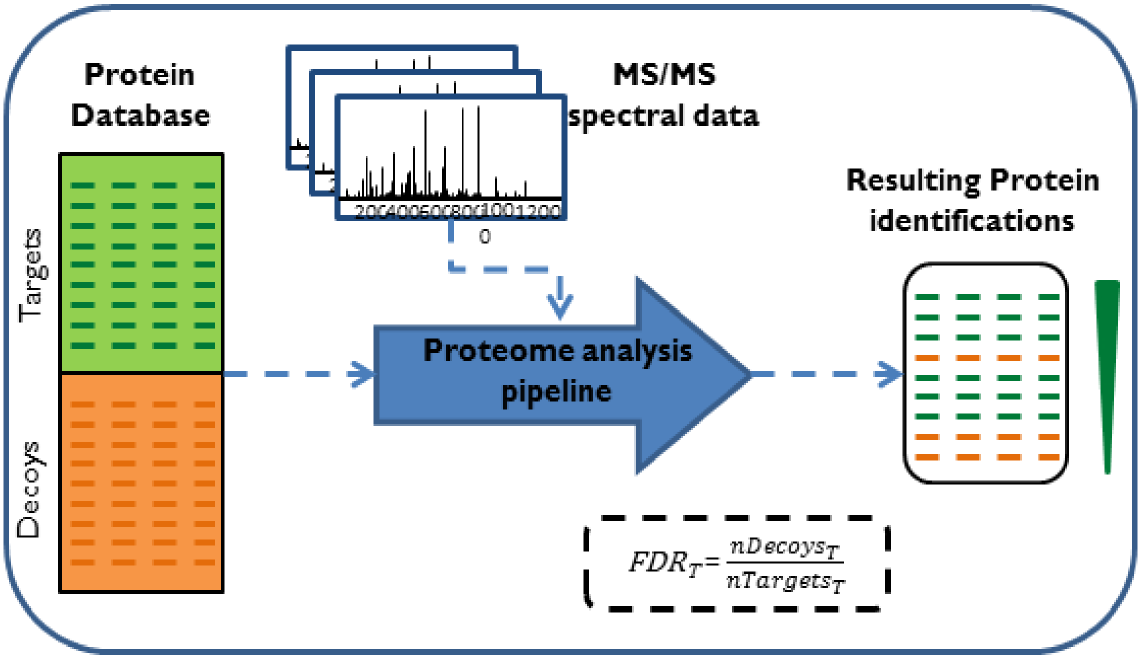

Protein Database
Targets
Decoys
MS/MS spectral data
200400600800100120
0
Proteome analysis pipeline
Resulting Protein identifications
$FDR_T = \frac{nDecoys_T}{nTargets_T}$

Figure 1.1: Target-decoy strategy for estimating error rates. Decoys or artificial sequences expected to not be present in the protein sample are included in the search database and processed through the proteome analysis pipeline as a way to estimate the rate of random / false positive matches.

Decoy sequences are artificial sequences that are not expected to be present in the protein sample, and are most commonly generated by reversing the target sequences in the database or sometimes by randomly shuffling them. In the target-decoy strategy, as illustrated in Figure 1.1, decoy sequences are appended to the database (FASTA file) of protein sequences, or targets, which is used for database searching and processed through the entire proteomic informatics pipeline. Since it is known that they are artificial sequences, any peptide or protein identifications from the decoys are likely false identification, presumably due to random spectral matching. Provided the size of the space of decoy sequences and that of the target sequences are the same, the number of decoys identified can be used as an estimate of the number of false positive target identifications at the same score level. Based on this, the FDR at a given threshold, T, can be estimated as $FDR_T = \frac{nDecoys_T}{nTargets_T}$.

MOTIVATION

While the sophisticated algorithms for spectral matching and analysis described above have allowed for high throughput peptide and protein identification, they can still be hampered by issues such as low efficiency of peptide ionization, low quality or noisy spectra, dynamic range of protein abundances and the complexity of protein samples[17] resulting in a loss of sensitivity. To deal with such issues, there have been continued attempts to incorporate additional information about the MS/MS experiment into analysis pipelines, such as peptide chromatographic retention time[18], pI[19] or mass accuracy[20], some of which are now a routine part of many proteomic analysis pipelines[18]. Studies have also investigated using matching MS2 and

MS3 information[21], match scores from multiple search engines[22,23] and various other information sources to re-score or adjust protein identification probability.

The above mentioned methods work to integrate additional information from the tandem MS experiment itself into the analysis pipeline. But considering the wealth of information available for most biological systems, integrating external, orthogonal information into proteomics analysis pipelines can be a very useful way of providing improved sensitivity in peptide and protein identification. There have already been studies that have investigated integration of some types of orthogonal information such as microarray data[24], protein-protein interaction networks[25] or gene functional networks[26].

In this dissertation, the utility of integrating orthogonal information and the best approaches for performing such integration has been analyzed, specifically for the integration of RNAseq[27] transcript abundance and GPMDB identification frequency data. Further, based the results, analysis frameworks that can be used for generalized integration of most types of orthogonal information into proteomics analysis pipelines have also been developed and described.

RNAseq uses short read sequencing technologies to sequence the RNA content (transcriptome profile) of a sample[27]. Based on the central dogma of molecular biology, it is a reasonable assumption that proteins and peptides corresponding to high abundance transcripts are more likely to be found in a sample, making it a useful source of orthogonal information for proteomics analysis. The Global Proteome Machine Database (GPMDB)[28] is one of largest repositories of the results of proteomics experiments in the world. With the large volume of data aggregated in GPMDB, the frequency of identification of a protein or peptide in GPMDB serves as a surrogate measure of its propensity to be observed in a MS/MS experiment. In other words, peptides or proteins with a high GPMDB identification frequency can be reasonably assumed to be more likely of being identified in an MS/MS experiment, suggesting potential utility for improving peptide and protein identifications in proteomic experiments.

In addition to the integration of orthogonal information, the integration of data from multiple MS/MS experiments has also been investigated as an alternative approach for improving peptide and protein identifications. In particular, such an approach is geared towards improving the peptide and protein coverage of the overall proteome i.e. identifying proteins and peptides that might not have been previously identified in MS/MS experiments, exemplified in a 2015 article by Wilhelm et al[29]. However, the large datasets that are analyzed in this type of integrative analysis can cause a breakdown in the assumptions of the target-decoy strategy for FDR estimation, leading to over-estimation of FDR. Therefore, an exploration of the challenges in estimating FDR in large datasets and an investigation of the various methods described in literature for adjusting FDR estimation in large datasets are also discussed in this dissertation.

OVERVIEW OF THE DISSERTATION

This dissertation is devoted to the development and testing of frameworks for integration of orthogonal data into proteomics pipelines and also the testing of methods for accurate FDR estimation in very large datasets such as those from integrative analysis of multiple proteomics experiments.

In chapter 2 of this dissertation a re-scoring based approach for integrating orthogonal information and improving the sensitivity of protein identification is presented. Re-scoring is performed by a naive bayes model that adjusts identification probabilities of proteins based on their RNAseq transcript abundance or GPMDB identification frequency. This re-scoring boosts probabilities of proteins that have high transcript abundance or high identification frequency in GPMDB, and conversely also penalizes the probabilities of proteins with low transcript abundance or GPMDB identification frequency. This can help proteins that fall below the FDR threshold based on MS/MS evidence alone, but have enough supporting evidence from RNAseq or GPMDB that re-scoring boosts their identification probability above the FDR threshold. Thus, re-scoring is expected to improve sensitivity of protein identification from MS/MS data, particularly in small to medium sized datasets.

In chapter 3, an alternative approach to integrating orthogonal information based on search space restriction is described. This approach involves using orthogonal information to identify proteins or peptides likely to be present in a sample, creating a targeted database containing only those proteins or peptides. The smaller search space of the targeted database helps to improve sensitivity of peptide identification in database search. However in applying this approach it is important not over-restrict the search space, leading to an incomplete search space and causing loss of true identifications. With this in mind, a hybrid approach that combines search space restriction along with a Bayes’ re-scoring was also developed and tested. This hybrid approach leverages the improved sensitivity of search space restriction while also minimizing loss of peptide identifications due to incomplete search space.

Finally in chapter 4, the challenges in estimating accurate false discovery rates for very large proteomic datasets are discussed. Different methods described in literature for modifying the typical target-decoy strategy for FDR estimation to account for the large dataset sizes, namely, R-factor correction[30], the picked FDR strategy[31] and MAYU[32],were compared with each other, and also implemented in a single R script to facilitate easy comparison in further studies.

REFERENCES

(1) Henzel, W. J.; Billeci, T. M.; Stults, J. T.; Wong, S. C.; Grimley, C.; Watanabe, C. Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proc. Natl. Acad. Sci. U. S. A.* **1993**, *90* (11), 5011–5015.

(2) Frank, A.; Pevzner, P. PepNovo: De Novo Peptide Sequencing via Probabilistic Network Modeling. *Anal. Chem.* **2005**, *77* (4), 964–973.

(3) Sadygov, R. G.; Cociorva, D.; Yates, J. R. Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nat. Methods* **2004**, *1* (3), 195–202.

(4) Lam, H.; Deutsch, E. W.; Eddes, J. S.; Eng, J. K.; King, N.; Stein, S. E.; Aebersold, R. Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **2007**, *7* (5), 655–667.

(5) McDonald, W. H.; Yates, J. R. Shotgun proteomics and biomarker discovery. *Dis. Markers* **2002**, *18* (2), 99–105.

(6) Craig, R.; Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20* (9), 1466–1467.

(7) Kim, S.; Pevzner, P. A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.* **2014**, *5*, 5277.

(8) Eng, J. K.; Jahan, T. A.; Hoopmann, M. R. Comet: an open-source MS/MS sequence database search tool. *Proteomics* **2013**, *13* (1), 22–24.

(9) Eng, J. K.; McCormack, A. L.; Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **1994**, *5* (11), 976–989.

(10) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20* (18), 3551–3567.

(11) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **2002**, *74* (20), 5383–5392.

(12) Käll, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **2007**, *4* (11), 923–925.

(13) Nesvizhskii, A. I.; Aebersold, R. Interpretation of shotgun proteomic data: the protein inference problem. *Mol. Cell. Proteomics* **2005**, *4* (10), 1419–1440.

(14) Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **2003**, *75* (17), 4646–4658.

(15) Zhang, B.; Chambers, M. C.; Tabb, D. L. Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. *J. Proteome Res.* **2007**, *6* (9), 3549–3557.

(16) Elias, J. E.; Gygi, S. P. Target-decoy search strategy for mass spectrometry-based proteomics. *Methods Mol. Biol.* **2010**, *604*, 55–71.

(17) Nesvizhskii, A. I. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J. Proteomics* **2010**, *73* (11), 2092–2123.

(18) Deutsch, E. W.; Mendoza, L.; Shteynberg, D.; Farrah, T.; Lam, H.; Tasman, N.; Sun, Z.; Nilsson, E.; Pratt, B.; Prazen, B.; et al. A guided tour of the Trans-Proteomic Pipeline. *Proteomics* **2010**, *10* (6), 1150–1159.

(19) Malmström, J.; Lee, H.; Nesvizhskii, A. I.; Shteynberg, D.; Mohanty, S.; Brunner, E.; Ye, M.; Weber, G.; Eckerskorn, C.; Aebersold, R. Optimized peptide separation and identification for mass spectrometry based proteomics via free-flow electrophoresis. *J. Proteome Res.* **2006**, *5* (9), 2241–2249.

(20) Li, Y. F.; Arnold, R. J.; Li, Y.; Radivojac, P.; Sheng, Q.; Tang, H. A bayesian approach to protein inference problem in shotgun proteomics. *J. Comput. Biol.* **2009**, *16* (8), 1183–1193.

(21) Ulintz, P. J.; Bodenmiller, B.; Andrews, P. C.; Aebersold, R.; Nesvizhskii, A. I. Investigating MS2/MS3 matching statistics: a model for coupling consecutive stage mass spectrometry data for increased peptide identification confidence.http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2734953/pdf/nihms121924.pdf. *Mol. Cell. Proteomics* **2008**, *7* (1), 71–87.

(22) Shteynberg, D.; Deutsch, E. W.; Lam, H.; Eng, J. K.; Sun, Z.; Tasman, N.; Mendoza, L.; Moritz, R. L.; Aebersold, R.; Nesvizhskii, A. I. iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Mol. Cell. Proteomics* **2011**, *10* (12), M111.007690.

(23) Sheng, Q.; Dai, J.; Wu, Y.; Tang, H.; Zeng, R. BuildSummary: using a group-based approach to improve the sensitivity of peptide/protein identification in shotgun proteomics. *J. Proteome Res.* **2012**, *11* (3), 1494–1502.

(24) Ramakrishnan, S. R.; Vogel, C.; Prince, J. T.; Li, Z.; Penalva, L. O.; Myers, M.; Marcotte, E. M.; Miranker, D. P.; Wang, R. Integrating shotgun proteomics and mRNA expression data to improve protein identification. *Bioinformatics* **2009**, *25* (11), 1397–1403.

(25) Li, J.; Zimmerman, L.; Park, B. Network-assisted protein identification and data interpretation in shotgun proteomics. *Mol. Syst. ...* **2009**, *5* (303), 303.

(26) Ramakrishnan, S. R.; Vogel, C.; Kwon, T.; Penalva, L. O.; Marcotte, E. M.; Miranker, D. P. Mining gene functional networks to improve mass-spectrometry-based protein identification. *Bioinformatics* **2009**, *25* (22), 2955–2961.

(27) Wang, Z.; Gerstein, M.; Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **2009**, *10* (1), 57–63.

(28) Craig, R.; Cortens, J. P.; Beavis, R. C. Open source system for analyzing, validating, and storing protein identification data. *J. Proteome Res. 3* (6), 1234–1242.

(29) Wilhelm, M.; Schlegl, J.; Hahne, H.; Moghaddas Gholami, A.; Lieberenz, M.; Savitski, M. M.; Ziegler, E.; Butzmann, L.; Gessulat, S.; Marx, H.; et al. Mass-spectrometry-based draft of the human proteome. *Nature* **2014**, *509* (7502), 582–587.

(30) Shanmugam, A. K.; Yocum, A. K.; Nesvizhskii, A. I. Utility of RNA-seq and GPMDB protein observation frequency for improving the sensitivity of protein identification by tandem MS. *J. Proteome Res.* **2014**, *13* (9), 4113–4119.

(31) Savitski, M. M.; WIlhelm, M.; Hahne, H.; Kuster, B.; Bantscheff, M. A scalable approach for protein false discovery rate estimation in large proteomic data sets. *Mol. Cell. Proteomics* **2015**, mcp.M114.046995.

(32) Reiter, L.; Claassen, M.; Schrimpf, S. P.; Jovanovic, M.; Schmidt, A.; Buhmann, J. M.; Hengartner, M. O.; Aebersold, R. Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Mol. Cell. Proteomics* **2009**, *8* (11), 2405–2417.

CHAPTER 2
A re-scoring based approach to integrating orthogonal information for improved protein identification

INTRODUCTION

As described previously, the proteomic analysis pipelines identify peptides from MS/MS spectra and map these to protein sequences to provide a list of protein identifications that are believed to be present in the sample under analysis. An identification probability is also determined for each protein, which reflects the confidence in the protein identification based on evidence from the MS/MS experiment. Integration of evidence from orthogonal sources of information can provide a boost to the identification probability of true positive identifications that do not have sufficient evidence, based on MS/MS data alone, to pass the FDR threshold. Thus integrative analysis can improve the sensitivity of protein identifications.

In this chapter, we describe a re-scoring based approach to integrate orthogonal data into proteomic analysis pipelines. While generally applicable to all orthogonal information, this approach was specifically tested on two types of orthogonal information, namely RNAseq transcript abundance and Global Proteome Machine database (GPMDB) protein identification frequency, the results of which are presented here. The re-scoring was seen to provide significant improvement over the basic analysis pipeline especially for small to medium sized datasets.

METHODS

Datasets description

Data from VCaP[1], a human prostate cancer cell line, and HEK293[2], a cell line derived from human embryonic kidney cells were used in this study. The MS/MS and RNAseq data for the VCaP cell line were generated in parallel at the same lab. This RNAseq data was also used as the control sample in a paper by L. Sam et al[3] and is available for download from the NCBI short read archive, SRA (SRA accession numbers: SRR090590, SRR090591). For the HEK293 cell line, MS/MS data was obtained from control samples in a publication by B. Fonslow et al[4] while the RNAseq data was downloaded from data generated by M.Sultan et al [5] (SRA accession numbers: SRR023583, SRR023584).

The GPMDB identification frequencies which are not cell line specific, were obtained by querying GPMDB for every Ensembl protein id using a perl script. This data was retrieved on April 30th 2012.

MS/MS experimental protocol for VCaP

The VCaP cell line was provided by Dr. Ken Pienta (University of Michigan, Ann Arbor). Collection of VCaP whole cellular protein extract was done in RIPA complete buffer supplemented with HALT Protease and Phosphatase Inhibitor Cocktail (Peirce Biotechnology). Total protein extract was quantified by bicinchoninic acid assay. 50 mg aliquots of total cellullar proteins were first separated by 1D SDS-PAGE (4-12 % Bis-Tris Novex-Invitrogen, Carlsbad, CA). Forty equal sized gel bands were excised and subjected to in-gel digestion as previously described[6]. Extracted peptides were reconstituted with mobile phase A prior to on-line reverse phase nanoLC-MS/MS (LTQ-Velos with Proxeon nanoHPLC, ThermoFinnigan). Peptides were eluted on-line to the mass spectrometer with a reverse phase linear gradient from 97 % A (0.1 % Formic acid in water) to 45 % B (0.1 % formic acid in acetonitrille. Peptides were detected and fragmented in the mass spectrometer in a data dependent manner sending the top 12 precursor ions, excluding singly charged ions, for collisional induced dissociation. Raw spectra files were converted into mzXML by an in-house version of ReAdW[7].

Preparation of datasets

Both of the MS/MS experiments (VCaP and HEK293) used in this study were seen to have very deep proteome coverage, with about 4000-6000 protein identifications at 1% FDR. However, most MS/MS experiments do not achieve this level of proteome coverage. To investigate performance in experimental conditions with varying depths of proteome coverage, the MS/MS data was sampled at the level of individual mzXML files to create various subsets of data of varying sizes (fewer files would be included for a smaller subset and more files to get a larger subset). The VCaP data had a total of 40 mzXML files while the HEK293 data consisted of 60 mzXML files. The number of protein identifications in these subsets ranged between about 500 to 5000 protein identifications at 1% FDR.

MS/MS data analysis pipeline

The MS/MS data was searched using the X!Tandem (CYCLONE (2010.12.01.1))[8] search engine with a K-score plugin[9,10] provided by the Trans-proteomic Pipeline. The search was performed against the Ensembl v.66 Human proteome with reversed protein sequences appended as decoys. Trypsin was specified as the enzyme with no missed cleavages allowed and cysteine carbamidomethylation and methionine oxidation were set as a fixed and variable modifications respectively. VCaP data was searched using a precursor mass error of -1 Da to +4 Da while the HEK293 data (high mass accuracy data) was searched with a precursor mass error of +/- 50 PPM. Fragment mass error was set to 0.8 Da for both searches.

Statistical validation of PSMs was performed using the Trans-proteomic Pipeline (TPP v4.6 OCCUPY rev 2) software suite[11]. VCaP data was processed with +1 charge state ions set to be ignored and using a semi-supervised model[12] for estimating negative distributions. HEK293 data was processed using the same settings as above along with additional parameters to use accurate mass binning and the PPM scale for the mass models. The output protXML files from TPP were processed using the Abacus software tool[13] to select a representative protein for each protein group, according to heuristic filters built into the tool.

RNAseq processing pipeline

RNAseq data was aligned to the Ensembl v.66 Human Genome (hg19 build 37) using the Tophat aligner (Tophat v.1.3.2)[14]. Parameters were set to allow up to one mismatch per alignment and a GTF file containing Ensembl v.66 gene annotations was provided to Tophat, using the '-G' option, to improve alignment accuracy. For aligning HEK293 reads, an additional parameter was used to set the segment length to 13 bases.

Transcript abundance, in the form of Reads Per Kilobase per Million mapped reads (RPKM)[15] (read count normalized to transcript length and total number of reads in the experiment), was calculated for each transcript from the BAM file output from Tophat. RPKM calculation was performed with a custom R script utilizing functions from the Bioconductor[16] packages Rsamtools (v.1.6.3)[17] and GenomicFeatures (v.1.6.9)[18].

Sampling methodology for assigning RPKM and GPMfreq values to decoys

In order to utilize decoys, which do not have inherent RPKM or GPMfreq values, in the pipeline a way to assign these values to the decoys is required. This was achieved in our study by randomly by sampling RPKM and GPMfreq values from target proteins to assign to the decoys. As will be discussed later in this chapter, the discriminatory power (between true and false positive identifications) provided by the orthogonal data is preserved in this sampling.

In our data, weak correlations were observed between protein length, RPKM and GPMfreq values. It was desirable to preserve this structure in the data when assigning values to the decoys. Therefore, in our sampling approach, RPKM and GPMfreq values are always sampled and assigned together (if they come from the same forward sequence, they are assigned to the same decoy sequence). This helps to preserve correlation between those values. To preserve the correlations with protein length RPKM/GPMfreq values are sampled and assigned to decoy sequences of a similar length.

The sampling process begins by selecting the shortest decoy sequence identified in the MS/MS experiment. All other decoys with length within 50 AA of the shortest decoy are grouped together with it. Then RPKM and GPMfreq values for all forward sequences (not just proteins identified in the MS/MS experiment but all proteins in the Ensembl human proteome) that fall within the same length range as the grouped decoys are randomly sampled and assigned to the decoys. To ensure sufficient population for proper random sampling, the set of forward sequences that we sample from is required to have at least 100 proteins. If a particular decoy group has less than 100 corresponding forward sequences in the length range(shortest decoy length + 50 AA), the length range is incremented in steps of 1 (and adding decoys within that length range into the group) until the corresponding forward sequences set has at least 100 sequences to sample from. After sampling is finished for the first decoy group, the next shortest decoy not yet assigned values is selected and the above process is repeated until all decoys have been assigned values.

Computing pProt values

The maximum peptide probability (maxPepProb) has been observed to be a better measure of protein identification confidence than the native protein probability reported by TPP (Figure 2.1). However, to use it as the prior probability for probability adjustment, it needed to be converted to a protein probability score, which we call *pProt.*

For this conversion, protein identifications were first sorted in decreasing order of the maxPepProb. This sorted list was then divided into bins of 50 proteins each and the local FDR was computed for each bin (*r* x *nDecoys* / *nForwards*, in that particular bin). This local FDR value estimates the probability of a protein, at a particular maxPepProb level, to be a decoy. The inverse of this value, 1 – local FDR, hence captures the probability of a protein at a given maxPepProb level for being a true positive identification. In other words it is a protein identification probability (*pProt*).

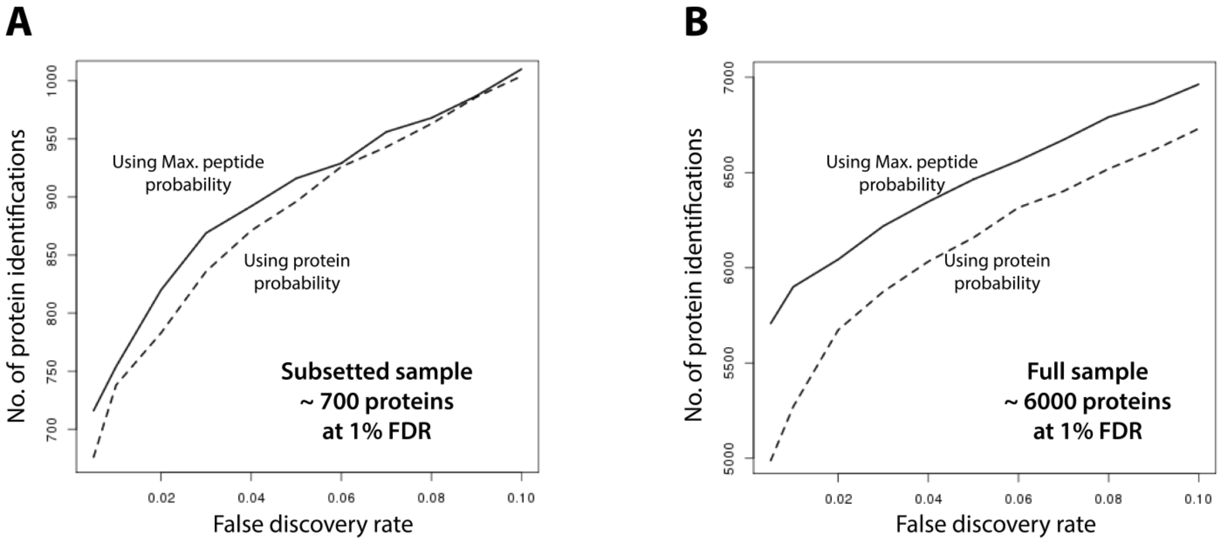


Figure 2.1: ROC curves using protein probability and maximum peptide prob. for ranking protein identifications. The maximum peptide prob. is seen to be a better measure of true protein identification, especially in large samples.

The raw local FDR values can show random fluctuations, as we move down the sorted list of protein identifications, as a result of local variations in bin composition. Loess smoothing is performed to smooth out the fluctuations before using local FDR to compute *pProt*. Also due to localized variations in the distribution of decoys, the local FDR value might exceed 1 in some bins (more decoys than forwards). This is dealt with by capping the local FDR values at 1. Also, there is usually a long tail of bins with high local FDR values (low maxPepProb region). To prevent these values from overweighting the loess smoothing, and inflating values for the low local FDR bins, only the first bin with local FDR 1 or more is included during smoothing. *pProt* values calculated as above are observed to vary monotonically with the maximum peptide probability. The conversion of the confidence score from maximum peptide probability to *pProt* was not seen to alter the ranking of proteins or the number of proteins identified above the 1% or 5% FDR thresholds.

Computing mean adjusted probability for decoys

Since RPKM/GPMfreq values are randomly sampled for assignment to decoy sequences (see 'Sampling methodology for assigning RPKM & GPMfreq values to decoys'), there is no stability in the value assigned to a particular decoy across iterations (i.e.) a particular decoy may be assigned a high RPKM/GPMfreq value in some iterations and low value in others. So directly averaging the adjusted probabilities for a particular decoy protein across all iterations is meaningless. However the distribution of all values assigned to decoys does show consistency across iterations, i.e. in every iterations there are always a certain number of decoys with high RPKM /GPMfreq values, a certain number with mid-range values and a certain number with low values etc. Therefore a mean distribution of adjusted probability values can be computed. We do this by sorting the adjusted probability values from each iteration and calculating the mean values across iterations for each position in the sorted list (i.e.) mean value of position 1 in all iterations, mean value of position 2 in all iterations etc. The mean adjusted probability for each decoy is obtained by again sampling from this mean distribution and assigning to the decoys.

R – Factor correction

When estimating the False Discovery Rate, FDR, it is typically assumed that each decoy identification represents the presence of one false positive forward protein identification at that protein score level. Under this assumption the FDR is estimated as,

$$FDR = \frac{N_d}{N_f}$$

, where N_d = Number of decoy identifications

N_f = Number of forward identifications

But this assumption does not always hold true, especially in samples with a large number of protein identifications. In such cases the 'r-factor', the true ratio of false positive forward identifications to decoy identifications in an MS/MS experiment should be estimated and used to estimate FDR more accurately.

At very low protein identification probabilities (on the order of 0.2 or less) it is reasonable to assume that all of the identifications (from both forward and decoy sequences) are likely to be false positive matches. Therefore the ratio of number of forward identifications to number of decoy identifications in this low probability region can be used as an estimate of the true ratio of false positive forward matches to decoy matches in the overall dataset i.e. the r-factor.

$$r = \frac{N_{f\_0.2}}{N_{d\_0.2}}$$

, where $N_{f\_0.2}$ = Number of forward identifications at identification probability 0.2 or less

$N_{d\_0.2}$ = Number of decoy identifications at identification probability 0.2 or less

By computing the number of false positive forward identifications at a particular protein score as the number of decoy identifications weighted by the r-factor, the FDR estimates will be more accurate.

$$FDR = \frac{r \times N_d}{N_f}$$

Replicating the customized database approach

To compare the results of this probability adjustment method with the customized database approach previously described by X. Wang et al, a perl script was developed replicating the pipeline as described in their manuscript. Given a fasta file, a file containing RPKM values of all proteins in the fasta file and an RPKM threshold, the perl script creates a customized fasta file containing only those proteins that were higher than the given RPKM threshold.

In their paper, RPKM thresholds were set empirically by comparing to matching mRNA and MS/MS data but were seen to correspond approximately to an inflection point in the bimodal distribution seen for the log-transformed RPKM values (presumably dividing a background level of expression and actual transcript expression). A similar bimodal distribution was observed in the RNA-seq data used in this work as well, and so an empirical threshold of RPKM 0.1 close to the inflection point was chosen (Figure 2.2). The manuscript also suggests manually adding histone proteins to the customized database since transcripts corresponding to histones are more likely to be missed by RNAseq. However in our data, 2/3rds of known histones (as listed by HGNC) were found in RNAseq and passed our RPKM threshold. The other 1/3rd of known histones were not identified in the full database search. Therefore histones were not manually added to the customized database.

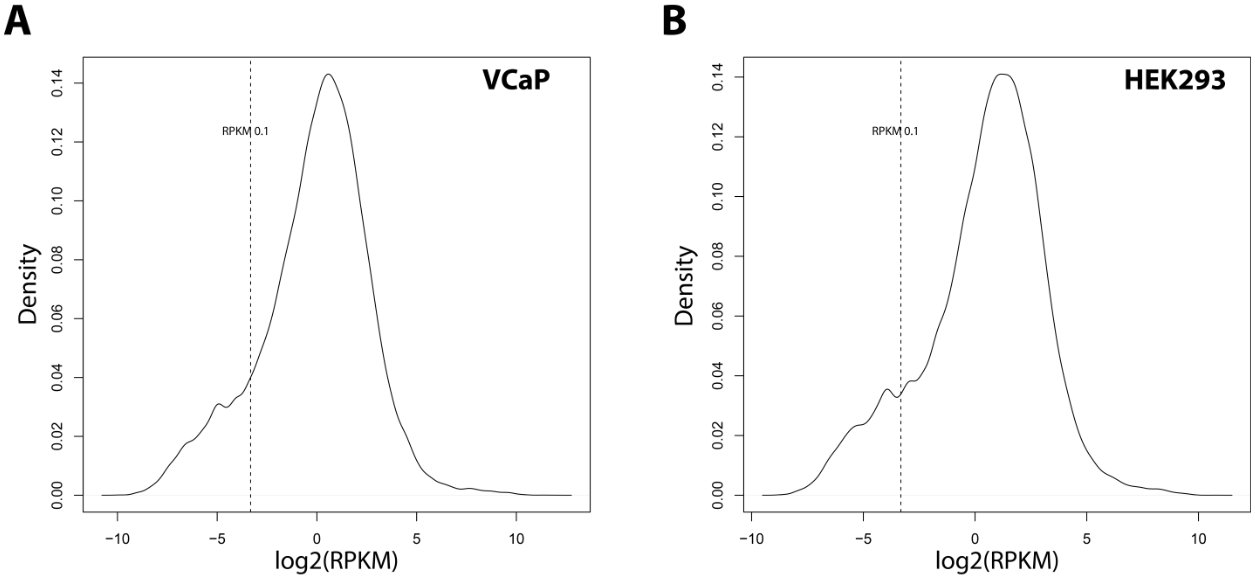
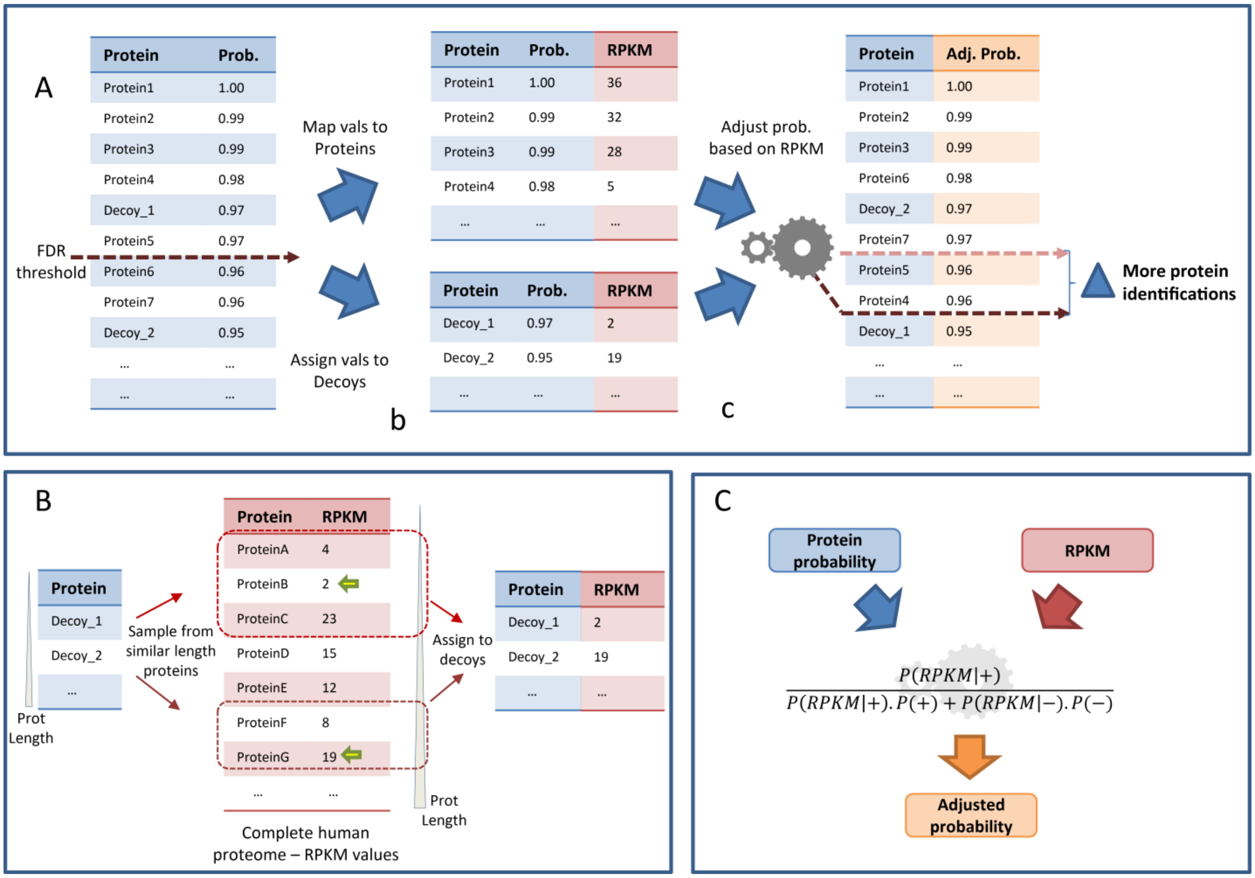


Figure 2.2: Density distribution (log-scaled) of RPKM values for all proteins. The RPKM thresholds used in our re-implementation of the customized database approach are shown.

RESULTS AND DISCUSSION

Overview of the approach

Our approach to incorporating RNA-seq or GPMDB frequency information (Figure 2.3) is built upon a statistical adjustment[19,20] of the protein probability. The probability adjustment (re-scoring) increases the identification confidence of proteins that have significant supporting evidence from external data (high RNAseq transcript abundance or high GPMDB identification frequency), relative to other proteins without such supporting evidence. Protein identifications that previously fell just below the FDR threshold based on MS/MS evidence alone, in a ‘grey zone’ of identification confidence, can be promoted above the threshold when ranked by adjusted probability. Therefore, we are able to obtain more protein identifications at the same FDR.


A
FDR threshold
Map vals to Proteins
Assign vals to Decoys
b
Adjust prob. based on RPKM
c
More protein identifications
B
Prot Length
Sample from similar length proteins
Assign to decoys
Prot Length
Complete human proteome – RPKM values
C
Protein probability
RPKM
$\frac{P(RPKM|+)}{P(RPKM|+).P(+)+P(RPKM|-).P(-)}$
Adjusted probability

Protein	Prob.
Protein1	1.00
Protein2	0.99
Protein3	0.99
Protein4	0.98
Decoy_1	0.97
Protein5	0.97
Protein6	0.96
Protein7	0.96
Decoy_2	0.95
…	…
…	…

Protein	Prob.	RPKM
Protein1	1.00	36
Protein2	0.99	32
Protein3	0.99	28
Protein4	0.98	5
…	…	…

Protein	Prob.	RPKM
Decoy_1	0.97	2
Decoy_2	0.95	19
…	…	…

Protein	Adj. Prob.
Protein1	1.00
Protein2	0.99
Protein3	0.99
Protein6	0.98
Decoy_2	0.97
Protein7	0.97
Protein5	0.96
Protein4	0.96
Decoy_1	0.95
…	…
…	…

Protein
Decoy_1
Decoy_2
…

Protein	RPKM
ProteinA	4
ProteinB	2
ProteinC	23
ProteinD	15
ProteinE	12
ProteinF	8
ProteinG	19
…	…

Protein	RPKM
Decoy_1	2
Decoy_2	19
…	…

Figure 2.3: Overview of the re-scoring based approach. (A) External information is added to protein identifications from the analysis pipeline. Protein probabilities are adjusted on the basis of transcript abundance (RPKM) or GPMDB identification frequency (GPMfreq). (B) Decoys sequences used to estimate FDR thresholds do not have native RPKM (or GPMfreq) values; they are assigned values by sampling from set of all forward sequences with similar length (see methods). (C) Protein identification probabilities are adjusted using Bayes' theorem.

RPKM/GPMfreq value assignment for decoys

FDR estimation for protein identifications is performed by the Target-Decoy approach. (Reversed 'decoy' sequences are appended to the 'forward' protein sequence database before performing database searching. Number of PSMs matched to decoy sequences is used to estimate the rate of random matches in the PSMs matched to forward sequences). Since decoy sequences do not have inherent RPKM/GPMfreq values their identification probabilities would be selectively decreased during probability adjustment based on the external information. To be able to un-biasedly estimate FDR after probability adjustment, a rational method for assigning RPKM/GPMfreq values to decoy sequences is necessary.

The density distribution of RPKM/GPMfreq values for proteins identified in the MS/MS experiment at 1% FDR (confident true positive identifications) is seen to be appreciably different from that of RPKM/GPMfreq for all proteins (Figure 2.4). Based on this observation, sampling from the 'all proteins' distribution allows us to unbiasedly assign RPKM/GPMfreq values to decoy sequences while maintaining discrimination between decoy and forward identifications. Weak correlations between the RPKM, GPMfreq values and protein length were also observed in the data (data not shown). To preserve this structure in the data our sampling approach was designed to sample RPKM/ GPMfreq values together from forward sequences and assign them only to decoys of similar length (See methods section for a detailed description of the sampling process).

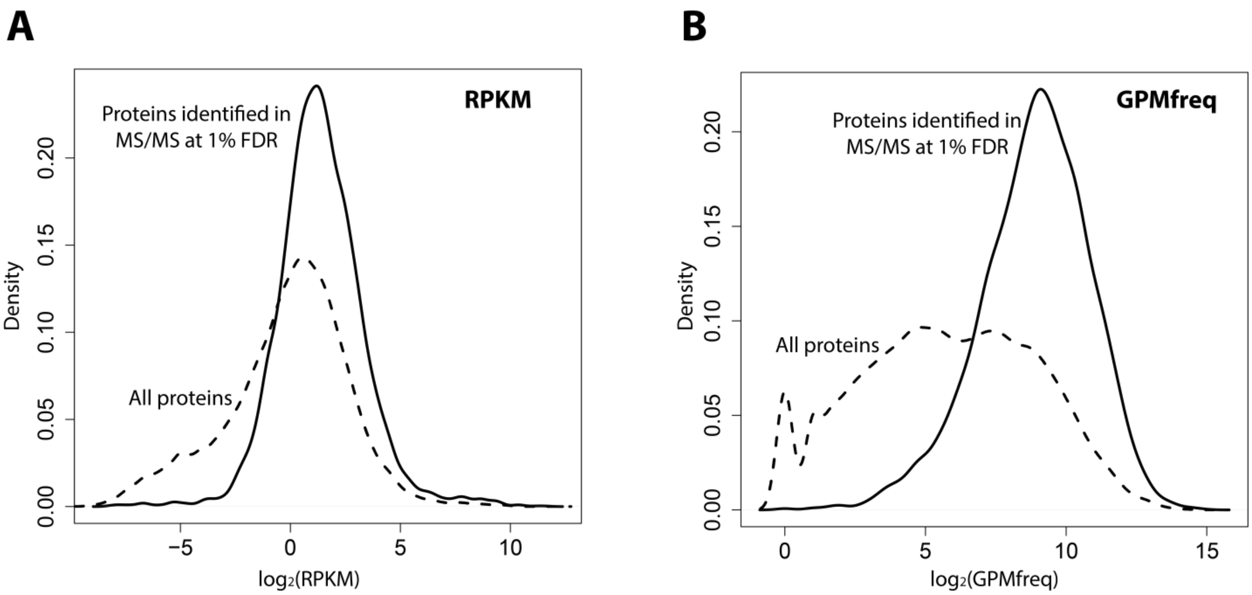


Figure 2.4: Density distribution (log-scaled) of RPKM (A) and GPMfreq (B) values – High confidence proteins vs All proteins. The difference between distributions of proteins identified in 1% FDR and all proteins in Ensembl, allows us to sample the 'all proteins' distribution to un-biasedly assign values to decoys while still maintaining discrimination between true positive and decoy identifications.

Probability adjustment

When performing probability adjustment, pProt, a protein probability score calculated on the basis of maximum peptide probability, was used as the prior probability(See methods section for details of *pProt* calculation). *pProt* was used instead of the native protein probability reported by TPP because it has been observed that the maximum peptide probability is a more reliable indicator of true protein identification than the protein probability value, especially for large samples.

Using Bayes' theorem the probability adjustment estimates the probability of a protein identification being a 'true positive' given its RPKM / GPMfreq value.

$$P(+|V) = \frac{P(V|+).P(+)}{P(V|+).P(+) \; + \; P(V|-).P(-)}$$

, where *V* can be either an RPKM or GPMfreq value.

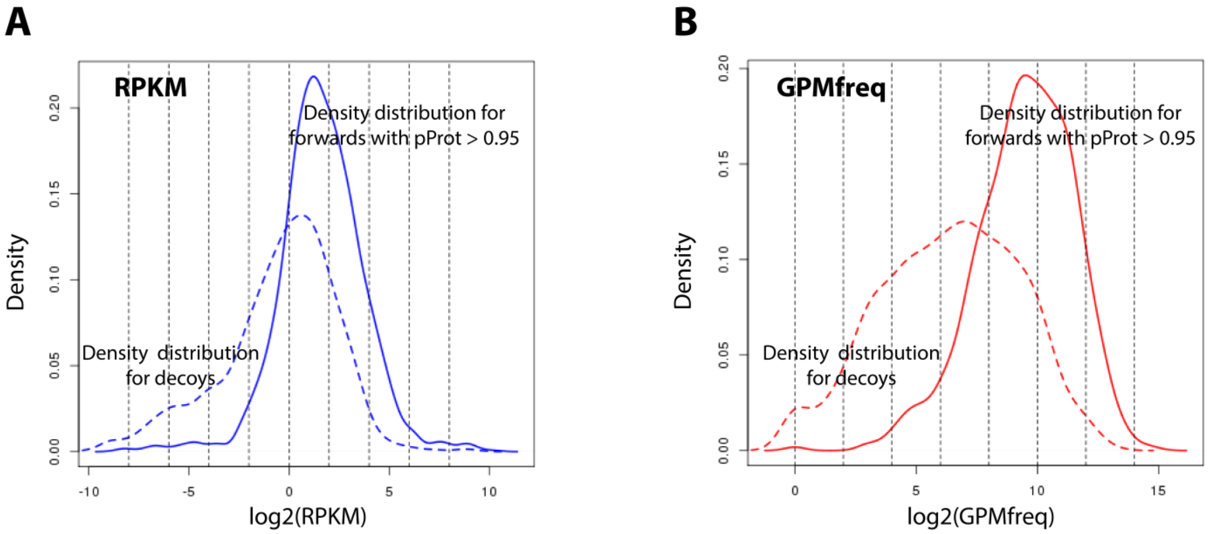


Figure 2.5: Density distribution (log-scaled) of RPKM and GPMfreq values - high confidence proteins vs. decoys . Shown from one of the iterations of sampling, with the bin thresholds that were used for computing conditional probabilities indicated.

The prior probability terms *P(+)* & *P(-)* were substituted for with *pProt* and *1- pProt* respectively. To estimate the conditional probabilities of decoy or forward identifications having value V, *P(V|-)* & *P(V|+),* the density distribution of log-scaled RPKM/GPMfreq values into bins of equal width (Figure 2.5). Conditional probabilities for each bin were calculated as $P(V_i|-) = nD_i / nD_t$, where V_i is any RPKM/GPMfreq value that falls within bin *i*, nD_i is the number of decoys having RPKM/GPMfreq values within bin *i and* nD_t is the total number of decoys in the sample. Values for $P(V_i|+)$ were also estimated similarly, but instead of number of decoys, the number of forward hits with pProt > 0.5 (i.e. forward identifications that are more likely to be true positive than false positive) were used.

Effect of the probability adjustment

Since the RPKM/GPMfreq values are assigned through random sampling, the assignment and probability adjustment (Fig. 1A), are repeated multiple times to nullify any sampling artifacts and obtain stable mean adjusted probability values. In our study, the mean values were typically seen to stabilize after about 200 iterations (Figure 2.6), but the process was repeated to 500 iterations for the results reported here.

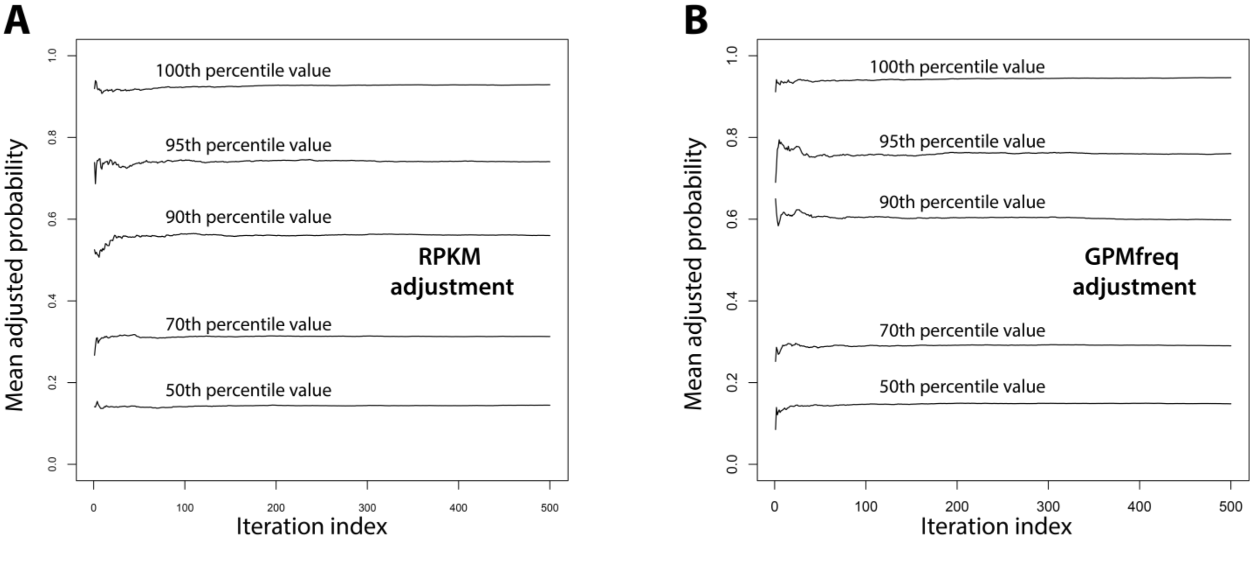


Figure 2.6: Stabilization of mean adjusted probabilities of decoys. The values in 100th, 95th, 90th, 70th and 50th percentiles of the non-zero mean adjusted probabilities of decoys were plotted, after each iteration, to track their stability. The values stabilize after about 200 iterations, but sampling is repeated to 500 iterations to ensure stability.

The effect of the probability adjustment was measured by comparing the number of protein identifications at 1% FDR without adjustment to the number of protein identifications at 1% FDR after probability adjustment (RPKM or GPMfreq based). The percent improvement from all of the various subsets was calculated and plotted, as shown in Figure 2.7A and Figure 2.7B. Loess smoothing was performed on the values to show trends clearly.

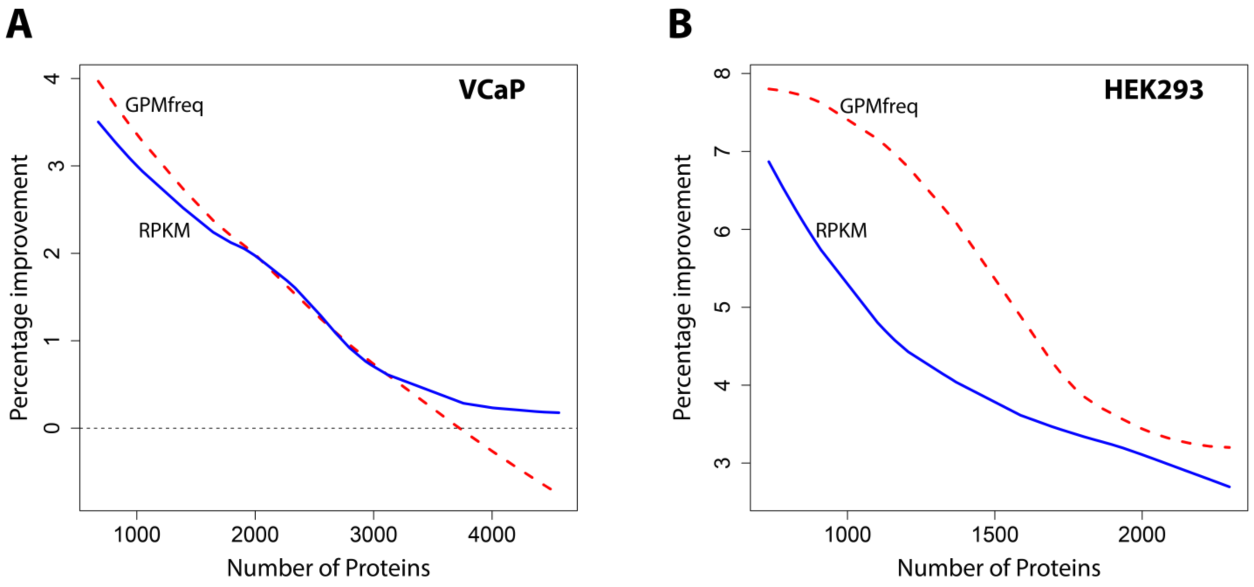


Figure 2.7: Percentage improvement in protein identifications due to probability adjustment. Results are shown for VCaP (A) and HEK293 (B) cell lines are plotted at various depths of proteome coverage (Number of proteins). The adjustment is seen to be more effective for low and medium coverage datasets.

The probability adjustment results in improvements of almost 8% in the HEK293 cell line and up to 4% in VCaP (Figure 2.7). Notably, the amount of improvement observed is similar for both RPKM adjustment and GPMfreq adjustment of the protein probability. Furthermore, it appears that using RNAseq data generated in parallel to the MS/MS data (VCaP) or RNAseq generated at a different time and location from the MS/MS data (HEK293) does not significantly affect the results.

We believe the probability adjustment works by boosting protein identifications that fall in a ‘grey-zone’ of confidence of identification. To test this hypothesis, the entire analysis described above was repeated using maximum hyperscore instead of maximum peptide probability as the identification confidence score. Hyperscore is a spectral matching score calculated and reported by the X!Tandem search engine. The maximum hyperscore for a protein can be used as an alternate, albeit less effective than the maximum peptide probability, confidence score for sorting protein identifications and estimating FDR thresholds. Since maximum hyperscore is a sub-optimal score compared to maximum peptide probability, the resulting protein identifications should have more proteins in the ‘grey-zone’ and therefore the probability adjustment on these identifications should provide higher improvement. As expected, Figure 2.8A and 2.8B show that the percentage improvement is much higher (7 – 20%) in the maximum hyperscore based analysis. These results support the idea that the amount of improvement obtained from probability adjustment is dependent on the number of proteins falling in the ‘grey-zone’ of confidence of identification.

In our analysis a clear trend of the percentage improvement from probability adjustment decreasing as the depth of proteome coverage (i.e.) number of proteins identified in the dataset, increases. With deeper coverage of the proteome, low abundance and rare proteins are increasingly identified. As per our assumptions, such proteins would have low RNAseq abundance and /or low frequency of identification in GPMDB. Therefore these proteins will not benefit from a probability adjustment based on RPKM/GPMfreq evidence and in fact may have their confidence scores decreased by it. Furthermore, increasing depth of proteome coverage not only increases the number of proteins identified but also increases the amount of MS/MS or spectral evidence collected for each identified protein. This would lead to a decrease in the number of proteins falling in the ‘grey-zone’. Based on this, we believe that the observation of decreased improvement in deeper coverage datasets reflects the fact that in these datasets, there are fewer proteins that would benefit from the probability adjustment.

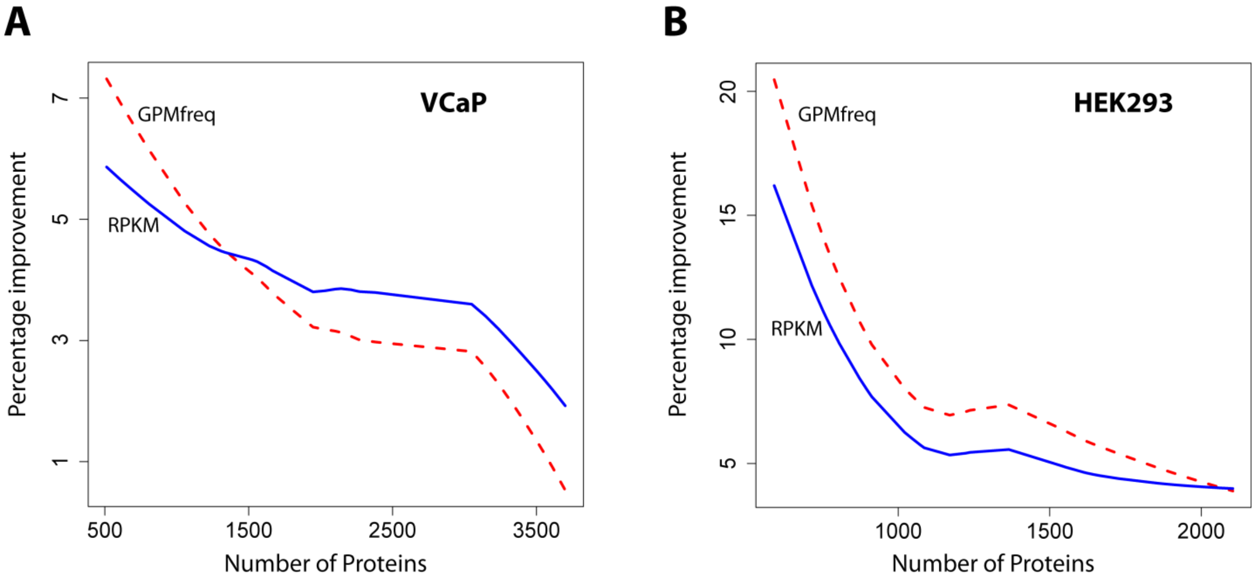


Figure 2.8: Percentage improvement due to maximum hyperscore based probability adjustment. As expected, improvement is significantly higher when the sub-optimal maximum hyperscore is used instead of maximum peptide probability (Fig. 2.7).

Validating proteins promoted by probability adjustment

A more detailed analysis of the effects of probability adjustment was carried out on one of the sampled data subsets from each cell line, the results of which are shown in Table 1. Proteins that were promoted above the 1% FDR threshold as a result of the probability adjustment were selected for manual validation. These selected proteins were compared with the list of proteins identified at 1% FDR in the complete dataset (largest dataset without any sub sampling) of that cell line. A promoted protein being found in the complete sample suggests that the protein is indeed a true identification. It is likely that there wasn't sufficient MS/MS evidence in the smaller sampled dataset for the protein to be confidently identified. But the probability adjustment using RPKM / GPMfreq information provided the necessary boost to promote it above the FDR threshold. In our analysis, 70-80% of the promoted proteins were indeed identified in the larger sample. Therefore, the probability adjustment was successful in promoting true positive identifications.

Cell Line	**Rescoring**	**No. of promoted proteins in sampled dataset**	**No. identified in complete dataset**
VCaP	RPKM	55	43
	GPMfreq	52	41
HEK293	RPKM	82	55
	GPMfreq	88	63

Table 2.1: **Number of promoted proteins validated in larger complete dataset.** The no. of proteins promoted by re-scoring in the smallest sampled dataset that were directly identified without re-scoring in larger complete dataset. Results provide a validation that the promoted proteins were indeed true positive identifications.

The remaining 20-30% of promoted proteins, which were not observed in the complete dataset, were seen to have high confidence scores in the same range as the validated proteins. In other words, unobserved proteins were not outliers (Figure 2.9). We believe that these unobserved proteins are also true positive protein identifications. It is possible that these proteins were not observed in the complete dataset because, even with the increased amount of MS/MS evidence collected in the complete dataset, there still is not sufficient evidence to confidently identify them solely by MS/MS evidence without the aid of external information. In our analysis, a larger proportion of proteins are validated in the VCaP sample, which has more MS/MS data collected (~6000 proteins), than in the HEK293 sample (~3000 proteins), which appears to support this interpretation.

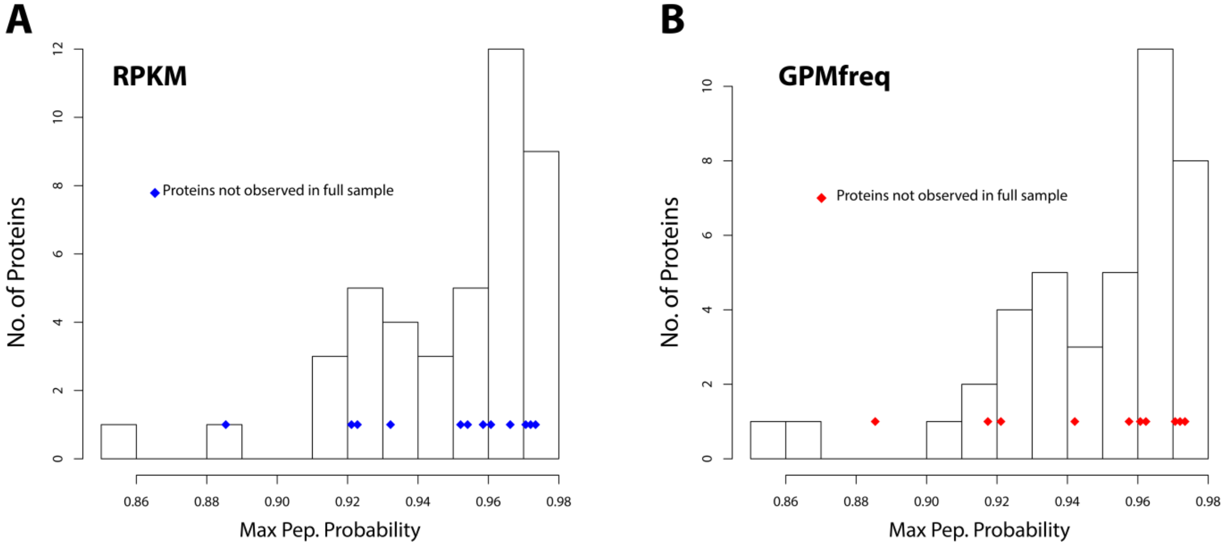


Figure 2.9: Histogram of maximum peptide probabilities of promoted proteins. Maximum peptide probabilities of proteins promoted above 1% FDR threshold by probability adjustment (RPKM or GPMfreq based) are plotted as a histogram with the locations of proteins not observed in the larger sample are marked. It is seen that the unobserved proteins aren't outliers from the validated proteins.

Attempted modifications to the probability adjustment workflow

In addition to individual probability adjustment by only RNAseq or GPMDB information, to analyze the effects of combining both, a combined re-scored probability based on both RNAseq and GPMDB information was also calculated as shown below.

$$P(+|Rval, Gval) = \frac{P(Rval|+) \times P(Gval|+) \times P(+)}{P(Rval|+) \times P(Gval|+) \times P(+) \;+\; P(Rval|-) \times P(Gval|-) \times P(-)}$$

, where *Rva*l and *Gval* represent RPKM value and GPMfreq value respectively.

However, analysis of results from combined re-scoring showed no marked improvement in protein identification over the individual probability adjustments (Figure 2.10), suggesting that RNAseq and GPMDB data capture similar types of information about a sample, for the purposes of probability adjustment.

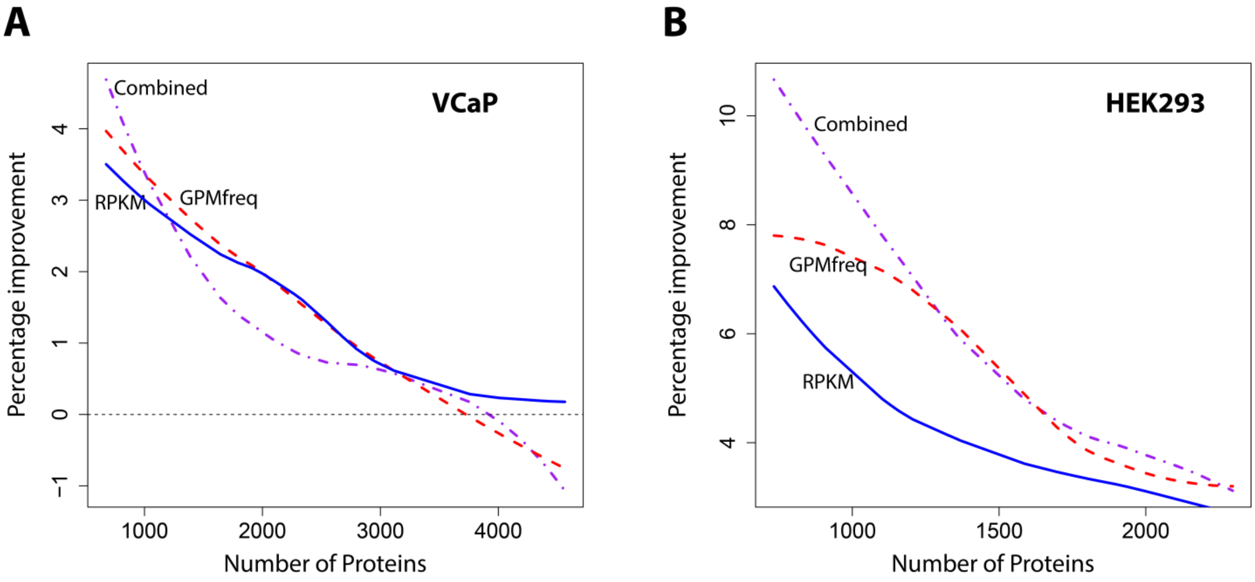


Figure 2.10: Percentage improvement due to combined adjusted probability. Results from a combined (RPKM and GPMfreq) probability re-scoring is shown, with the improvements from individual probability adjustments also shown for comparison.

In further analysis, the sampling methodology used to assign values to decoys was weighted, to preferentially assign decoy values from proteins not identified in the MS/MS experiment, instead of a completely random sampling (having an equal probability of sampling values from proteins with high confidence protein identifications as it has for any other protein). With this type of sampling, it was observed that the improvement from rescoring was slightly increased. However, this methodology might introduce some bias in the overall process. Hence, only the more statistically rigorous approach of completely random sampling was adopted in our study. But the results from this weighted sampling are reported here in the interest of potential further developments that might make use of it.

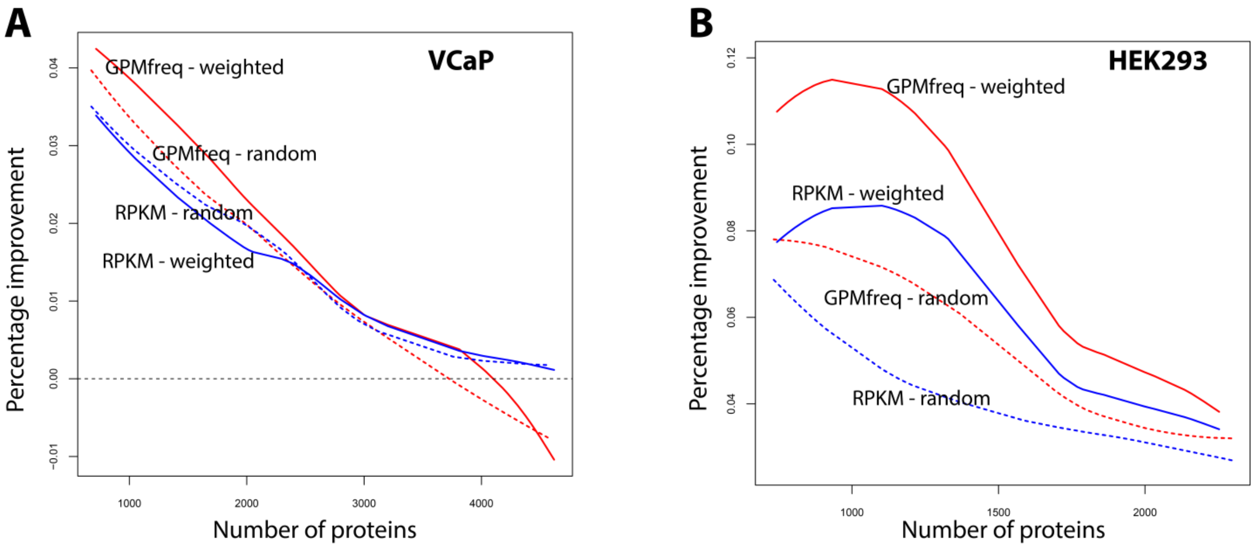


Figure 2.11: Comparison of results from weighted sampling and random sampling for decoy value assignment. Weighting the sampling methodology to preferentially assign decoy values from proteins not identified in high confidence is seen to provide a slight improvement in the results.

CONCLUDING REMARKS

The probability adjustment method described here allows us to utilize external data, RNAseq abundance or GPMDB identification frequency, to improve the sensitivity of protein identification through database searching. While some studies generate RNAseq data in parallel to proteomics data, large amounts of RNAseq data for many common organisms and/or cell lines used in biological research are already freely available from public resources such as the Sequence Read Archive (SRA)[21]. As we can see from Figure 2.7, whether RNAseq data is generated parallel with proteomics data (VCaP) or independently (HEK293) does not appear to significantly affect its utility for probability adjustment. This will allow us effectively leverage the large amounts of publicly available RNAseq data.

Furthermore, the improvement obtained by adjusting probability based on GPMfreq is similar to, and sometimes better than, improvement from RPKM adjusted probability. This is very convenient, allowing us to make use of readily available GPMDB information in our proteome analysis pipelines. Of course, this requires that the GPMDB repository contains enough experiments for the organism of interest for the GPMfreq values to be meaningful. But for commonly studies organisms of interest such as Human or Mouse, with numerous experiments in GPMDB, this can be a useful source of external information.

As mentioned earlier, the improvement obtained from probability adjustment decreases as the depth of proteome coverage of the experiment increases because there are fewer proteins in the ‘grey-zone’ that would benefit from the probability adjustment and more rare and low abundance proteins that could be penalized by it. This is an inherent upper limit to the amount of improvement possible from this method and must be taken into consideration when applying this to large samples. However it remains useful for MS/MS data of low to medium levels of proteome coverage. Hence, one potential application of this method may be for data from older instruments or experiments where instrument time available was low. Comparing to the customized database approach described by X. Wang et al[22], the probability adjustment method was seen to provide better improvement for low to medium coverage samples, while the customized database approach performed better for deep coverage samples (Figure 2.12), suggesting non-overlapping scenarios of usage for the two methods.

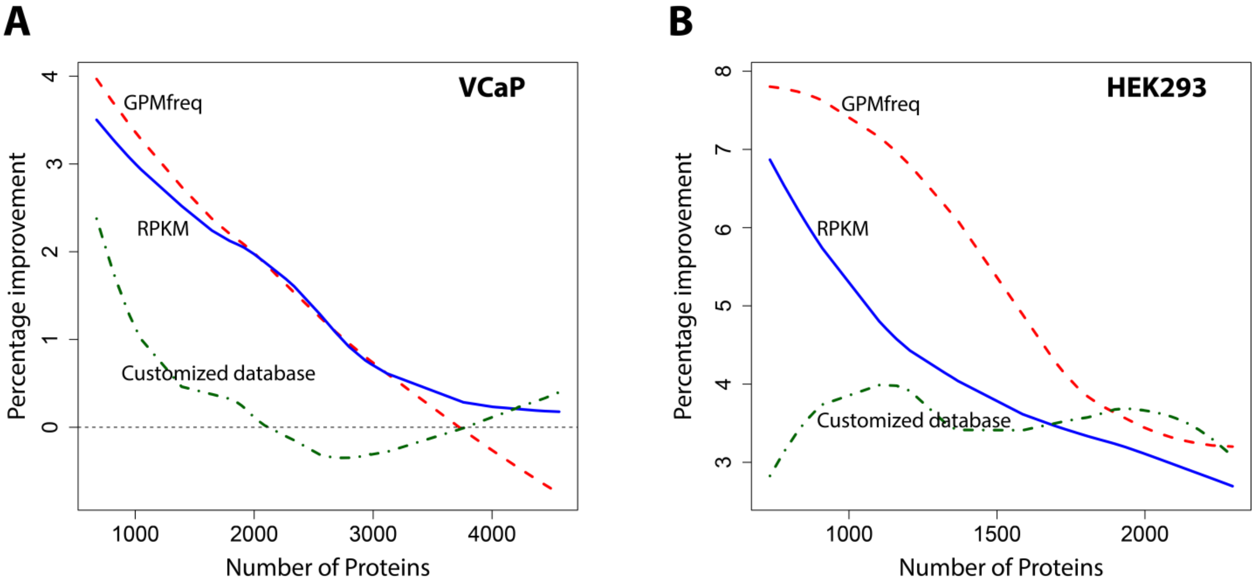


Figure 2.12: Comparison of percentage improvement from probability adjustment approaches (RPKM & GPMfreq) and the customized database approach (described by X. Wang et al).

While the probability adjustment approach described above has been demonstrated using RNAseq and GPMDB data, it does not involve any assumptions that would limit it to only these two kinds of data. So this approach provides a general framework that can be used to incorporate any external source of data, with a significant association with protein presence or abundance, into proteomic analysis pipelines for improving the sensitivity of protein identification.

REFERENCES

(1) Korenchuk, S.; Lehr, J. E.; MClean, L.; Lee, Y. G.; Whitney, S.; Vessella, R.; Lin, D. L.; Pienta, K. J. VCaP, a cell-based model system of human prostate cancer. *In Vivo 15* (2), 163–168.

(2) Graham, F. L.; Smiley, J.; Russell, W. C.; Nairn, R. Characteristics of a human cell line transformed by DNA from human adenovirus type 5. *J. Gen. Virol.* **1977**, *36* (1), 59–74.

(3) Sam, L. T.; Lipson, D.; Raz, T.; Cao, X.; Thompson, J.; Milos, P. M.; Robinson, D.; Chinnaiyan, A. M.; Kumar-Sinha, C.; Maher, C. A. A comparison of single molecule and amplification based sequencing of cancer transcriptomes. *PLoS One* **2011**, *6* (3), e17305.

(4) Fonslow, B. R.; Stein, B. D.; Webb, K. J.; Xu, T.; Choi, J.; Park, S. K.; Yates, J. R. Digestion and depletion of abundant proteins improves proteomic coverage. *Nat. Methods* **2013**, *10* (1), 54–56.

(5) Sultan, M.; Schulz, M. H.; Richard, H.; Magen, A.; Klingenhoff, A.; Scherf, M.; Seifert, M.; Borodina, T.; Soldatov, A.; Parkhomchuk, D.; et al. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* **2008**, *321* (5891), 956–960.

(6) Yocum, A. K.; Khan, A. P.; Zhao, R.; Chinnaiyan, A. M. Development of selected reaction monitoring-MS methodology to measure peptide biomarkers in prostate cancer. *Proteomics* **2010**, *10* (19), 3506–3514.

(7) Pedrioli, P. G. A. Trans-proteomic pipeline: a pipeline for proteomic analysis. *Methods Mol. Biol.* **2010**, *604*, 213–238.

(8) Craig, R.; Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20* (9), 1466–1467.

(9) Keller, A.; Eng, J.; Zhang, N.; Li, X.; Aebersold, R. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol. Syst. Biol.* **2005**, *1*, 2005.0017.

(10) MacLean, B.; Eng, J. K.; Beavis, R. C.; McIntosh, M. General framework for developing and evaluating database scoring algorithms using the TANDEM search engine. *Bioinformatics* **2006**, *22* (22), 2830–2832.

(11) Deutsch, E. W.; Mendoza, L.; Shteynberg, D.; Farrah, T.; Lam, H.; Tasman, N.; Sun, Z.; Nilsson, E.; Pratt, B.; Prazen, B.; et al. A guided tour of the Trans-Proteomic Pipeline. *Proteomics* **2010**, *10* (6), 1150–1159.

(12) Choi, H.; Nesvizhskii, A. I. Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics. *J. Proteome Res.* **2008**, *7* (1), 254–265.

(13) Fermin, D.; Basrur, V.; Yocum, A. K.; Nesvizhskii, A. I. Abacus: a computational tool for extracting and pre-processing spectral count data for label-free quantitative proteomic analysis. *Proteomics* **2011**, *11* (7), 1340–1345.

(14) Trapnell, C.; Pachter, L.; Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **2009**, *25* (9), 1105–1111.

(15) Mortazavi, A.; Williams, B. A.; McCue, K.; Schaeffer, L.; Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **2008**, *5* (7), 621–628.

(16) Gentleman, R. C.; Carey, V. J.; Bates, D. M.; Bolstad, B.; Dettling, M.; Dudoit, S.; Ellis, B.; Gautier, L.; Ge, Y.; Gentry, J.; et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **2004**, *5* (10), R80.

(17) Morgan, M.; Pages, H. Rsamtools: Binary alignment (BAM), variant call (BCF), or tabix file import.

(18) Carlson, M.; Pages, H.; Aboyoun, P.; Falcon, S.; Morgan, M.; Sarkar, D.; Lawrence, M. GenomicFeatures: Tools for making and manipulating transcript centric annotations.

(19) Ulintz, P. J.; Bodenmiller, B.; Andrews, P. C.; Aebersold, R.; Nesvizhskii, A. I. Investigating MS2/MS3 matching statistics: a model for coupling consecutive stage mass spectrometry data for increased peptide identification confidence. *Mol. Cell. Proteomics* **2008**, *7* (1), 71–87.

(20) Shteynberg, D.; Deutsch, E. W.; Lam, H.; Eng, J. K.; Sun, Z.; Tasman, N.; Mendoza, L.; Moritz, R. L.; Aebersold, R.; Nesvizhskii, A. I. iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Mol. Cell. Proteomics* **2011**, *10* (12), M111.007690.

(21) Leinonen, R.; Sugawara, H.; Shumway, M. The sequence read archive. *Nucleic Acids Res.* **2011**, *39* (Database issue), D19–D21.

(22) Wang, X.; Slebos, R. J. C.; Wang, D.; Halvey, P. J.; Tabb, D. L.; Liebler, D. C.; Zhang, B. Protein identification using customized protein sequence databases derived from RNA-Seq data. *J. Proteome Res.* **2012**, *11* (2), 1009–1017.

CHAPTER 3
Search-space restriction based and hybrid approaches to improve peptide identification

INTRODUCTION

As mentioned previously, database searching is the most commonly used technique for peptide identification from tandem mass (MS/MS) spectra in discovery based proteomics. It compares experimental MS/MS spectra to peptide sequences derived from a protein sequence database, such as the standard reference databases from Uniprot, Refseq or Ensembl, to find the best matching peptides, referred to as peptide to spectrum matches (PSMs). It is well known that the sensitivity of peptide identification is affected by size of the search space[1]. Conventional database searching itself involves elements of a targeted strategy in the sense that the search is typically restricted to sequences most likely to be present in the analyzed sample (e.g., restricting to sequences from the organism of interest, allowing tryptic peptides only with one or no missed cleavages, etc.). Any additional strategies for further decreasing the search space to only those proteins / peptides likely to be found in a particular sample or experiment should allow for higher database search sensitivity and thus, if significant loss of true sample peptides from the search space is avoided, more peptide identifications.

Targeted search space strategies in general attempt to preferentially retain proteins (or peptides) that are likely to be found in the sample while excluding those unlikely to be found from the search database. Use of RNA-Seq based transcript abundances for targeting the search space is one such strategy that has been investigated in previous studies[2]. The global proteome machine database (GPMDB)[3] is the largest repository of the results of proteomics experiments. The large volume of data aggregated in GPMDB allows the global frequencies of identification of proteins / peptides in GPMDB to be used as a reasonable surrogate measure of their propensity to be identified in an MS/MS experiment (for human, PeptideAtlas database[4] can be used equally well). Results from the previous chapter[5] have also indicated that GPMDB protein identification frequencies are comparable to RNA-Seq transcript abundance with respect to predicting protein identification propensity in a sample, suggesting that search space restriction based on GPMDB identification frequencies merits further investigation. Given the quantity and level of detail of data available in GPMDB, search space restriction can be effectively performed at the peptide level. Peptide level restriction is more advantageous than that at the protein level reflecting the fact that within a given protein sequence not all peptides are equally likely to be identified by MS/MS[6,7].

In this study we explore creation of peptide-level targeted databases based on GPMDB identification frequencies, and investigate their effect on peptide identification through database search. Importantly, to be practically useful, the computational method should allow direct and easy integration of the targeted peptide databases into existing proteomics analysis pipelines. Furthermore, while taking advantage of the increased sensitivity offered by targeted databases, it is important to address the potential limitations of search space reduction. Due to inherent limitations in how much the external information (i.e. global information accumulated in

GPMDB) correlates with protein / peptide presence in a particular biological sample under investigation, the targeted search space might be incomplete. Therefore approaches that can effectively deal with this potential search space incompleteness are critical for ensuring robust performance of the method across a wide range of experimental datasets. In this study, we investigated workflows for leveraging the increased sensitivity offered by a targeted database while also minimizing potential peptide loss due to search space incompleteness. These strategies were tested on different types of MS/MS data and were found to consistently perform at least as well and often significantly better that the conventional database search strategy.

METHODS

Datasets

The workflow development and testing described in this study was primarily performed on data from a K562 cell line lysate (Promega) acquired on a AB/Sciex TripleTof 5600 instrument by Tsou et al[8] (Accn: PXD001587). The workflows developed here were tested further on data independent acquisition (DIA) data from a K562 human cell lysate (Promega) acquired on a AB/Sciex TripleTof 5600 instrument (SWATH mode) from the same Tsou et al study; an affinity purification mass spectrometry (AP-MS) dataset generated on an LTQ instrument using MEPCE protein as bait from Mellacheruvu et al[9]; and deep coverage HeLa cell lysate acquired on QExactive HF instrument by Scheltema et al[10] (Accn: PXD001203).

Dataset	Accn.	File(s) used	N.Scans	N.Peps	N.Prots
Tsou et al; K562 lysate; **DDA**	PXD001587	*18299_REP2_500ng_HumanLysate_IDA_1.mzXML,* *18301_REP2_500ng_HumanLysate_IDA_2.mzXML*	141460	8824	1639
Tsou et al; K562 lysate; **DIA**	PXD001587	*18300_REP2_500ng_HumanLysate_SWATH_1.mzXML,* *18302_REP2_500ng_HumanLysate_SWATH_2.mzXML*	473317 (pseudo MS/MS)	7089	1384
Mellacheruvu et al; MEPCE bait; **AP-MS**	Available on request; unhosted.	*ACG_BM_7594_MEPCE.mzXML*	26370	1609	404
Scheltema et al; HeLa lysate; **Deep coverage DDA**	PXD001203	*20140201_EXQ00_RiSc_SA_STEVENHELA_01.mzXML,* *20140201_EXQ00_RiSc_SA_STEVENHELA_02.mzXML,* *20140201_EXQ00_RiSc_SA_STEVENHELA_03.mzXML,* *20140201_EXQ00_RiSc_SA_STEVENHELA_04.mzXML*	414989	62758	6200
Cabili et al; HeLa lysate; **RNAseq**	SRR309265	*SRR309265.sra*	*NA*	*NA*	*NA*

Note: NPeps and NProts determined based on searches against the full proteome database, filtered at 1% FDR.

Table 3.1: Dataset Details

GPMDB peptide identification frequencies for search space restriction were retrieved on August 10th 2014, using a MySQL database dump provided on the GPMDB FTP site. Application of workflows to RNA-Seq based search space restriction was tested using data generated from HeLa cell line on an Illumina Genome Analyzer II instrument (~ 30.4 million paired end 76 bp reads) by Cabili et al. (Accn: SRR309265)[11]. The human genome and proteome reference sequence database used for this study were obtained from Ensembl[12] release 76. Further details about the datasets and specific data files used are provided in Table 3.1.

MS/MS data analysis pipeline

The primary database search engine used in this study was MS-GF+[13] (v. 9949 2/10/2014). Searches were run with trypsin as the cleaving enzyme, a minimum peptide length of 7 amino acids, cysteine carbamidomethylation specified as a fixed modification and methionine oxidation as a variable modification. Mass tolerances were set to 30 ppm for TripleTof 5600 data searches, 20 ppm for QExactive HF searches, and 4.0 Da for the searches of AP-MS data generated using LTQ. Further testing of the methods were also carried out with the X! Tandem search engine[14] (from TPP release Jackhammer 2013.06.15.1) using the same parameters as for the TripleTof 5600 MSGF+ searches, with an additional parameter of fragment mass error set to 40 ppm.

Searches were run against the Ensembl v.76 Human proteome, and restricted search space databases derived from it, with an equal number of decoy sequences appended. Decoy sequences were created by reversing the sequence between all tryptic sites in the protein, but keeping the positions of the tryptic sites themselves unchanged. In contrast to creating decoy sequencing by reversing the entire protein sequence, this method results in decoy peptides with exact same masses as the target peptides. Further, this method also ensures that decoy peptides are consistent between the full proteome database search and the various restricted search space databases.

In the case of DIA (SWATH) data, spectra were first processed using the DIA-Umpire tool[8]. DIA-Umpire performs de-convolution of the multiplex MS/MS spectra and extracts pseudo MS/MS spectra. These pseudo MS/MS spectra are equivalent to conventional MS/MS spectra generated using data dependent acquisition (DDA) data, except they are noisier. The pseudo

MS/MS spectra were subjected to database search as described above, and further processing just as the rest of the data generated using conventional DDA strategy. DIA-Umpire provides three categories of pseudo MS/MS spectra (Q1, Q2 & Q3), corresponding to three levels of evidence. Each category of pseudo MS/MS spectra is processed separately through the pipeline and the results are combined after PSM validation.

Downstream PSM validation and protein inference was performed using the Trans-Proteomic Pipeline[15] (TPP v4.7 POLAR VORTEX rev 0) software suite. PeptideProphet[16] was run with the option to use a semi-supervised model[17] for estimating negative distributions. Except for the AP-MS data, which is of low mass accuracy, all data was processed using accurate mass binning option and the PPM scale for mass models. During processing using iProphet[18], for the results reported in this study all models except the number of sibling searches (NSS) model were turned off in order to clearly observe the effects of combining searches alone in isolation. However, a comparative analysis of results with all iProphet models (except the NSP model) turned on was also performed separately. When processing search results from the restricted search space databases the full human proteome was specified as the database in TPP. This makes sure that all peptide identifications are mapped to the full protein database prior to ProteinProphet analysis, ensuring consistent peptide to protein mapping across different analyses.

Targeted peptide sequence databases

Peptide identification frequencies in GPMDB were derived from a MySQL dump of all GPMDB data as of August 10th 2014. All peptides extracted from GPMDB were compared to the Ensembl v.76 human proteome fasta file to retain human peptide sequences only. This resulted in a list of about 1.4 billion PSMs, corresponding to 1.48 million unique peptides. To maintain a consistent comparison with the full proteome database searches, this list of human peptides from GPMDB was further filtered to only retain fully tryptic peptides containing no more than 1 missed cleavage, resulting in a filtered list of approximately 850,000 unique peptides.

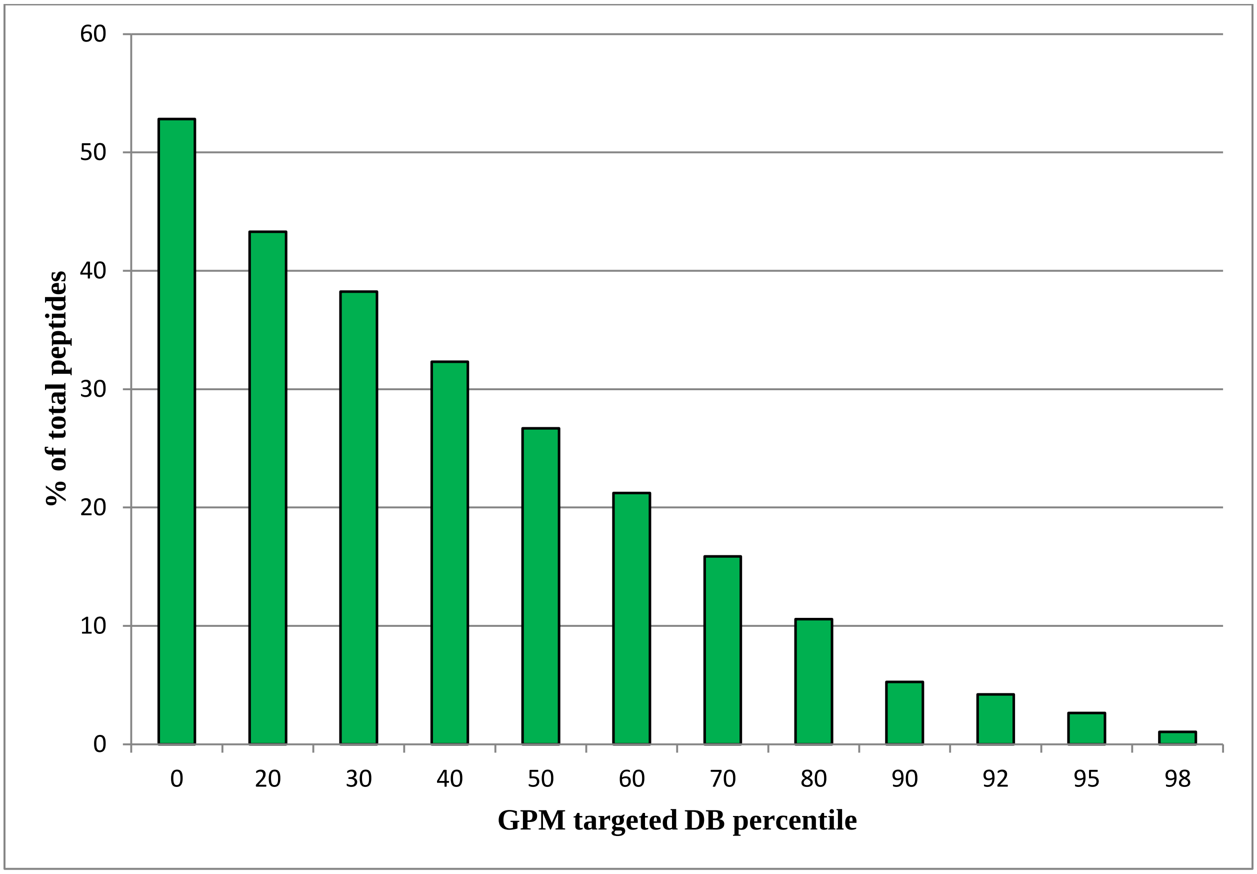


Figure 3.1: Amount of search space reduction from GPMDB based peptide databases. Number of peptides in the targeted databases as a percentage of the number of all possible tryptic peptides fulfilling our filtering criteria (fully tryptic, maximum 1 missed cleavage, length >= 7 AA).

Targeted databases were created from this filtered list by selecting peptides with frequency of identification above a certain threshold and creating a peptide sequence fasta file (each sequence in the file is an individual peptide unlike typical protein sequence database). Twelve different targeted databases, at different levels of search space reduction, were created with the frequency threshold at quantile 0%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 92%, 95% and 98%. These targeted databases ranged in size from about 52% of the full proteome database (0% quantile) to approximately 1% of it (98% quantile) in terms of number of unique peptides satisfying the filtering criteria above (see Figure 3.1).

To validate the efficacy of our targeted search space restriction, results from targeted search space restriction were compared with results from databases created by random search space restriction. These were created by randomly sampling, the same number of peptides as present in the targeted databases, from the set of all possible tryptic peptides (with up to 1 missed cleavage, & peptide length >= 7AA) in the Ensembl human proteome. Ten random databases were created for each GPM Targeted DB percentile (for a total of 120 random databases). Results from using these random databases with the basic targeted pipeline are shown in Figure 3.2 (each point is the mean improvement from 10 random samples at that database size, with error bars). As seen, a random restriction of the search space results in a significant reduction in the number of peptides identified in database search. These results demonstrate that an appropriate targeted method of performing search space restriction is critical to obtaining improvement in peptide identification.

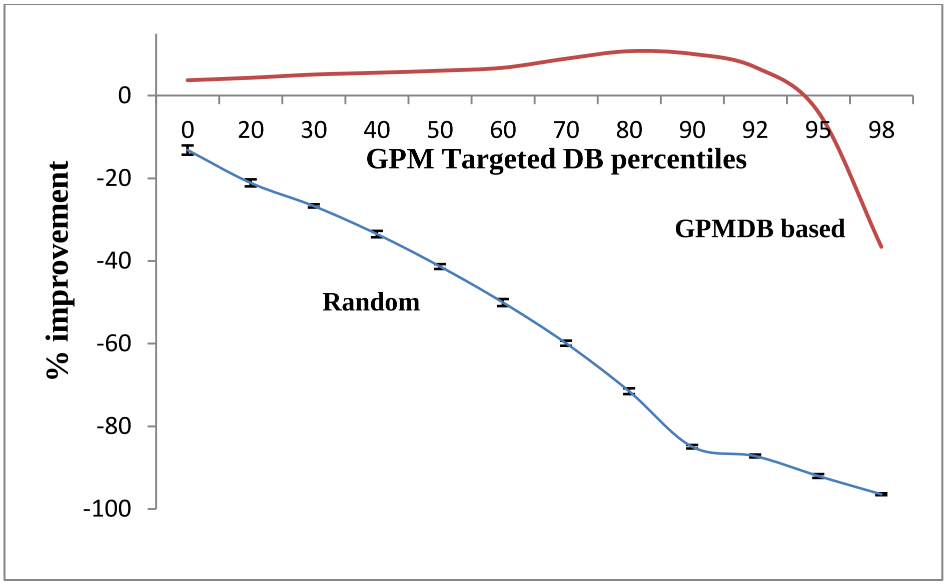


Figure 3.2: Comparison of GPMDB based and random search space restriction. Results of searching against the GPMDB targeted databases compared with results of searching against databases of the same size created by random search-space restriction. Results show the importance of using an appropriate targeted method when performing search space restriction.

Targeted protein sequence databases from RNA-Seq data

RNA-Seq data was aligned to the Ensembl v.76 human genome using Tophat[19] (v. 2.0.13) and Bowtie2[20] (v. 2.2.4). Gene annotations from Ensembl were used to improve alignments and all other Tophat options were set to default values. Transcript abundances, normalized to Reads Per Kilobase per Million mapped reads (RPKM) were computed for each transcript using a custom R (v. 3.1.0) script that utilizes functions from the Bioconductor packages Rsamtools[21] (v. 1.18.2) and GenomicFeatures[22] (v. 1.18.2). Restricted search space databases based on the RNA-Seq data were created by filtering the human proteome to only retain proteins with transcript abundances at or above a threshold as described by Wang et al[2]. While a default threshold of 30th percentile was suggested by Wang et al[23], for the data used in this study this was too stringent a threshold (see Figure 3.3). So a 20th percentile threshold was used instead.

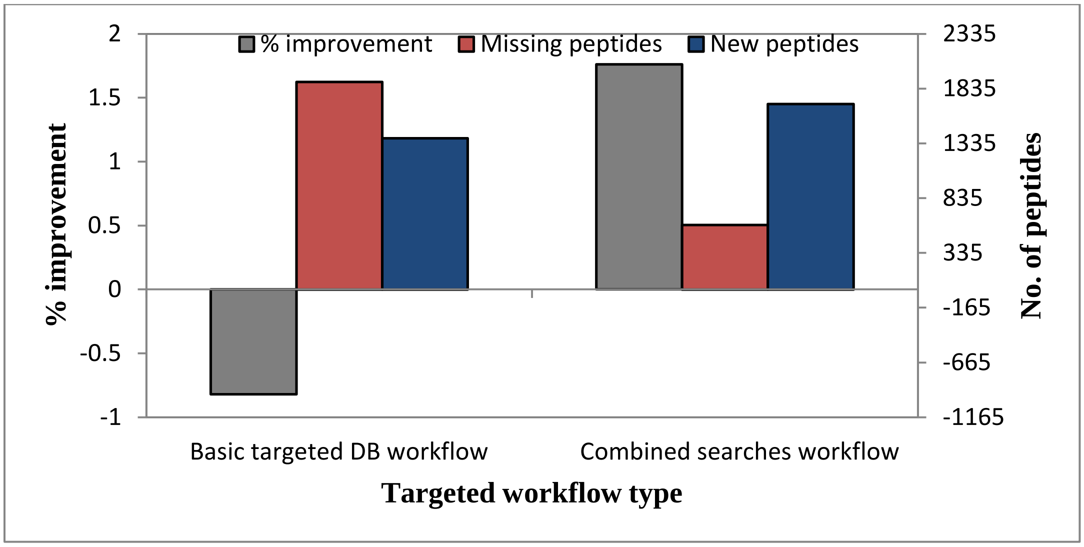


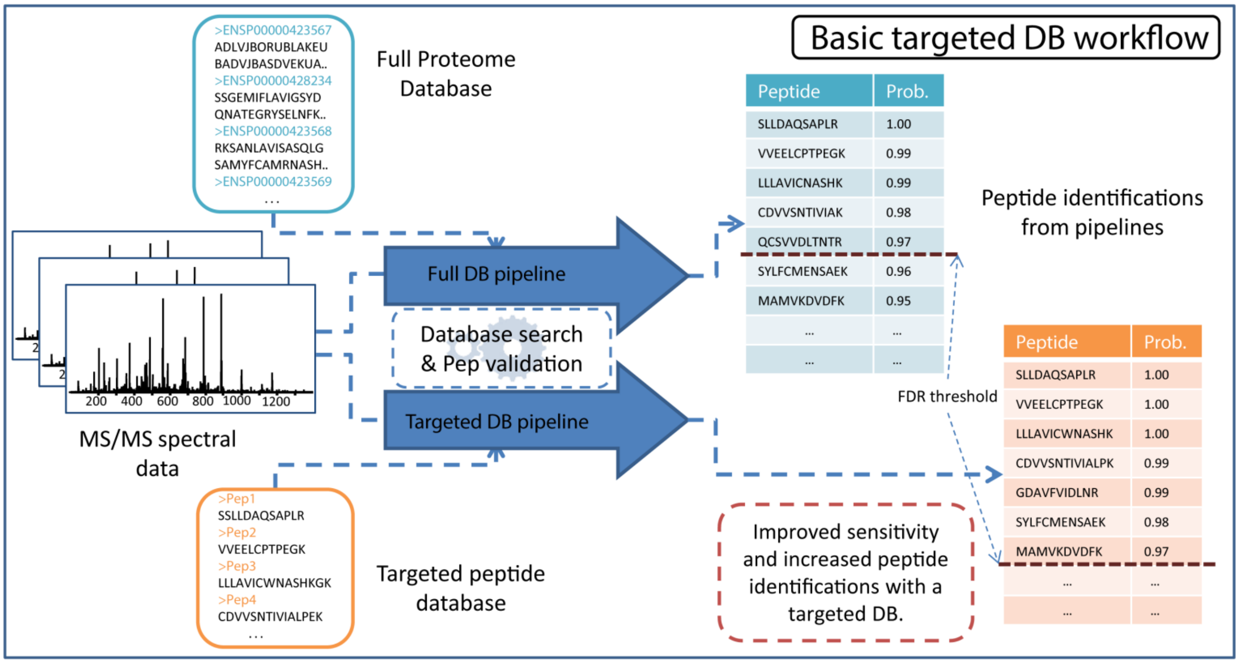
Figure 3.3: 30th percentile threshold for RNA-seq based targeted database is too stringent. Using a 30th percentile RPKM threshold causes a reduction in the number of peptides identified in the basic targeted DB workflow suggesting that it is too stringent. So a 20th percentile threshold was used in our analysis instead. [See 'Results and Discussion' section for a description of the Basic targeted DB and Combined searches workflows]

RESULTS AND DISCUSSIONS

Targeted peptide databases for peptide identification

The large number of proteomic experiments present in the GPMDB repository allows us to observe which peptides have been identified more frequently, and thus are also more likely to be identified in any new experiment. Therefore, information from the GPMDB repository was used to create targeted peptide databases (containing only peptide sequences and not protein sequences as is typical) to be used for identification of peptides from MS/MS spectra data by database searching.

The degree of database search space restriction can be adjusted by varying the frequency threshold above which peptides are included in the targeted peptide database. A higher frequency threshold corresponds to a more restrictive search space (i.e. only the most frequent peptides are included in the targeted database). Database searches of the MS/MS data were performed against targeted peptide databases filtered at thresholds ranging from 0 to 98th percentiles (see Methods), and also against the full database, as illustrated in Figure 3.4. Importantly, targeted peptide databases can be directly integrated into existing proteomic analysis pipelines with little to no modifications. Since peptides in the targeted database are a subset of the peptides in the full database, it is possible to directly map peptide identifications from the targeted database to their respective proteins in the full protein database (e.g., such mapping is done in TPP by explicitly specifying the path to the full database).



Peptide	Prob.
SLLDAQSAPLR	1.00
VVEELCPTPEGK	0.99
LLLAVICNASHK	0.99
CDVVSNTIVIAK	0.98
QCSVVDLTNTR	0.97
SYLFCMENSAEK	0.96
MAMVKDVDFK	0.95
...	...
...	...

Peptide	Prob.
SLLDAQSAPLR	1.00
VVEELCPTPEGK	1.00
LLLAVICWNASHK	1.00
CDVVSNTIVIALPK	0.99
GDAVFVIDLNR	0.99
SYLFCMENSAEK	0.98
MAMVKDVDFK	0.97
...	...
...	...

Figure 3.4: Basic targeted database workflow. Searching MS/MS spectra against a targeted peptide database results in improved sensitivity and increased peptide identifications in comparison to a search against a full protein database.

Results from the basic targeted database workflow

Improvement in peptide identification through the use of targeted peptide databases was measured by comparing the number of peptides identified from the targeted database search against those from the full protein database search. Tracking the improvement in the number of peptides from the different targeted databases (Figure 3.5), we can notice a clear trend of increased percent improvement, starting at 3.72% for the 0th percentile database and steadily increasing as we move to more and more restrictive targeted databases until reaching a peak at 10.75% for the 80th percentile database. After that, percent improvement begins rapidly decreasing and crosses into negative territory (i.e. fewer peptides are identified than that using the full protein database) for the most restricted (95th and 98th percentile) databases. This is consistent with our expectations, since peptide identification would initially benefit from the increased sensitivity that comes with a targeted database. However, as the targeted databases become too restrictive, we begin to lose true positive peptide identifications because they are no longer present in the search space, decreasing overall performance. In this data, the 90th percentile represents the level of search space restriction at which peptide loss due to search space incompleteness begins to outweigh the gain in peptide identifications due to increased sensitivity. Curves corresponding to the number of new peptides found (peptides not previously identified in the full database search) and the number of missed peptides (peptides identified in the full database search but missed in the targeted databases) are included in the figure.

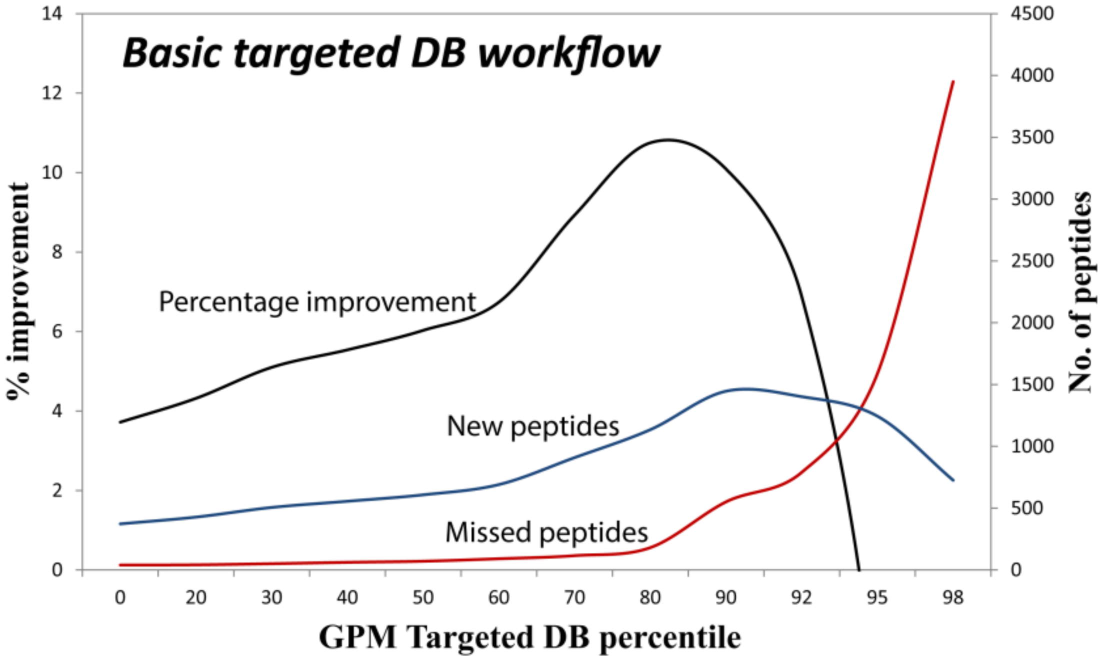
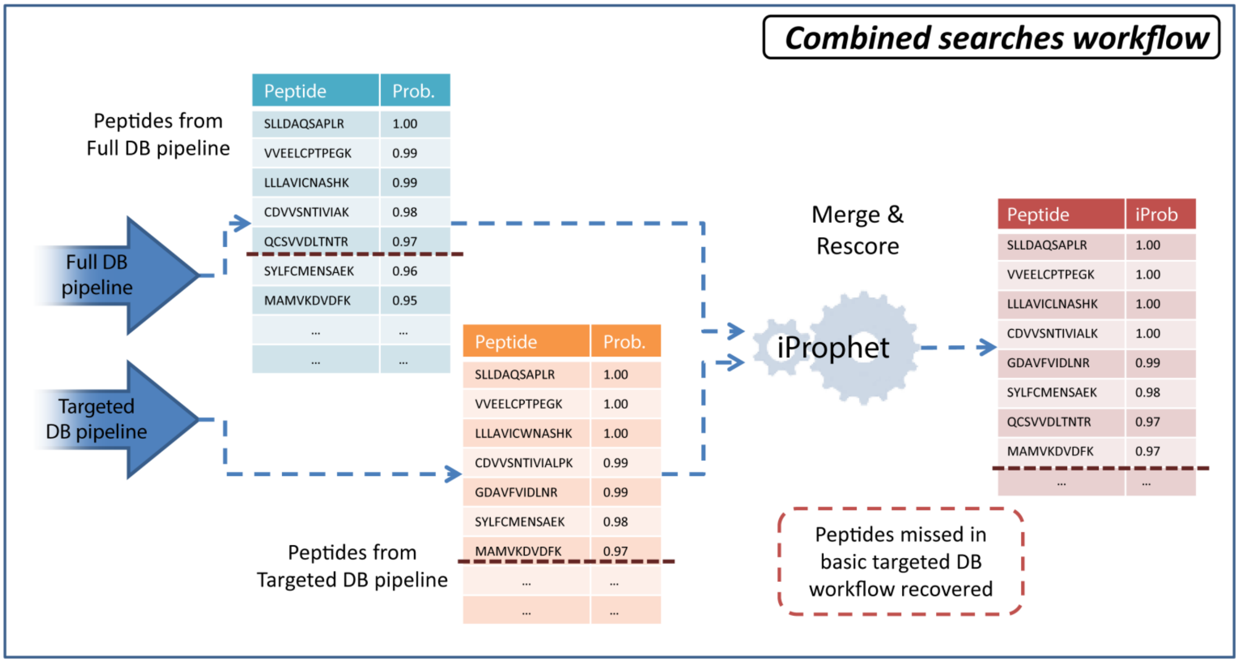


Figure 3.5: Results from the basic targeted database workflow. Percentage improvement from the basic targeted database workflow for varying levels of search space restriction (K562 lysate data). The number of missed and new peptides compared to the full database search is also plotted. Maximum improvement is obtained at the balance between high number of new peptides and low number of missed peptides.

Dealing with search space incompleteness: Combined searches workflow

The basic targeted database workflow provides an improvement in peptide identification over a typical full protein database search. However, as discussed above, targeted database searches also result in a number of peptides being missed due to search space incompleteness. Minimizing the loss of these peptide identifications is important for effective leveraging of targeted databases for proteomics analysis. Thus, we also designed a workflow that combines, using iProphet, the search results from the two independent searches - against the full databases and the targeted database. iProphet[18], a relatively recent addition to the Trans-Proteomic Pipeline, allows combining multiple levels of MS/MS evidence for scoring peptide identification, and specifically combining the results from multiple searches. While the typical use-case for the number of sibling searches (NSS) model, a statistical model implemented in iProphet for combining results from different searches, has been to combine results from multiple different database search tools, it contains no assumptions regarding the orthogonality of the different searches to be combined. Therefore, in this study we applied it to combine and rescore results from searches against the different databases (Figure 3.6).



Peptide	Prob.
SLLDAQSAPLR	1.00
VVEELCPTPEGK	0.99
LLLAVICNASHK	0.99
CDVVSNTIVIAK	0.98
QCSVVDLTNTR	0.97
SYLFCMENSAEK	0.96
MAMVKDVDFK	0.95
...	...
...	...

Peptide	Prob.
SLLDAQSAPLR	1.00
VVEELCPTPEGK	1.00
LLLAVICWNASHK	1.00
CDVVSNTIVIALPK	0.99
GDAVFVIDLNR	0.99
SYLFCMENSAEK	0.98
MAMVKDVDFK	0.97
...	...
...	...

Peptide	iProb
SLLDAQSAPLR	1.00
VVEELCPTPEGK	1.00
LLLAVICLNASHK	1.00
CDVVSNTIVIALK	1.00
GDAVFVIDLNR	0.99
SYLFCMENSAEK	0.98
QCSVVDLTNTR	0.97
MAMVKDVDFK	0.97
...	...

Figure 3.6: Combined searches workflow. Peptide identifications from the targeted database search and the full database search are combined using iProphet to recover peptides missed in the targeted database search.

However, the majority of rarely observed peptides identified with a high score (i.e. PeptideProphet probability) when using the full database, but missed in the targeted database search, would still score high enough to pass the specified FDR threshold (here, 1%) in the combined iProphet results. Thus, the computational strategy of performing two separate searches followed by combining the results using iProphet effectively retains the increased sensitivity advantages of using targeted databases while mitigating the negative potential impact of their incompleteness.

Multi-pass strategies involving searching against various search spaces and using different tools for more comprehensive interrogation of MS/MS data have been utilized previously[24–26]. However, the best way of estimating error rates for peptide identifications from these strategies has not yet been fully understood[1]. The combined search workflow presented here is partly related to such multi-pass strategies in that it utilizes multiple searches against differing search spaces. However, in our strategy the same spectra are searched against the different search spaces each containing its own set of decoys and the targeted database is created in unbiased way using external data. Furthermore, the search results are merged using iProphet previously extensively tested in a multiple database search tool setting. Thus, we believe the error rate concerns typical to multi-pass strategies are satisfactory addressed here.

Results from this combined searches workflow are shown in Figure 3.7. As can be seen, this strategy outperforms the basic targeted database workflow in terms of the peak level of improvement over the full protein database search, obtaining 12.5% improvement for the 92nd percentile database. It significantly reduces the number of missed peptides (i.e. peptides not identified because they are not in the targeted database) at the expense of only a slightly reduced number of additional (compared to the full database search) peptide identifications that one can obtain using the basic targeted database workflow. As a result, while the percent improvement decreases beyond the peak value, it does not drop into the negative territory even using the most restricted, 98th percentile targeted database. In fact, due to the merging of the two search results and rescoring carried out by iProphet, this workflow is not expected to result in a reduction in the number of peptide identifications compared to performing the full database search alone, irrespective of the degree of completeness of the targeted database.

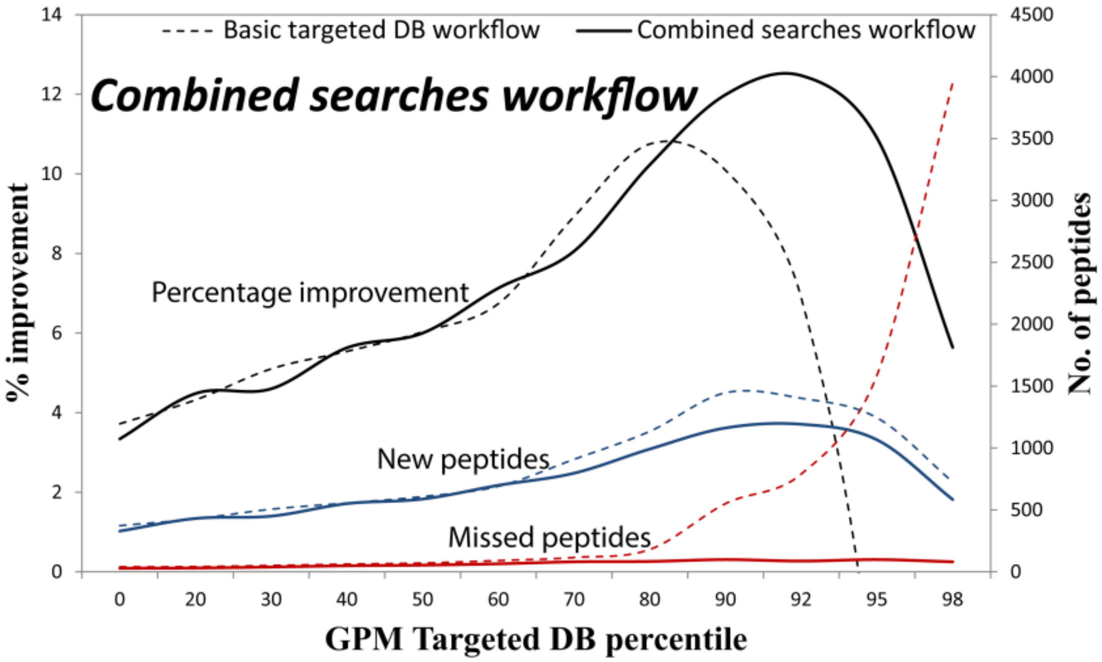


Figure 3.7: Results from combined searches workflow. Percentage improvement from the combined searches workflow, at varying levels search space restriction is plotted. The combined searches workflow outperforms the basic targeted database workflow, in maximum improvement, by minimizing the number of missed peptides.

As described earlier, the level of search space restriction corresponding to maximum improvement is a balance between the gain in peptide identifications due to improved sensitivity and the loss due to search space incompleteness. Since this combined searches workflow reduces the effect of search space incompleteness, the point of maximum improvement becomes more tightly linked with the increased sensitivity and hence is expected to occur at a higher percentile targeted database. Indeed, Figure 3.7 shows that the peak improvement in the iProphet based workflow occurs when using the 92nd percentile database, compared to the 80th percentile database in the basic targeted database search workflow.

In the above analysis, to observe the effects of combining these searches in isolation, all models in iProphet apart from the NSS model were turned off. A further comparative analysis with other models (except the NSP model) turned on was also performed (See Figure 3.8). While turning on other models did not provide much improvement in our data, in regular analysis it may be advisable to turn on other iProphet models and also use recommended strategies with iProphet, such as combining multiple search engine results[27], to take full advantage of any potential improvements.

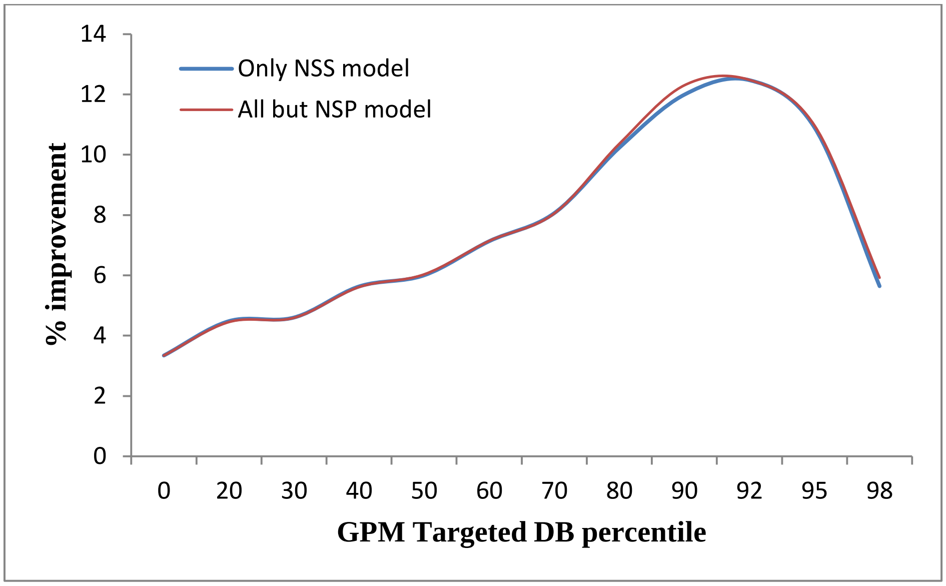


Figure 3.8: Comparison of iProphet runs with only NSS model vs. all models. In the rest of the study iProphet is run with only the NSS model. The comparison shows that the other models do not make much difference to the results.

The NSP model is not used in our analysis because the ProteinProphet tool, run next in the pipeline, also implements an NSP model. Therefore using the NSP model in iProphet would result in the same adjustment being applied twice to the data. However, if the analysis is planned to stop at the peptide level, or a different protein inference tool without an NSP model is used, the NSP model may be turned on in iProphet.

Source of missed peptides

As discussed in the above sections, the peptides missed during the targeted database workflows are assumed to be primarily due to search space incompleteness i.e. the missed peptides are absent in the smaller targeted database and not due to peptides being present in the database but getting lower scores. To validate this, the numbers of the two were plotted out for both the basic and combined workflows (Figure 3.9). Peptides missed due to a lower score are a very small, negligible portion of the total missed peptides, thus confirming our assumption.

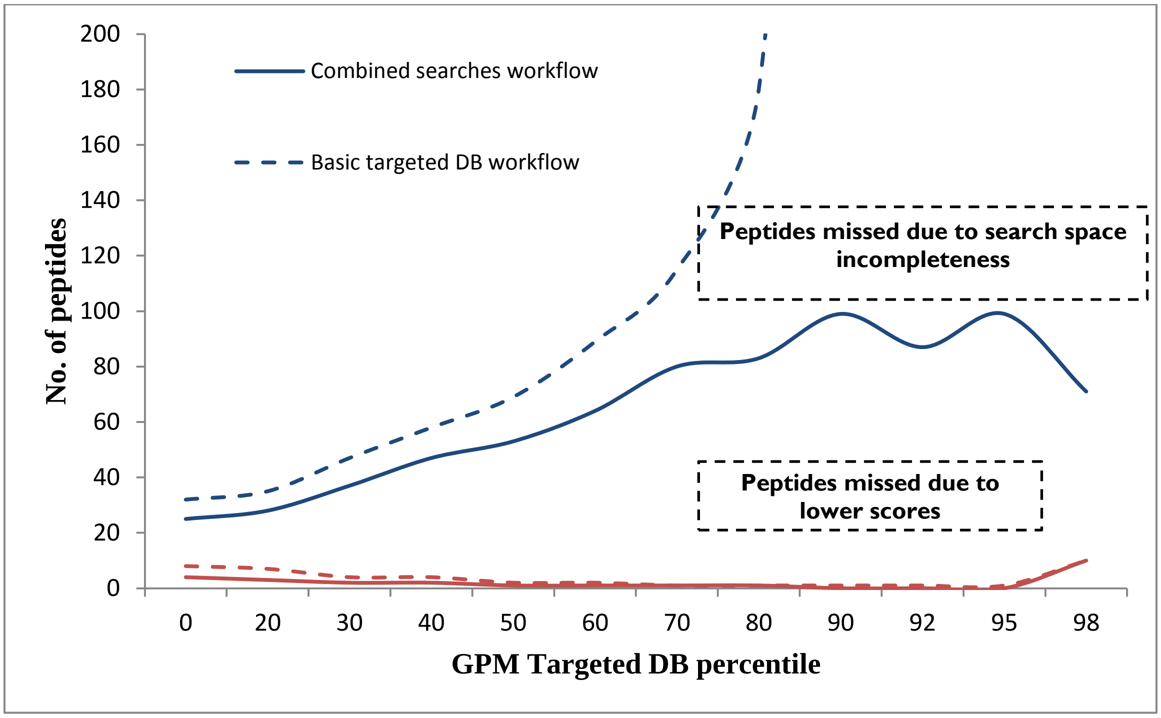


Figure 3.9: Source of missed peptides. Numbers of peptides missed in the targeted database searches due to search space incompleteness and the number of peptides present in the search space but were missed due to getting scored lower in the targeted search.

Applying workflows to other data

The performance of the computational strategies described above was further tested using three additional datasets (see Methods for detail): (i) data acquired on the same sample and instrument as above (K562 cell lysate, AB/Sciex 5600 instrument) but using a data independent acquisition (SWATH) strategy, with pseudo MS/MS spectra extracted using DIA-Umpire; (ii) data from an AP-MS experiment, in which the sample is enriched for a specific bait protein and its interacting partners; (iii) data from a deep proteome coverage experiment on a HeLa cell lysate containing about 60,000 peptide identifications (in contrast to about 8000 peptides identified in K562 dataset used above). These datasets represent a fairly diverse sampling of the different types of data that might be encountered in a modern proteomics experiment.

Figure 3.10 shows that the overall trends are largely similar across all datasets. In the DIA pseudo MS/MS data (Figure 3.10A), the improvement in peptide identification is even higher than that seen earlier in the corresponding conventional DDA data (compare with Figure 3.5 & Figure 3.7), with a peak improvement of 16.5% in the basic targeted database search workflow and 17.9% using combined full plus targeted searches. The de-convolution process applied to convert the multiplex DIA MS/MS spectra into pseudo MS/MS spectra results in spectra containing more noise than normal MS/MS spectra from DDA data. Peptide identification using noisier MS/MS spectra would be expected to benefit more from the increased sensitivity provided by targeted search space strategies. The AP-MS data (Figure 3.10B) shows a peak improvement of 11.3% using the basic targeted database search and 13.8% using the combined search workflow. While the same overall trends are observed, these data shows a higher degree of fluctuation which is likely due to a much smaller size of the dataset (~1000 peptide identifications).

While the increased sensitivity from a targeted database search results in better peptide identification scores, translation of these better scores into an increase in the number of peptide identifications passing a certain FDR threshold is dependent on the number of peptides in the sample that are in the 'grey-zone'. As we discussed previously[5], high quality deep proteome coverage samples are expected to contain less of such 'grey-zone' identifications, since they collect enough spectral data to confidently identify most identifiable peptides in the sample. Therefore the amount of improvement possible in such data is expected to be less than that observed for shallower coverage sample. Figure 3.10C shows that in the deep coverage HeLa dataset the maximum improvement is only about 1% using the basic targeted database search workflow and 2.4% using the combined search workflow.

Figure 3.10C also illustrates that in deep datasets like the one used here there are likely to be more rarely identified peptides (according the frequency of observation in GPMDB), leading to a higher number of missed peptides even at lower levels of search space restriction. This can be seen in the fact that the peak improvement occurs at lower percentile databases, 40th percentile for the basic targeted database search workflow and 50th percentile for the combined search workflow. At the same time, these results also demonstrate the robustness of the combined database search workflow. Even with a dataset where the basic workflow shows negative performance by the 60th percentile database, the combined search workflow provides some (albeit non-significant) improvement in the number of identified peptides across the entire set of targeted databases tested.

The workflows described in this study are, by design, neutral to the source of the targeted databases. In order to demonstrate this aspect, the workflows were also tested with targeted databases created using other types of information. Specifically, we used targeted protein sequence databases derived using RNA-Seq data[2]. The deep coverage HeLa cell lysate data was used as the MS/MS data for this analysis. Figure 3.10D shows that the results are similar to those seen with the GPMDB based targeted peptide databases. The basic targeted database search workflow results in a high number of missed peptides and essentially no overall improvement (0.4%), while the combined search workflow results in 1.7% overall improvement and less missing peptides.

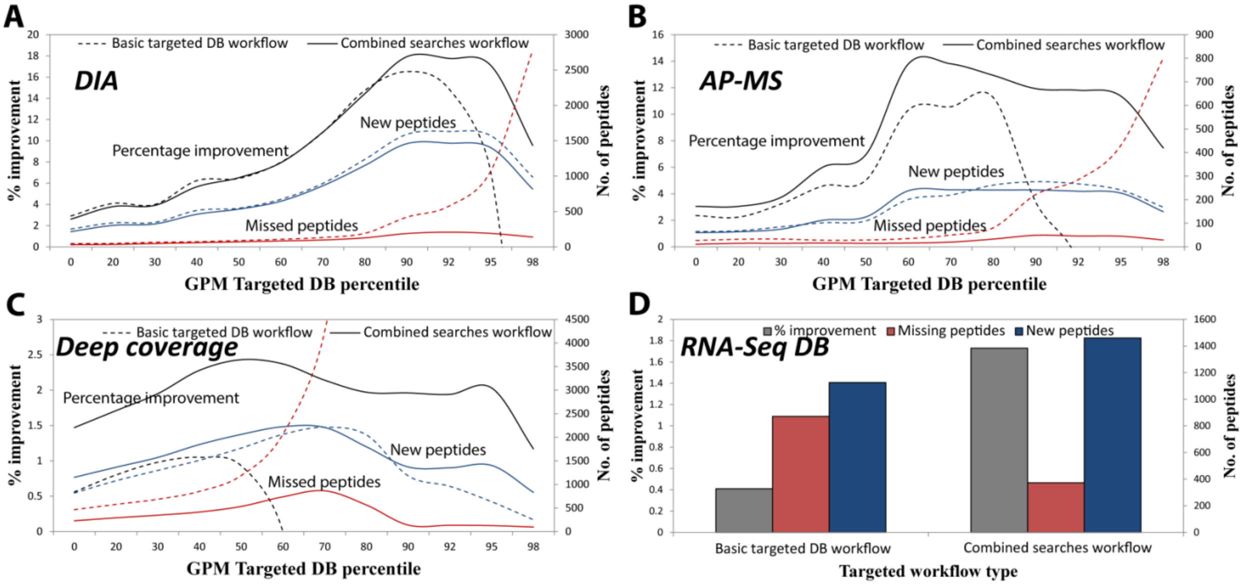


Figure 3.10: Applying workflows to other data. Results from applying the two workflows to (A) DIA extracted pseudo MS/MS spectra; (B) AP-MS data; (C) deep proteome coverage data. (D) Results of using a targeted DB derived from RNA-Seq transcript abundances.

Using targeted databases with an X! Tandem based pipeline

The results presented above were obtained the MSGF+ database search engine. The analyses were repeated using X! Tandem on the main dataset (K562 cell lysate; AB/Sciex 5600; DDA data). The overall trends for the basic targeted database search workflow were similar to those seen with MSGF+, with a peak improvement of 7.7% (Figure 3.11A). However, the number of peptides missed in the basic targeted database search compared to the full database search was notably more than that seen with MSGF+. A closer examination of the missed peptides revealed that a significant portion of them were missed in spite of actually being present in the targeted database. This issue was further investigated and was traced to an underlying problem of the E-value estimation approach implemented in X! Tandem.

X! Tandem (and several other search engines including Comet[28]) estimates E-values from the original scores (e.g. hyperscores in X! Tandem) using a null distribution fitted based on the non-top scoring (i.e. assumed to be random) matches to each spectrum. An insufficient number of random matches can cause the E-value estimation to be inaccurate or fail altogether. We have previously commented on the possibility of such issues arising with highly constrained database searches (e.g. searches with a very narrow precursor peptide mass tolerance)[1]. In this work, the additional reduction of the search space (via the use of targeted databases) further exacerbated the issue. Note, however, that the combined search strategy mitigated this problem as discussed above, resulting in a higher overall improvement, up to 15.5% (Figure 3.11B). It must also be noted that the problem of highly constrained search space was not an issue with MSGF+ searches altogether, which takes an alternative approach to computing the scores using the so-called generating functions[13,29] that is not as sensitive to the size of the search space.

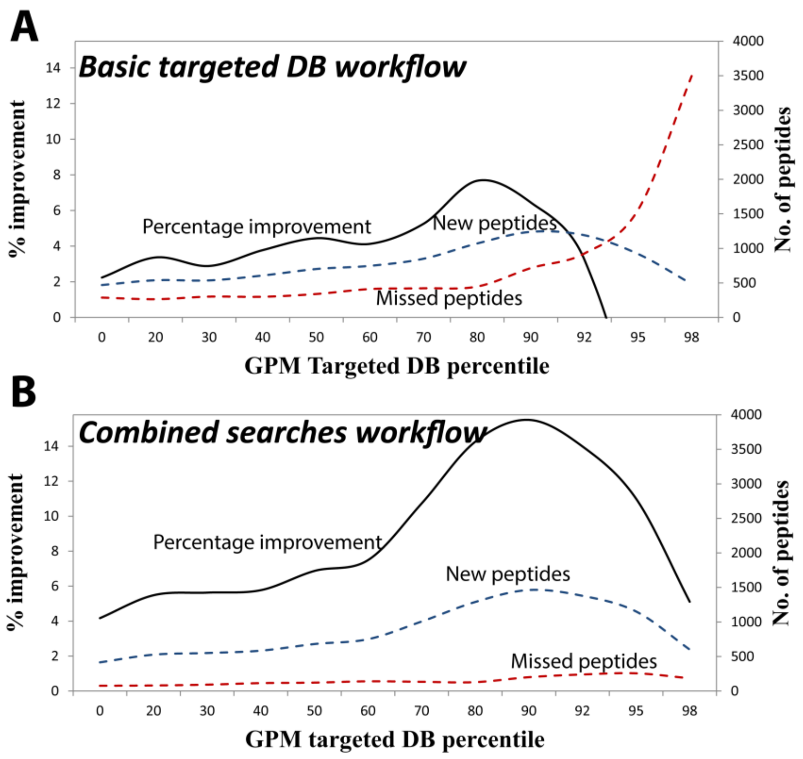


Figure 3.11: Targeted databases with an X!Tandem based workflow. Results from applying the targeted database workflows, using the X! Tandem database search engine on data from K562 cell lysate using (A) the basic targeted database workflow; (B) Combined searches workflow.

Selecting targeted database percentile thresholds

As can be seen from above results, the choice of percentile threshold for the targeted database is critical in determining the amount of improvement achieved. While in this study multiple percentile thresholds for the targeted databases were tested to identify the point of maximum improvement in peptide identification, it might be too time consuming to do routinely as part of a proteomic analysis pipeline.

Since the degree of search space completeness (number of peptides in the sample that are present in the targeted database) is a key determinant of improvement from the targeted database, the percentage of high confidence (1% FDR) peptides in the full database search retained at various targeted database percentile thresholds were analyzed (See Figure 3.12).

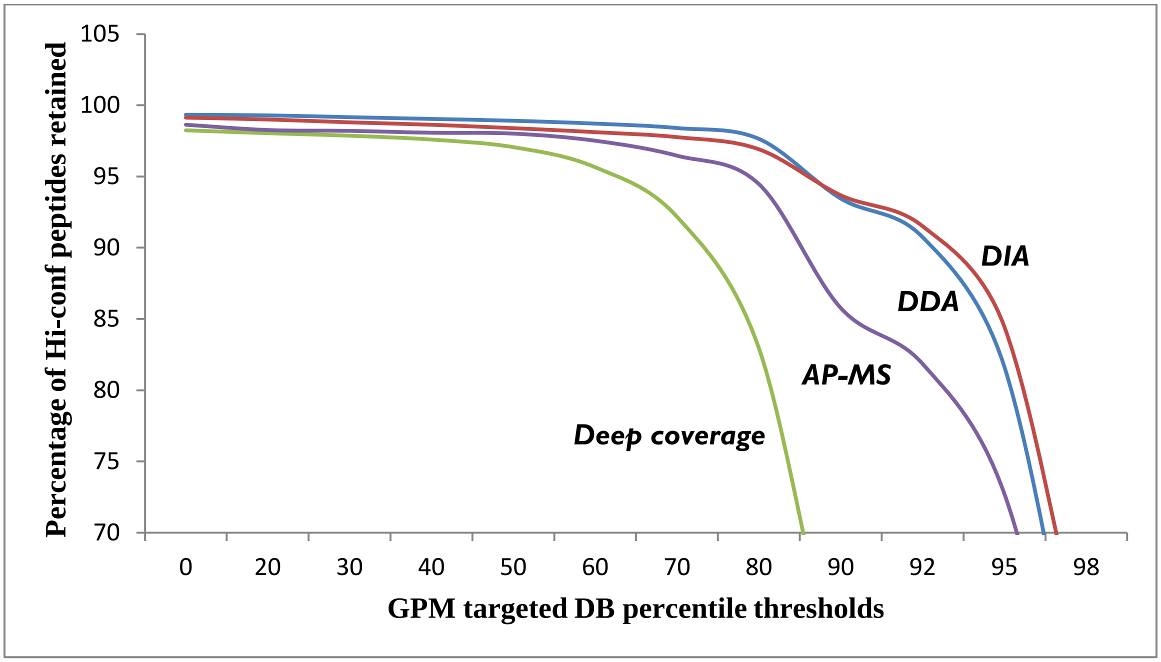


Figure 3.12: Percentage of high confidence peptides retained in targeted databases. Peptides identified at 1% FDR in a full database search that are retained in various targeted databases, for the different datasets used in our study.

Plotting percentage improvement achieved by the combined searches workflow on different datasets against the percentage of peptides retained, we observe that while the point of peak improvement in terms of percentile threshold varies among the different datasets, in terms of the percentage of peptides retained they are fairly clustered in the 90% - 97% region. These results, and the fact that the combined searches workflow is fairly robust to over-restriction (and consequential incompleteness) of search space, suggest that a lower percentage of peptides retained threshold around 90% or 92% would be a good empirical threshold to obtain improvements close to the maximum possible improvement. Since the combined searches workflow already requires performing a separate search with the full database, no extra time would be needed for this strategy. And in cases where running multiple searches would not be time consuming, multiple thresholds within this smaller range (say 90%, 92%, 95% and 97% of peptides retained) may be tested to more accurately determine the best targeted database to use.

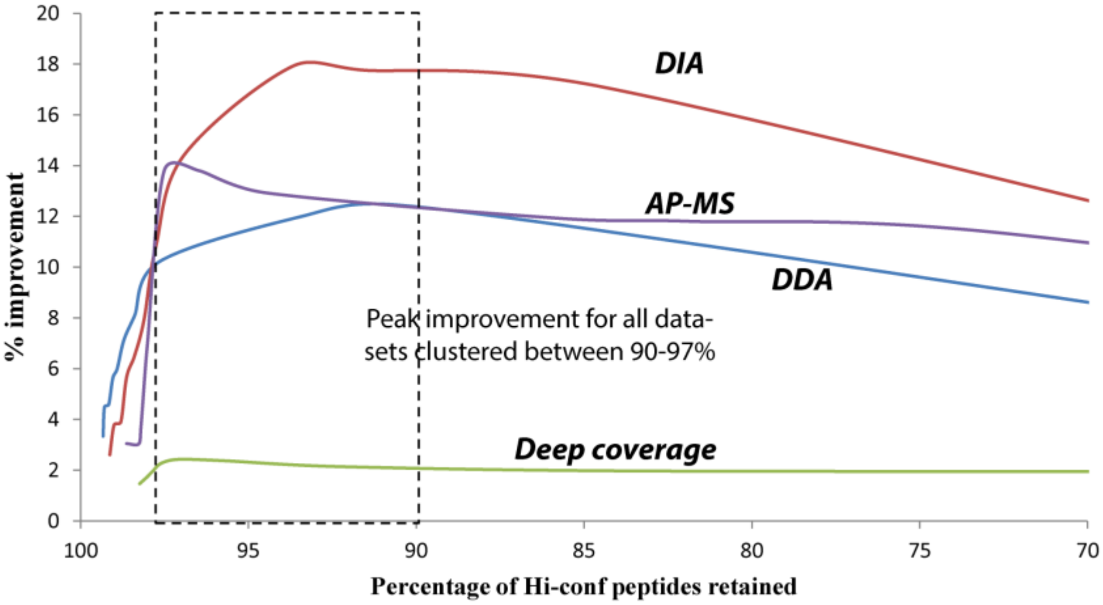


Figure 3.13: Selecting percentile thresholds for combined searches workflow. Percentage improvement for various datasets from the combined searches workflow plotted against the percentage of 1% FDR peptides from full database search, retained in the targeted databases. Peak improvement occurs in the 90%-97% range for all datasets.

The same analysis when repeated with the basic targeted DB workflow, which is more sensitive to search space incompleteness, shows the peak improvement points for all datasets occurring at a higher percentage of peptides retained (93%-98%) and also improvement dropping off much more steeply than with the combined searches workflow, which is to be expected. Based on the results and considering the reduced robustness of the workflow to search space over-restriction, a higher threshold of around 97% of peptides retained might be a good empirical threshold to use with the basic targeted database workflow. However, testing a couple of thresholds within this range would be highly advisable if it is practical.

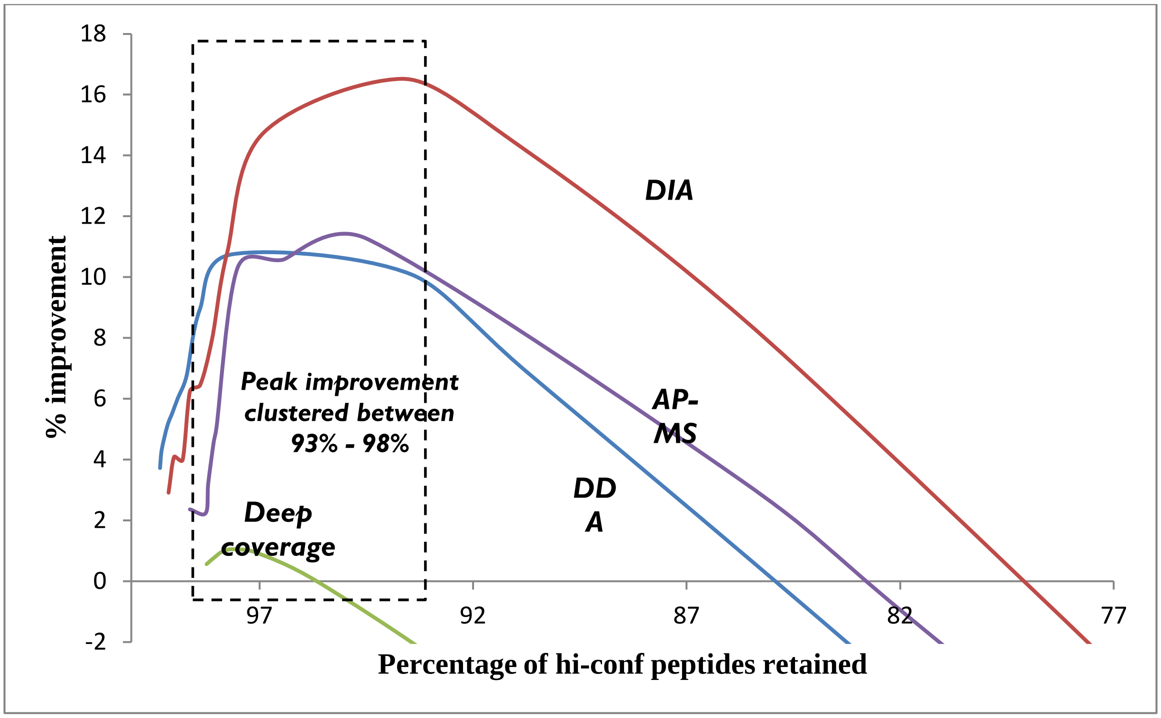
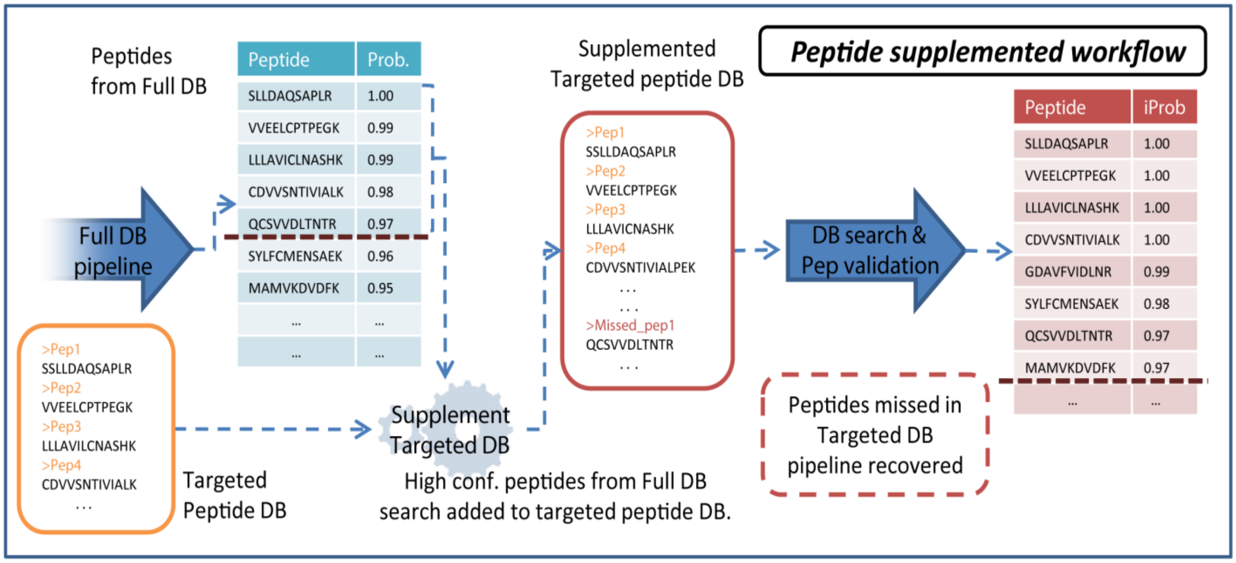


Figure 3.14: Selecting percentile thresholds for basic targeted workflow. Percentage improvement for various datasets from the basic targeted database workflow plotted against the percentage of 1% FDR peptides from full database search, retained in the targeted databases. Peak improvement occurs in the 93%-98% range for all datasets.

Peptide supplemented workflow

As an alternative to the combined searches workflow, a third workflow which we termed peptide supplemented workflow, was also tested. This workflow deals with the missing peptides problem by a straightforward approach of supplementing the targeted peptide database by directly adding the missing peptides to it. Peptides that were identified at 1% FDR in the full proteome database search were compared to the targeted peptide database file and any peptides not present in it were added to the database. Database search was then carried out using this supplemented database.



Peptide	Prob.
SLLDAQSAPLR	1.00
VVEELCPTPEGK	0.99
LLLAVICLNASHK	0.99
CDVVSNTIVIALK	0.98
QCSVVDLTNTR	0.97
SYLFCMENSAEK	0.96
MAMVKDVDFK	0.95
...	...
...	...

Peptide	iProb
SLLDAQSAPLR	1.00
VVEELCPTPEGK	1.00
LLLAVICLNASHK	1.00
CDVVSNTIVIALK	1.00
GDAVFVIDLNR	0.99
SYLFCMENSAEK	0.98
QCSVVDLTNTR	0.97
MAMVKDVDFK	0.97
...	...

Figure 3.15: Peptide supplemented workflow. High confidence peptides from the full proteome database search are compared to the targeted database and any peptides missing are added to the targeted database. Database search is then carried out using this supplemented targeted database.

One important point to note when using this workflow is that it is important to include decoys which passed the 1% FDR threshold in the list of peptides being added. Based on the assumptions of the target decoy approach, 1% of the target (forward) peptides that are being added to the targeted peptide database are estimated to be false positive matches. Not including the decoys that also scored in the same range would be akin to specifically removing high scoring decoys. That would lead to an under-estimation of FDR in the search results and hence bias the results.

But because of the addition of high scoring decoys, the final targeted database after supplementing would contain more decoys than target peptides, meaning there is a possibility of the FDR being slightly over-estimated in the results. But in the absence of a different, readily apparent, method to ensure non under-estimation of FDR, it was decided to take the cautious route of accepting a possibility of slight FDR over-estimation rather than allow any FDR under-estimation.

As can be seen in the results from applying this data to the K562 DDA data, (see Figure 3.16), the peptide supplemented workflow is indeed successful in minimizing missed peptides (lesser number of missed peptides than from combined searches workflow). While it was noticed there were still a small number of missed peptides, even though all missing peptides have been added to the search database, closer analysis of these results showed that these missed peptides are not due to search space incompleteness but rather mainly due to increased stringency in FDR filtering. The purposeful addition of high scoring decoys to the supplemented database makes the FDR filtering more stringent, resulting in the loss of a few of the peptide identifications which previously passed the FDR threshold in the full database search. A further effect of this stringency in filtering is that the number of new peptides too is reduced compared to the basic targeted database workflow (shown as the lightly dotted line). As a result, the overall improvement seen from the peptide supplemented workflow is also lower than that seen in the basic targeted database workflow.

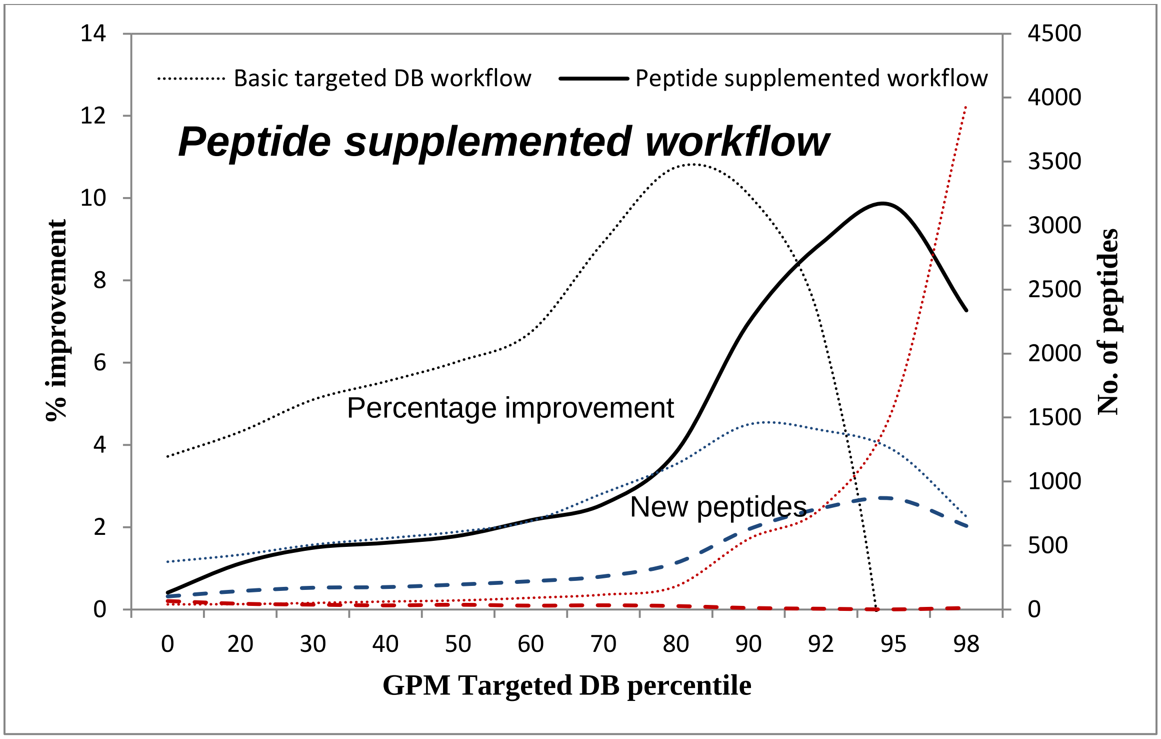


Figure 3.16: Results from the peptide supplemented workflow. Percentage improvement from applying the peptide supplemented workflow to K562 lysate DDA data. The peptide supplemented workflow is outperformed by basic targeted database workflow in maximum improvement. Even though the workflow is able to minimize the number of missed peptides, it is unable to identify many new peptides due to the increased stringency of the FDR threshold.

Thus, the peptide supplemented workflow would not be useful as a way to utilize targeted databases to obtain the most number of peptide identifications. But in scenarios where not missing peptides identified in the full proteome database search is important, this workflow might be of interest. Further, work on minimizing the FDR over-estimation seen in this workflow while taking care to avoid FDR under-estimation might help make this approach more useful.

CONCLUDING REMARKS

In this study, we have demonstrated the utility of targeted peptide databases derived with the help of GPMDB for providing a significant improvement in peptide identification in many types of MS/MS datasets. While the basic targeted database search workflow attempts to maximize the identification sensitivity, the combined database search workflow retains this increased sensitivity while also preventing any loss of peptides due to incomplete search spaces. Both workflows described in this study can be integrated into existing proteomics analysis workflows with little to no modifications. Furthermore, iProphet used here for integrating the results of different searches can be applied in other similar scenarios requiring merging of searches from different search spaces.

In addition to the two workflows described above, an additional workflow (the peptide supplemented workflow) was also designed and tested. While it was quite successful in reducing the number of missed peptides, it was seen to provide lesser improvement in comparison to the above two workflows (Figure 3.16). However, in the interest of potential future improvements to the workflow that might make it more useful, a description of the workflow and results from it has been included in the results.

One area of particular utility for targeted peptide databases could be in the identification of post-translational modifications (PTMs). Since searching for PTMs in MS/MS data can lead to exponential expansion of the search space, using a small targeted initial search space can be useful for maintaining sensitivity in the PTM expanded search space. This would be an alternative to approaches that improve PTM identification by post-search rescoring such as described in Li et al[30]. Proteogenomics, which typically involves creating large custom protein databases (i.e. obtained using six-frame translations of potential novel transcripts to databases of known sequences) is another area where the size of search space is seen to cause sensitivity issues[31–33]. The combined searches strategy described here could be extended to proteogenomics, by performing separate searches (e.g. first against the reference database, and then against a larger custom database of predicted sequences) prior to merging the results using iProphet. For proteogenomics applications, however, it will be necessary to perform subsequent searches only using spectra that remain unidentified based on the initial analysis (i.e. using the reference database of known sequences).Such a strategy would account for a much lower likelihood of identification of any novel peptide (as compared to known peptide), and ensuring that the estimation of posterior peptide probabilities in iProphet is performed separately for these two different types of peptides.

In addition to identification frequencies, GPMDB also stores spectral matching information for all the identified PSMs. As shown in Zhang et al[34], spectral library information can provide improved sensitivity in peptide identification in addition to that achieved just due to the search space reduction in spectral libraries. However, the spectral libraries provided by GPMDB earlier are no longer updated, and extracting the spectral information from GPMDB directly is technically difficult. In contrast, the method of creating targeted databases described in this work is relatively simple and can be re-done periodically as the GPMDB database continues growing in size.

REFERENCES

(1) Nesvizhskii, A. I. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J. Proteomics* **2010**, *73* (11), 2092–2123.

(2) Wang, X.; Slebos, R. J. C.; Wang, D.; Halvey, P. J.; Tabb, D. L.; Liebler, D. C.; Zhang, B. Protein identification using customized protein sequence databases derived from RNA-Seq data. *J. Proteome Res.* **2012**, *11* (2), 1009–1017.

(3) Craig, R.; Cortens, J. P.; Beavis, R. C. Open source system for analyzing, validating, and storing protein identification data. *J. Proteome Res. 3* (6), 1234–1242.

(4) Desiere, F.; Deutsch, E. W.; King, N. L.; Nesvizhskii, A. I.; Mallick, P.; Eng, J.; Chen, S.; Eddes, J.; Loevenich, S. N.; Aebersold, R. The PeptideAtlas project. *Nucleic Acids Res.* **2006**, *34* (Database issue), D655–D658.

(5) Shanmugam, A. K.; Yocum, A. K.; Nesvizhskii, A. I. Utility of RNA-seq and GPMDB protein observation frequency for improving the sensitivity of protein identification by tandem MS. *J. Proteome Res.* **2014**, *13* (9), 4113–4119.

(6) Craig, R.; Cortens, J. P.; Beavis, R. C. The use of proteotypic peptide libraries for protein identification. *Rapid Commun. Mass Spectrom.* **2005**, *19* (13), 1844–1850.

(7) Mallick, P.; Schirle, M.; Chen, S. S.; Flory, M. R.; Lee, H.; Martin, D.; Ranish, J.; Raught, B.; Schmitt, R.; Werner, T.; et al. Computational prediction of proteotypic peptides for quantitative proteomics. *Nat. Biotechnol.* **2007**, *25* (1), 125–131.

(8) Tsou, C.-C.; Avtonomov, D.; Larsen, B.; Tucholska, M.; Choi, H.; Gingras, A.-C.; Nesvizhskii, A. I. DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat. Methods* **2015**, *12* (3), 258–264.

(9) Mellacheruvu, D.; Wright, Z.; Couzens, A. L.; Lambert, J.-P.; St-Denis, N. A.; Li, T.; Miteva, Y. V; Hauri, S.; Sardiu, M. E.; Low, T. Y.; et al. The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. *Nat. Methods* **2013**, *10* (8), 730–736.

(10) Scheltema, R. A.; Hauschild, J.-P.; Lange, O.; Hornburg, D.; Denisov, E.; Damoc, E.; Kuehn, A.; Makarov, A.; Mann, M. The Q Exactive HF, a Benchtop mass spectrometer with a pre-filter, high-performance quadrupole and an ultra-high-field Orbitrap analyzer. *Mol. Cell. Proteomics* **2014**, *13* (12), 3698–3708.

(11) Cabili, M. N.; Trapnell, C.; Goff, L.; Koziol, M.; Tazon-Vega, B.; Regev, A.; Rinn, J. L. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* **2011**, *25* (18), 1915–1927.

(12) Flicek, P.; Amode, M. R.; Barrell, D.; Beal, K.; Billis, K.; Brent, S.; Carvalho-Silva, D.; Clapham, P.; Coates, G.; Fitzgerald, S.; et al. Ensembl 2014. *Nucleic Acids Res.* **2014**, *42* (Database issue), D749–D755.

(13) Kim, S.; Pevzner, P. A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.* **2014**, *5*, 5277.

(14) Craig, R.; Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20* (9), 1466–1467.

(15) Deutsch, E. W.; Mendoza, L.; Shteynberg, D.; Farrah, T.; Lam, H.; Tasman, N.; Sun, Z.; Nilsson, E.; Pratt, B.; Prazen, B.; et al. A guided tour of the Trans-Proteomic Pipeline. *Proteomics* **2010**, *10* (6), 1150–1159.

(16) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **2002**, *74* (20), 5383–5392.

(17) Choi, H.; Nesvizhskii, A. I. Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics. *J. Proteome Res.* **2008**, *7* (1), 254–265.

(18) Shteynberg, D.; Deutsch, E. W.; Lam, H.; Eng, J. K.; Sun, Z.; Tasman, N.; Mendoza, L.; Moritz, R. L.; Aebersold, R.; Nesvizhskii, A. I. iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Mol. Cell. Proteomics* **2011**, *10* (12), M111.007690.

(19) Trapnell, C.; Pachter, L.; Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **2009**, *25* (9), 1105–1111.

(20) Langmead, B.; Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **2012**, *9* (4), 357–359.

(21) Morgan, M.; Pages, H. Rsamtools: Binary alignment (BAM), variant call (BCF), or tabix file import.

(22) Carlson, M.; Pages, H.; Aboyoun, P.; Falcon, S.; Morgan, M.; Sarkar, D.; Lawrence, M. GenomicFeatures: Tools for making and manipulating transcript centric annotations.

(23) Wang, X.; Zhang, B. customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. *Bioinformatics* **2013**, *29* (24), 3235–3237.

(24) Tharakan, R.; Edwards, N.; Graham, D. R. M. Data maximization by multipass analysis of protein mass spectra. *Proteomics* **2010**, *10* (6), 1160–1171.

(25) Nesvizhskii, A. I.; Aebersold, R. Interpretation of shotgun proteomic data: the protein inference problem. *Mol. Cell. Proteomics* **2005**, *4* (10), 1419–1440.

(26) Ning, K.; Fermin, D.; Nesvizhskii, A. I. Comparative analysis of different label-free mass spectrometry based protein abundance estimates and their correlation with RNA-Seq gene expression data. *J. Proteome Res.* **2012**, *11* (4), 2261–2271.

(27) Shteynberg, D.; Nesvizhskii, A. I.; Moritz, R. L.; Deutsch, E. W. Combining results of multiple search engines in proteomics. *Mol. Cell. Proteomics* **2013**, *12* (9), 2383–2393.

(28) Eng, J. K.; Jahan, T. A.; Hoopmann, M. R. Comet: an open-source MS/MS sequence database search tool. *Proteomics* **2013**, *13* (1), 22–24.

(29) Howbert, J. J.; Noble, W. S. Computing exact p-values for a cross-correlation shotgun proteomics score function. *Mol. Cell. Proteomics* **2014**, *13* (9), 2467–2479.

(30) Li, S.; Arnold, R. J.; Tang, H.; Radivojac, P. Improving phosphopeptide identification in shotgun proteomics by supervised filtering of peptide-spectrum matches. In *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics - BCB'13*; ACM Press: New York, New York, USA, 2007; pp 316–323.

(31) Nesvizhskii, A. I. Proteogenomics: concepts, applications and computational strategies. *Nat. Methods* **2014**, *11* (11), 1114–1125.

(32) Blakeley, P.; Overton, I. M.; Hubbard, S. J. Addressing statistical biases in nucleotide-derived protein databases for proteogenomic search strategies. *J. Proteome Res.* **2012**, *11* (11), 5221–5234.

(33) Krug, K.; Carpy, A.; Behrends, G.; Matic, K.; Soares, N. C.; Macek, B. Deep coverage of the Escherichia coli proteome enables the assessment of false discovery rates in simple proteogenomic experiments. *Mol. Cell. Proteomics* **2013**, *12* (11), 3420–3430.

(34) Zhang, X.; Li, Y.; Shao, W.; Lam, H. Understanding the improved sensitivity of spectral library searching over sequence database searching in proteomics data analysis. *Proteomics* **2011**, *11* (6), 1075–1085.

CHAPTER 4
Challenges of protein false discovery rate estimation in very large proteomic datasets

INTRODUCTION

False discovery rate (FDR) estimation, at both peptide and protein level, is a critical tool in proteomics data analysis for discriminating between true and false positive identifications and for defining the threshold of acceptable and spurious identifications. While several different probabilistic approaches for estimating protein level FDR have been developed and used previously[1–4], the target-decoy strategy[5], which uses hits to decoy sequences to estimate the rate of random false positive target hits, remains the most widely used and simplest to implement approach for accurate estimation of FDR. But the implicit assumption used in the target-decoy strategy, that the number of decoy hits at a given score threshold is a close estimate of the number of false positive target hits at that same threshold, is conditional on the overall size of the 'potential decoy hits space' (i.e. the number of decoy sequences available to become a random hit during proteomic analysis) being approximately equal to that of the 'potential false positive target hits space'.

For protein level FDR estimation, the above condition is readily fulfilled in the analysis of relatively small datasets. However, with increasing sizes of the datasets, and consequently increasing number of true positive identifications being identified from those datasets, the size of the potential false positive space steadily decreases while the size of the potential decoy hits space remains largely the same (See Figure 4.1).

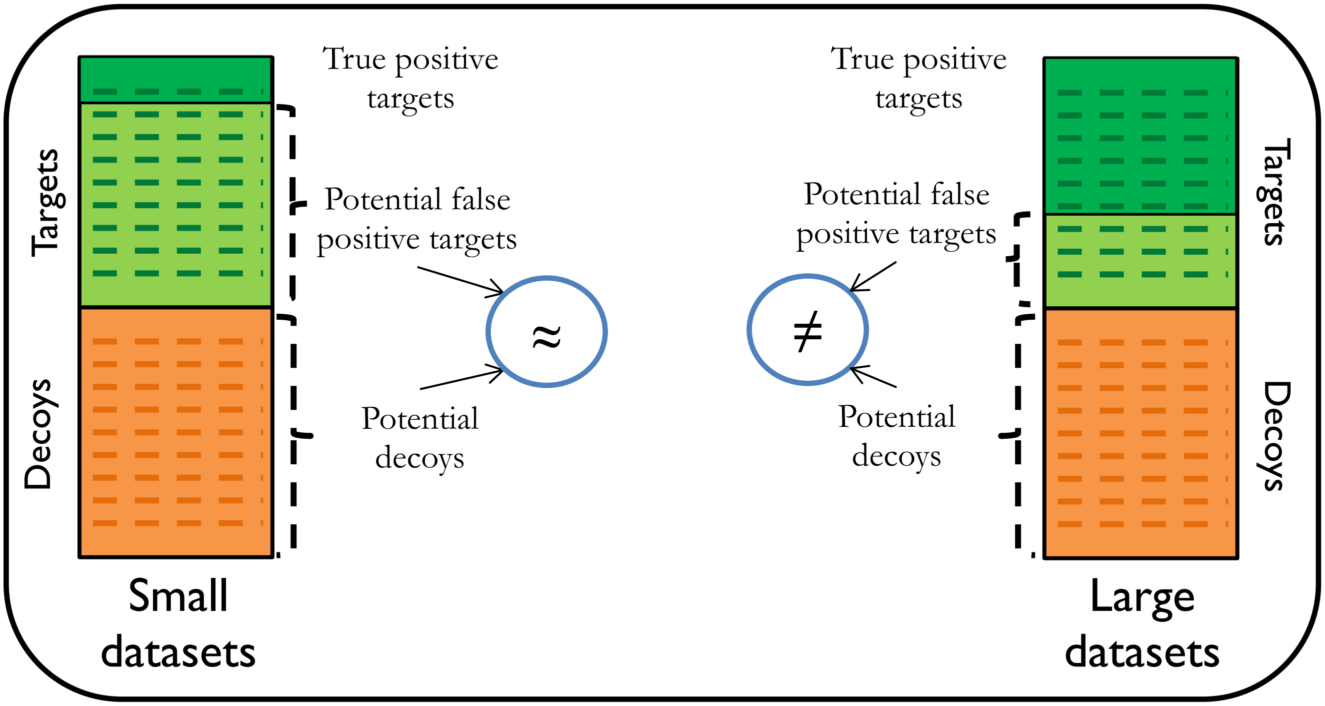


Figure 4.1: Decoy and False positive spaces inequality in large datasets. The number of potential false positives and potential decoy hits at the protein level are approximately equal in small datasets, allowing for the assumptions of the target-decoy strategy to hold true. But in large datasets, since a significant part of the target space is occupied by true positive identifications, the size of the potential false positives is smaller and no longer equal to the potential decoy space. As a result the assumption in the target-decoy strategy, that the number of decoy hits is a reasonable estimate of the number of false positive target hits, is no longer valid.

Due to the significant difference in the sizes of the potential decoy and false positive spaces, the assumption that the number of decoy hits closely estimates the number of false positive hits starts to break down in large datasets. The number of false positive hits is expected to only be a fraction of the number of decoy hits detected at a given score threshold. Applying the target-decoy strategy to estimate FDR in such datasets without accounting for this fact would result in an over-estimation of the number of false positives and consequently an over-estimation of the FDR.

Over the past decade, improvements in instrument speed and resolution have provided a steady increase in the number of proteins being identified by a single mass spectrometry experiment. But recently, studies have investigated integrating data from multiple experiments to create very large datasets. This approach was particularly exemplified by a recent article by Wilhelm et al.[6] demonstrating large scale integration of proteomic datasets, as part of the ProteomicsDB database, to create a draft of the human proteome. The development of such approaches, and the very large number of proteins being analyzed in them, increases the significance of the over-estimation problem in the basic target-decoy strategy for FDR estimation. In fact, recognizing the high impact of FDR over-estimation in such datasets some of the recent studies[6,7] dealing with such large datasets eschewed protein-level FDR altogether, preferring to utilize varying levels of peptide-level FDR filtering to control error rates. However, as pointed out in a recent article[8], such peptide-level filtering provides an inaccurate measure of protein level error rates and could result in larger than expected error rates at the protein level.

In this context, there is a critical need for robust methods that can accurately estimate protein false discovery rates in large scale proteomics datasets. In this chapter, three different methods, namely R-factor correction[9], the picked FDR approach[10], and MAYU[11], which have been proposed to address the problems of protein FDR over-estimation in large scale proteomics datasets are reviewed and compared. The results from applying these methods on large scale datasets and further challenges that will need to be addressed are also discussed.

METHODS

Datasets description

The methods described in this chapter were tested primarily on data generated by Kim et al[7]. This data consists of multiple replicates from 30 different human samples including 17 adult tissues, 7 fetal tissues and 6 purified primary haematopoietic cells, which taken together were used to create a draft map of the human proteome in their article. The data was generated on LTQ-Orbitrap Elite and LTQ-Orbitrap Velos instruments following fractionation by SDS-PAGE and basic reversed-phase liquid chromatography. The number of proteins identified in each individual sample ranged from about 1000 proteins to about 8000 proteins. All of the samples, when combined at the PSM level and protein inference performed for the combined data, yielded close to 15,000 proteins. While combined protein inference of PSMs from such diverse tissues might not be biologically sound, combined protein inference is justified here since we are only interested in the behavior of the large datasets and not biological inferences in our analyses.

Additionally, data generated by Geiger et al.[12] was also used for some analyses. This data consisted of MS/MS spectra collected from cell lysates of 11 different human cell lines (A549, GAMG, HEK293, HeLa, HepG2, Jurkat, K562, LnCap, MCF7, RKO and U2OS) in 3 replicates each; on a LTQ-Orbitrap Velos mass spectrometer. The number of proteins identified from each cell line ranged between 6000 to 7000 proteins and combined protein inference on all the data yielded close to 11,000 proteins.

MS/MS data analysis pipeline

Proteomic analysis of the MS/MS data was performed using the X!Tandem search engine[13] (JACKHAMMER TPP 2013.06.15.1) for database searching. Kim et al. data was searched against the Ensembl[14] (v.78) human proteome database with reversed protein decoy sequences while the Geiger et al. data was searched against the Uniprot human proteome[15] (as retrieved on March 4th 2015) with reversed peptide decoys appended. Searches were run with a precursor mass error of 15 PPM, fragment mass error of 20 PPM, Methionine oxidation as a fixed modification and only fully tryptic peptides with a maximum of 1 missed cleavage allowed. Trans-proteomic pipeline[16] (TPP v.4.7.0 POLAR VORTEX) was used for downstream analysis with the same parameters as described in the previous chapter.

Combined protein inference was performed by running ProteinProphet[1] on the all the pepXML files resulting from running TPP on individual samples. FDR estimation was performed by sorting identifications on the basis of ‘best peptide probability’, after first filtering out proteins groups having protein identification probability (*localPw* in the ProteinProphet results) less than 0.9. The decoy adjustment methods (R-factor, Picked FDR and MAYU) for FDR estimation were all implemented in an in-house R-script and these were applied to parsed protXML result files from ProteinProphet (in TPP). While the MAYU approach was already implemented and available as a Perl script, the output of it was not amenable to our analysis. Hence it was re-implemented in a custom R script.

RESULTS AND DISCUSSION

Best peptide probability vs Protein probability

The ranking score, or the score used to rank protein identifications before estimating false discovery rate, has a major effect on the discriminatory power of the analysis and the number of high confidence proteins identified. As mentioned previously in Chapter 2 of this dissertation, the best peptide probability has been seen to be a better ranking score than the protein probability that is computed by ProteinProphet. Both ranking scores were compared for 4 different samples of different sizes to verify this (See Figure 4.2). As can be seen in the figure, there is a marked difference in the number of proteins identified at 1% FDR between these approaches and the differences are more pronounced in larger datasets.

This is likely due to the approach taken to calculate protein identification probabilities by ProteinProphet [$P(Protein) = 1 - \prod_i (1 - P(Pep_i))$], in which as more peptides are mapped to a protein its protein probability moves closer to 1. It is easy to note that even spurious identifications might achieve a high probability if enough low scoring peptides map to them, which is often the case in large datasets with numerous peptide identifications. This leads to an inflation in the number of decoy hits with high protein probabilities, thereby reducing the effectiveness of the ProteinProphet protein probability as a discriminating score. Best peptide probability on the other hand does not suffer from this drawback and retains its effectiveness as a discriminating score even in large datasets. This can be observed in Figure 4.3, showing the density plots of best peptide probability and protein probability in the combined dataset from Geiger et al, as a pronounced peak for decoy hits in the protein probability distribution but not in the best peptide probability distribution. Based on this confirmation, the best peptide probability was used as the ranking score for all further analyses in this study.

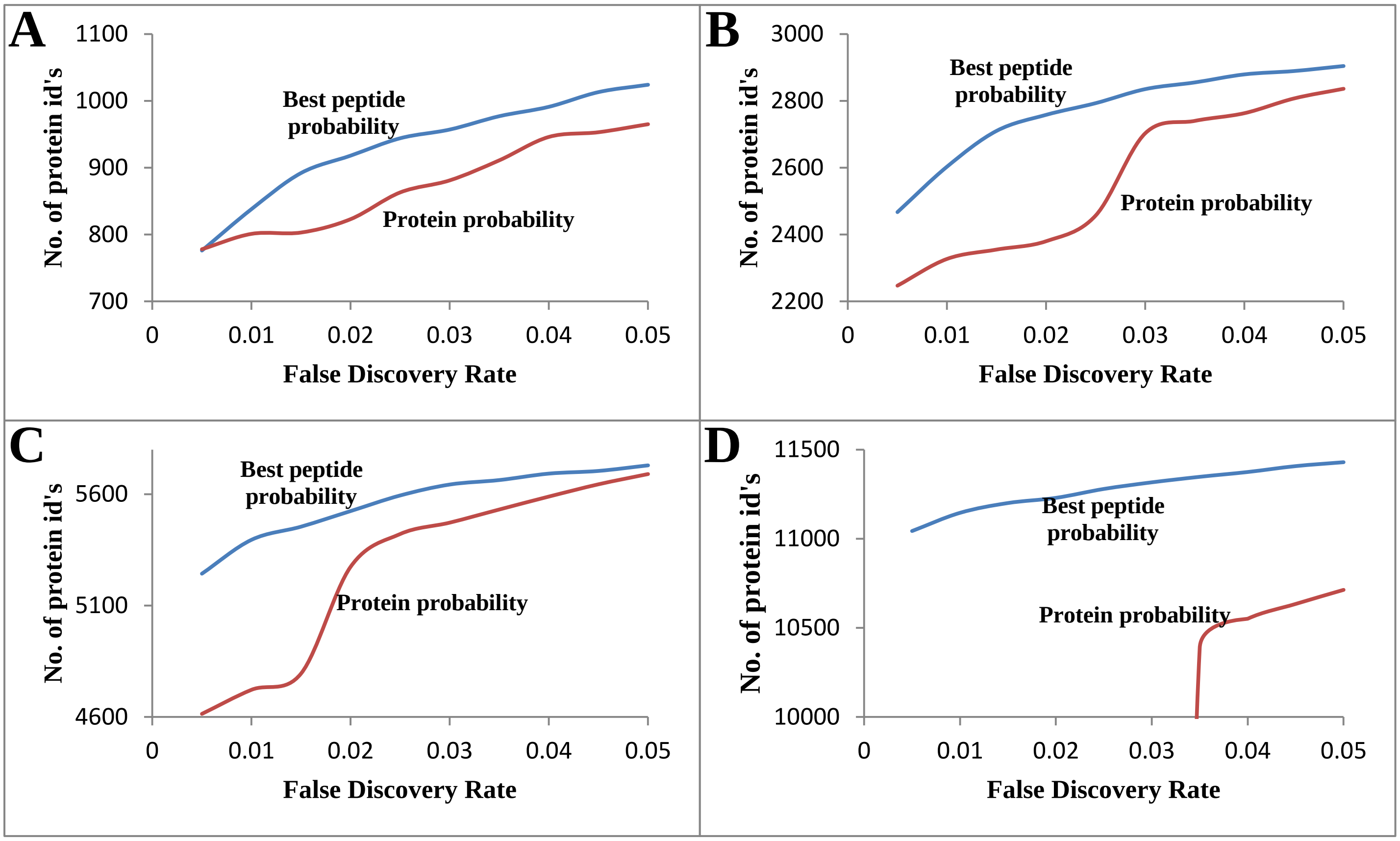


Figure 4.2: Best peptide probability vs. protein probability at different dataset sizes. The difference in number of proteins from best peptide probability and protein probability at 1% FDR increases as size of the dataset increases. Results are shown for dataset sizes ~800 (A), ~2600 (B), ~5300 (C) & ~11100 (D) proteins identified at 1% FDR.

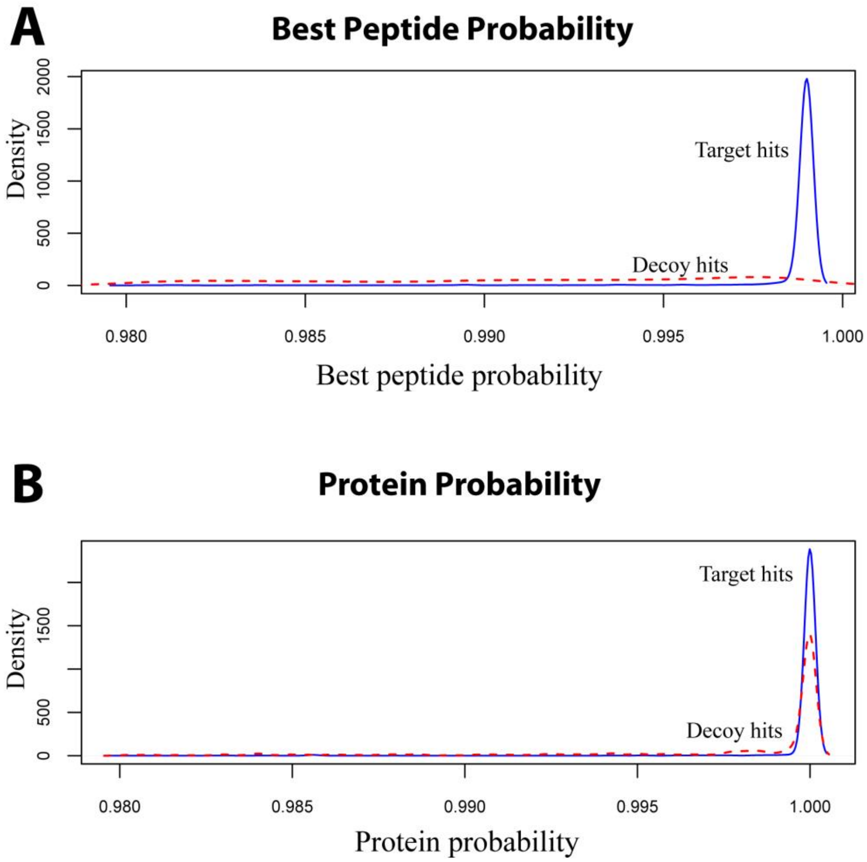


Figure 4.3: Best peptide probability vs. protein probability – density distributions. Density profiles of best peptide probability (A) and protein probability (B) of high scoring protein identifications shows more clustering of proteins closer to 1 with protein probability as the ranking score, observed as a narrower peak for target hits and more pronounced peak for decoy hits in (B) than (A). Hence protein probability is a less effective discriminatory score.

Adjusted FDR estimation methods

i) R-factor correction

As mentioned previously, due to unequal sizes of the potential decoy and false positive target spaces in very large datasets, the number of false positives expected to be present at a given score threshold would only be a fraction of the number of decoys detected at the same threshold.

Therefore applying a basic target-decoy strategy in such datasets would lead to an over-estimation of the FDR. However it would be possible to correct for this bias by estimating the fractional relationship between decoys and false positive targets.

The R-factor correction, which was briefly described previously in Chapter 2 of this dissertation, is based on the assumption that in the low probability range, all target hits are likely to be false positive identifications. Therefore in a dataset with nearly equal decoy and false positive spaces, the numbers of decoy and target hits in the low probability range must also be near equal, or in other words their ratio must be approximately 1. Conversely, in a dataset with unequal decoy and false positive spaces the ratio between target and decoy hits at low probabilities, the 'R-factor', would provide a close estimate of the fractional relationship between decoys and false positives, as illustrated in Figure 4.4. During FDR calculation, the number of decoys detected is adjusted by the R-factor to provide a more accurate estimate of the number of false positives and thereby a more accurate estimate of FDR.

To set the threshold below which low probability target and decoy hits are considered for estimating R-factor, multiple thresholds in the 0.2 - 0.4 range were tested and the one providing the lowest R-factor was selected. This allows for inclusion of the entirety of the low probability region 'hump' seen in both the target and decoy density distributions.

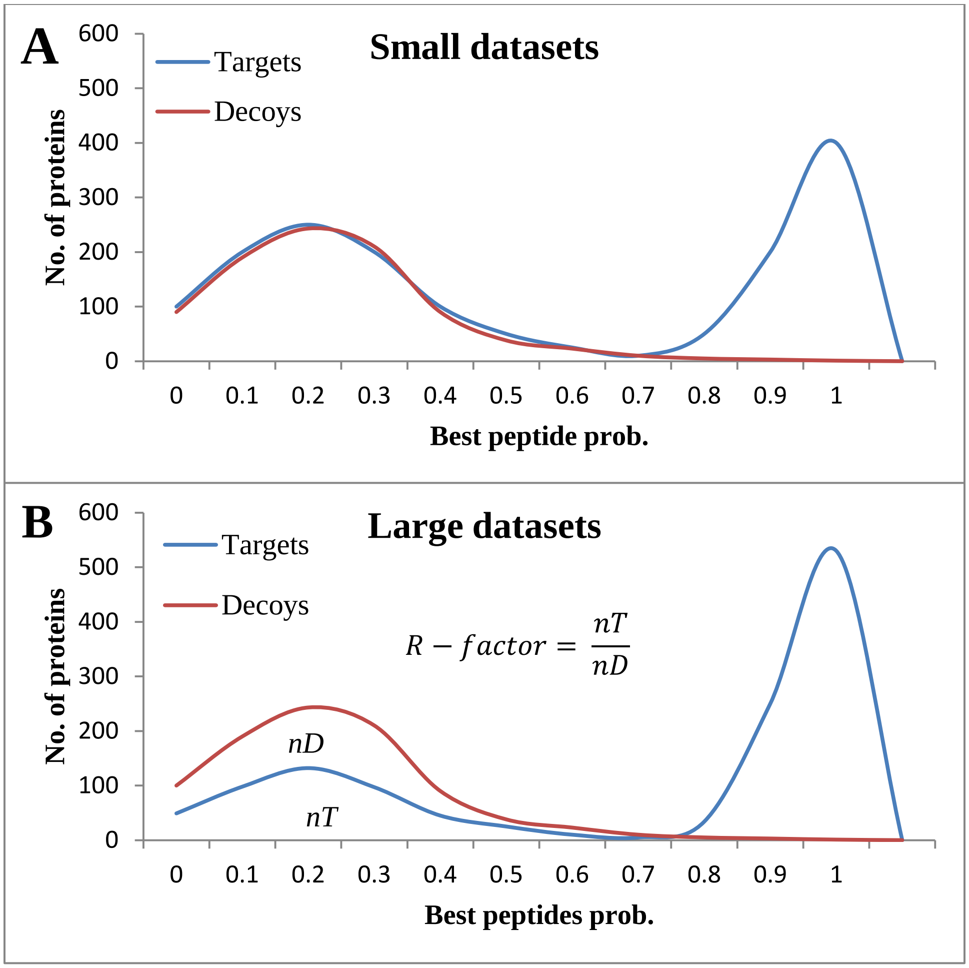


Figure 4.4: Toy example to illustrate R-factor correction. In small datasets (A) with nearly equal decoy and false positive spaces, the decoy and target hits numbers are nearly equal. In larger datasets (B) with more unequal decoy and false positive spaces, the ratio of number of targets to decoys in the low probability range, 'R-factor' gives an estimate of the fractional relationship between decoys and false positive target hits in the high probability range.

ii) Picked FDR approach

The picked FDR approach, described in a 2015 article by Savitski et al., attempts to address FDR over-estimation due to unequal decoy and false positive spaces by un-biasedly reducing the both the decoy and false positive spaces so as to equalize their sizes. To do this, the picked FDR approach treats the target sequence and the decoy sequence, created by reversing the target sequence, as a pair. After MS/MS data is processed through the proteomics pipeline, the ranking scores (best peptide probability in this case) of the target and decoy sequences, in each target decoy pair, are compared with each other. The sequence with the higher score is ‘picked’ to be retained while the sequence with the lower score is removed from the results.

While this process reduces both target and decoy spaces, true positive target hits, which are likely to be higher scoring, are more likely to be ‘picked’. If a decoy sequence is ‘picked’, its corresponding target sequence must be lower scoring and hence likely to be a false positive hit. It is claimed that in this way the decoy space is reduced till it is almost equal in size to the false positive target space. FDR estimation is then carried as normal using only the ‘picked’ targets and decoys.

iii) MAYU

The MAYU strategy works on the assumption that false positive PSMs are uniformly distributed over the target space, but some of these false positive PSMs may be mapped to target hits which are also identified independently by other high confidence true positive PSMs. In other words, not all target hits that contain a false positive PSM are false positive identifications themselves. The over-estimated number of false positives in a basic target-decoy strategy is hence explained to be estimating the number of false positive PSM containing target identifications and not the number of false positive target identifications itself.

MAYU uses a hyper geometric distribution to model the expected number of false positive target hits, given the number of target hits, number of decoy hits and the total number of targets in the database. This modeled number false positive target hits is equivalent to the r-adjusted number of false positives estimated by the R-factor approach and the number of ‘picked’ decoys determined by the picked FDR strategy. And just as in those approaches this adjusted value is used to estimate a more accurate FDR.

Comparison of FDR estimation methods

The three adjusted FDR estimation methods were applied to the combined datasets created from the MS/MS data of 11 cell lines (Geiger et al.) and data comprising the human proteome draft (Kim et al.). The number of proteins identified at 1% FDR from each of the three methods was compared with that obtained from a basic target-decoy strategy for FDR estimation. The results of this comparison are shown in Figure 4.5.

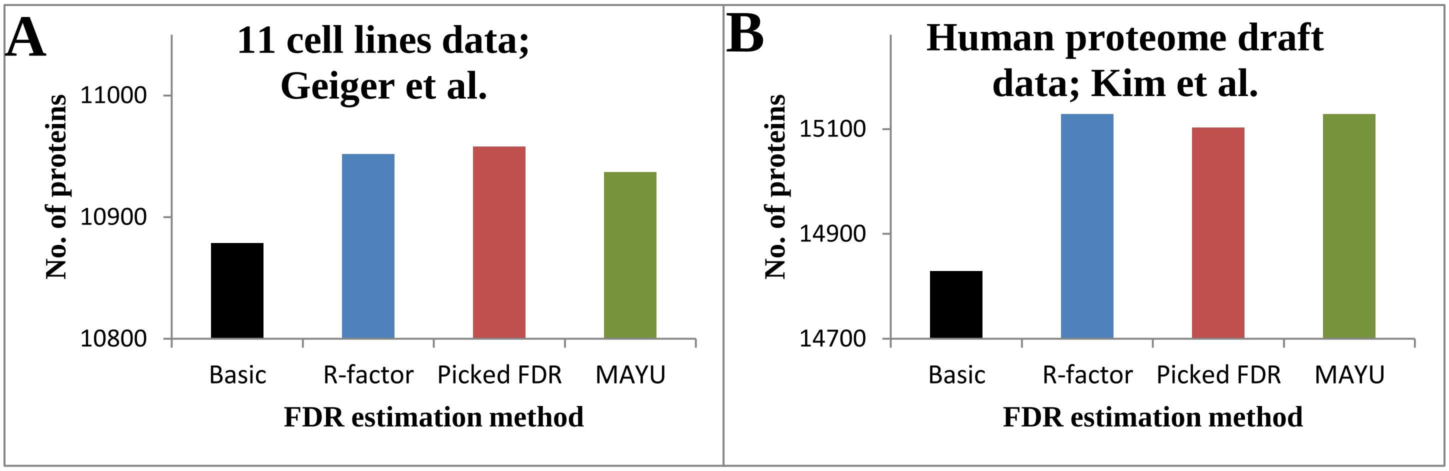


Figure 4.5: Effect of adjusted FDR estimation. Comparison of number of proteins identified at 1% FDR by various FDR estimation methods from 11 cell lines data by Geiger et al. (A) and from Human proteome draft data by Kim et al. (B). The adjusted FDR estimation methods provide a modest improvement over the basic target-decoy strategy.

As can be seen in these results, all three adjusted FDR estimation methods provide modest improvement over the basic target-decoy strategy. The improvement seen is somewhat higher in the larger human proteome draft data, which is understandable, since the effects of FDR over-estimation in the basic target-decoy strategy are expected to be more pronounced in the larger dataset. The results from the three methods are fairly comparable and no trends of significant difference can be inferred from these results.

In further analysis, the three adjusted FDR estimation methods were applied to individual samples from the human proteome draft dataset. These samples ranging in size from about 1000 proteins to about 8000 proteins allow us to observe the effects of these methods across a wide range of dataset sizes. However, owing to the small size of these samples the effects in terms of number of proteins identified at 1% FDR are likely to be quite small and hard to tell apart. Therefore, the comparison was performed in terms of the 'decoy adjustment ratio' which would be more sensitive to variation.

We define the decoy adjustment ratio, as the ratio of the effective number of decoy hits considered by the adjusted FDR estimation to the total number of decoy hits in the data. In the picked FDR approach this would be ratio of the number of 'picked' decoys to the total number of decoys (referred to hereafter as P-ratio), while in MAYU this would be the ratio of the expected number of false positives (from the hypergeometric distribution) to the total number of decoys (referred to hereafter as M-ratio). In the R-factor approach, this would be the R-factor itself. While the R-factor is constant across all FDR thresholds, the decoy adjustment ratio for the other two methods would vary across FDR thresholds. In our analysis, the ratios were compared at 1% FDR, with the plot of the comparison shown in Figure 4.6.

As expected, the decoy adjustment ratios show a steady decrease with increase in size of the datasets, since the false positive over-estimation is more pronounced in larger datasets. It is also seen that the P-ratio has higher values (lower is better) than the other methods, R-factor and

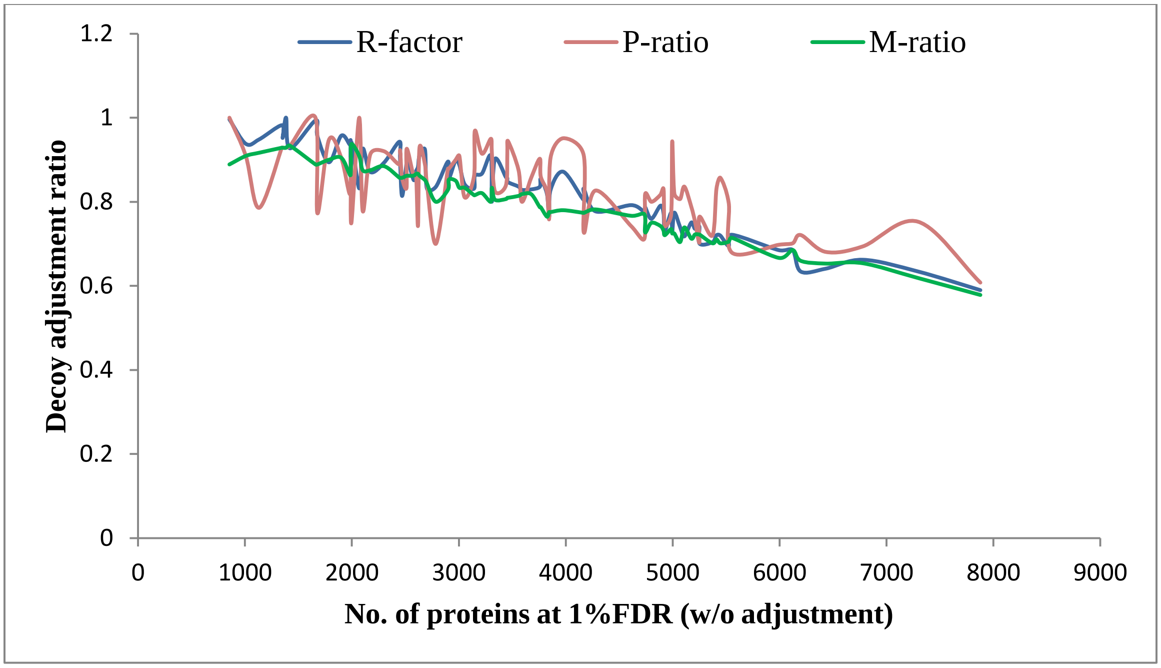


Figure 4.6: Comparison of decoy adjustment ratios (at 1% FDR) across various dataset sizes. Dataset sizes are shown in terms of number of protein identifications at 1% FDR (by basic target-decoy strategy).

M-ratio, in the largest few datasets. However, we believe that this is due to a difference in how protein groups are treated which is discussed in the next section. Additionally the M-ratio (from MAYU) is seen to have a smoother trend than the R-factor or P-ratio which show more variation. This is likely due to the fact that the M-ratio is the result of the hypergeometic model while the R-factor and P-ratio are empirically determined from the number of targets and decoys potentially causing more fluctuation. But overall, even in this comprehensive analysis, no significant differences in results between the different methods can be observed.

Impact of protein grouping on Adjusted FDR estimation

Protein grouping is an important aspect of protein inference. Protein groups are sets of proteins having shared peptides (peptides mapping to multiple proteins). While there are multiple possible scenarios in protein grouping[17], of particular interest to us are indistinguishable groups of proteins. In an indistinguishable group, all proteins in the group are only identified based on the same group of shared peptides. In such scenarios, based on the peptide evidence alone it can only be inferred that at least one of the proteins in the group are present in the sample. But there is no evidence to determine which of them, or if more than one of them, is present in the sample. Typically, during protein inference such groups are collapsed into a single protein identification, with one protein from the group chosen as a representative protein for the group.

When implementing the picked FDR approach, only the representative proteins from each indistinguishable group were compared with their respective target or decoy pair. Therefore, decoy proteins whose corresponding target is part of an indistinguishable protein group with a higher score would get removed if their corresponding target is the representative protein of the group. However if the corresponding target is not a representative protein, even though the target has just as much evidence for its presence as the representative protein, the decoy would be picked. Therefore, collapsing indistinguishable protein groups might be causing a bias against removal of decoys leading to the higher P-ratios in large datasets seen in Figure 4.6. On the other hand, treating every protein in an indistinguishable group as identified at the same score, especially considering that there are a lot more target hits collapsed into indistinguishable groups than decoys, might cause an inverse bias towards the removal of decoys hits.

To investigate this, the picked FDR approach was re-implemented with target-decoy picking performed before collapsing of indistinguishable groups and the P-ratios from this re-implementation were compared with those from previous results (Figure 4.7). As expected, performing target-decoy picking before collapsing of indistinguishable groups lowers the P-ratios. But comparing the two curves with decoy adjustment ratios from the other methods, it seems that the optimal P-ratio should lie somewhere between those from the two implementations (on the assumption that decoy adjustment ratios from all three methods should be approximately equal). Further research may be required to identify the best practice for treating indistinguishable protein groups in this situation.

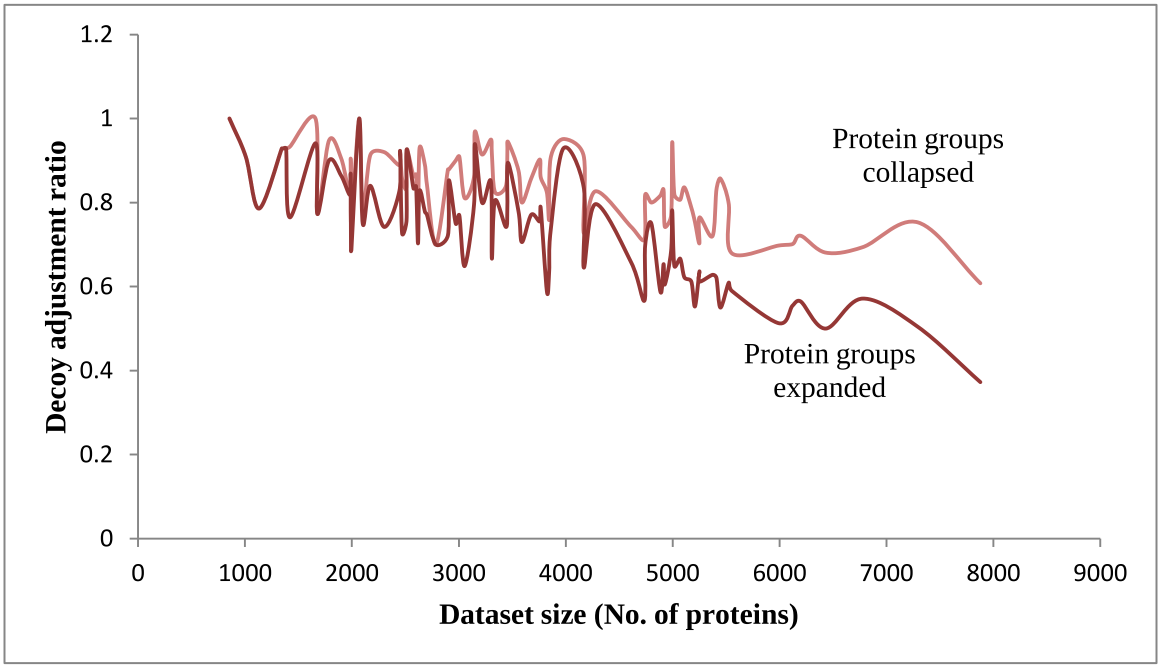


Figure 4.7: Comparison of P-ratio with different treatments of protein grouping. Not collapsing indistinguishable groups before target-decoy picking is seen to lead to lower P-ratios.

Another analysis that was performed was to evaluate the impact of protein grouping on the MAYU strategy. As mentioned previously, the MAYU model estimates the expected number of false positive target hits based on the number of target hits, number of decoy hits and the total number of targets in the database. When using a database with redundant protein sequences, such as Ensembl that was used in this analysis, using the exact total number of sequences in the database would be roughly equivalent to not collapsing protein groups while only using the non-redundant number of sequences in the database is roughly equivalent to treating the protein groups as collapsed. In the results reported previously, the MAYU model was implemented using the number of non-redundant protein sequences. But it was also re-implemented using the redundant number of proteins and the results compared (Figure 4.8). As can be seen in the figure, using a redundant number of proteins causes a very significant increase in the M-ratio, making the MAYU model barely more effective than the basic target-decoy strategy. Even though there

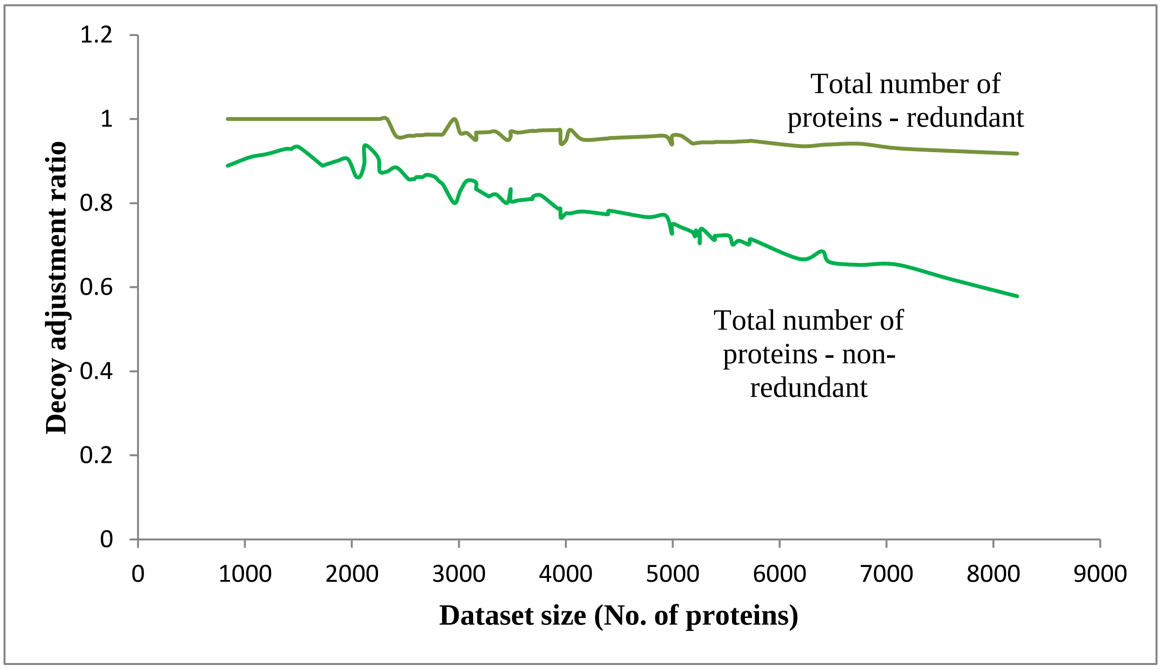


Figure 4.8: Comparison of M-ratio with redundant and non-redundant no. of total proteins. MAYU model was implemented using either a redundant or non-redundant number of proteins as number of proteins in the database. As can be seen, using a redundant number of proteins severely limits the effectiveness of the MAYU model.

is no doubt on what is the right way to implement the MAYU model here, the analysis is still useful to demonstrate the importance of treating protein grouping properly.

CONCLUDING REMARKS

In this chapter, we have discussed important issues to consider when estimating protein false discovery rates in large scale proteomic datasets. The unequal decoy and false positive target spaces when analyzing such datasets break basic assumptions used in the target-decoy strategy for estimating FDR, and not accounting for this can lead to over-estimation of FDR and consequently under identification of proteins from the dataset.

The three methods that have currently been proposed in the literature to adjust FDR estimation for the unequal decoy – false positive spaces have been implemented and tested on two large datasets and multiple smaller datasets. As seen from the results presented, there do not appear to be any significant differences in results from the three methods on the datasets that they were tested on. However, further analysis on more diverse datasets would be useful to confirm this. The R script developed implementing all three FDR estimation methods is expected to be useful to facilitate further comparative analysis on these methods.

While no significant differences in terms of results were observed between the methods, other considerations may weigh upon the choice of FDR estimation method used in proteomics pipelines. For instance, a common strategy to accelerate protein inference for large datasets is to only use PSMs above a certain score threshold, since very low scoring PSMs are unlikely to contribute to true positive protein identifications. However, to implement R-factor correction low scoring protein identifications are required, which may be reduced by this strategy. So if time required for protein inference is an important concern in the analysis, R-factor correction might not be an appropriate choice for FDR estimation.

Finally, the comparisons of the FDR estimation methods in our analyses have been mostly based on the number of proteins identified. But a direct validation of whether the proteins being identified by the adjusted FDR estimation methods are indeed true positive identifications was not performed. While the best way to perform such a validation is not currently obvious, developing a good validation of these methods would be important for widespread adoption of these methods.

REFERENCES

(1) Nesvizhskii, A. I.; Keller, A.; Kolker, E.; Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **2003**, *75* (17), 4646–4658.

(2) Fenyö, D.; Beavis, R. C. A Method for Assessing the Statistical Significance of Mass Spectrometry-Based Protein Identifications Using General Scoring Schemes. *Anal. Chem.* **2003**, *75* (4), 768–774.

(3) Sadygov, R. G.; Yates, J. R. A Hypergeometric Probability Model for Protein Identification and Validation Using Tandem Mass Spectral Data and Protein Sequence Databases. *Anal. Chem.* **2003**, *75* (15), 3792–3798.

(4) Adamski, M.; Blackwell, T.; Menon, R.; Martens, L.; Hermjakob, H.; Taylor, C.; Omenn, G. S.; States, D. J. Data management and preliminary data analysis in the pilot phase of the HUPO Plasma Proteome Project. *Proteomics* **2005**, *5* (13), 3246–3261.

(5) Elias, J. E.; Gygi, S. P. Target-decoy search strategy for mass spectrometry-based proteomics. *Methods Mol. Biol.* **2010**, *604*, 55–71.

(6) Wilhelm, M.; Schlegl, J.; Hahne, H.; Moghaddas Gholami, A.; Lieberenz, M.; Savitski, M. M.; Ziegler, E.; Butzmann, L.; Gessulat, S.; Marx, H.; et al. Mass-spectrometry-based draft of the human proteome. *Nature* **2014**, *509* (7502), 582–587.

(7) Kim, M.-S.; Pinto, S. M.; Getnet, D.; Nirujogi, R. S.; Manda, S. S.; Chaerkady, R.; Madugundu, A. K.; Kelkar, D. S.; Isserlin, R.; Jain, S.; et al. A draft map of the human proteome. *Nature* **2014**, *509* (7502), 575–581.

(8) Serang, O.; Käll, L. The solution to statistical challenges in proteomics is more statistics, not less. *J. Proteome Res.* **2015**.

(9) Shanmugam, A. K.; Yocum, A. K.; Nesvizhskii, A. I. Utility of RNA-seq and GPMDB protein observation frequency for improving the sensitivity of protein identification by tandem MS. *J. Proteome Res.* **2014**, *13* (9), 4113–4119.

(10) Savitski, M. M.; WIlhelm, M.; Hahne, H.; Kuster, B.; Bantscheff, M. A scalable approach for protein false discovery rate estimation in large proteomic data sets. *Mol. Cell. Proteomics* **2015**, mcp.M114.046995.

(11) Reiter, L.; Claassen, M.; Schrimpf, S. P.; Jovanovic, M.; Schmidt, A.; Buhmann, J. M.; Hengartner, M. O.; Aebersold, R. Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry. *Mol. Cell. Proteomics* **2009**, *8* (11), 2405–2417.

(12) Geiger, T.; Wehner, A.; Schaab, C.; Cox, J.; Mann, M. Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Mol. Cell. Proteomics* **2012**, *11* (3), M111.014050.

(13) Craig, R.; Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20* (9), 1466–1467.

(14) Flicek, P.; Amode, M. R.; Barrell, D.; Beal, K.; Billis, K.; Brent, S.; Carvalho-Silva, D.; Clapham, P.; Coates, G.; Fitzgerald, S.; et al. Ensembl 2014. *Nucleic Acids Res.* **2014**, *42* (Database issue), D749–D755.

(15) The UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res.* **2014**, *43* (D1), D204–D212.

(16) Deutsch, E. W.; Mendoza, L.; Shteynberg, D.; Farrah, T.; Lam, H.; Tasman, N.; Sun, Z.; Nilsson, E.; Pratt, B.; Prazen, B.; et al. A guided tour of the Trans-Proteomic Pipeline. *Proteomics* **2010**, *10* (6), 1150–1159.

(17) Nesvizhskii, A. I.; Aebersold, R. Interpretation of shotgun proteomic data: the protein inference problem. *Mol. Cell. Proteomics* **2005**, *4* (10), 1419–1440.

CHAPTER 5
Concluding remarks

Proteomics, with the potential to allow study of the overall protein complement of a cell in all its dynamism and interactivity, holds particular promise for resulting in interesting biological discoveries. Tandem mass spectrometry being the method of choice for high throughput protein identification in proteomics studies, development of methods for improving peptide and protein identification from MS/MS data remains an important and valuable area of research. Integration of orthogonal data from the other established 'omics' technologies and public data repositories can prove to be an effective way to achieve such improvement.

In this dissertation, three different approaches for the integration of orthogonal data (re-scoring based, search space restriction based and a hybrid approach) have been discussed.

The re-scoring based approach, discussed in chapter 2, boosts protein identification probabilities for proteins that have strong supporting evidence from orthogonal data, thereby improving the number of protein identifications from the experiment. However, a limitation of this approach is that it also penalizes the identification probabilities of proteins that do not have supporting evidence in orthogonal data. In some deep coverage datasets, the penalizing of protein identifications may outweigh the improvement due boosted probabilities of other proteins.

Overall, the re-scoring based approach is seen to provide significant improvement in small and medium size datasets.

The search space restriction based approach, discussed in chapter 3, attempts to improve identification sensitivity by restricting the search space based on orthogonal data. We demonstrate this approach using identification frequency data from GPMDB. Restriction was performed at the peptide level to provide an additional level of restriction, taking advantage of the fact that not all peptides are equally identifiable in MS/MS experiments. Importantly, this peptide level search space restriction is designed to be implemented in existing proteomic analysis pipelines with little or no modifications, facilitating the easy adoptability of this approach. Search space restriction was seen to provide significant improvements in peptide identification (up to 11%) in small to medium size datasets. The limitation of this approach, as one might imagine, is the possibility of over-restriction resulting in an incomplete search space, which in turn can cause loss of true positive peptide identifications. However, the hybrid approach was developed to address this limitation.

The hybrid approach or the ‘combined searches workflow’, also described in chapter 3, builds on the search space restriction approach of using a smaller targeted search space for database search, but adds an additional step of merging it (using iProphet) with results from a search against the full proteome. This allows the recovery of peptides missed due to an incomplete search space. The iProphet model used for merging performs a re-scoring that boosts the probabilities of peptides identified in both searches while penalizing the probabilities of those found in only one. This has the indirect effect of performing a more stringent filtering of peptides only found in one of the searches, but recovers most peptides in the full proteome search that have a ‘high enough’ identification probability. This hybrid approach has been shown to provide larger improvements (up to 12.5%) in peptide identification than the search space restriction approach alone, since it is not as limited by missing peptides due to incomplete search spaces. Further, this robustness of hybrid approach to over-restriction also allows for greater flexibility in selecting a threshold for search space restriction.

The search space restriction and hybrid approaches are seen to be particular useful when the MS/MS data being analyzed is ‘noisy’. For instance, the amount of improvement from these approaches was almost 50% higher in pseudo MS/MS data from data independent acquisition (DIA) than in comparable data dependent acquisition (DDA) data (18% in DIA compared to 12% in DDA). So, with the increasing popularity of DIA based proteomics experiments, the search space restriction and hybrid approaches could prove especially useful.

The hybrid approach or combined searches workflow also has relevance beyond just the integration of orthogonal data. It can be useful for any applications that require the merging of database searches that are performed on multiple search spaces. Potential applications in proteogenomics datasets and for the identification of Post-translational modifications (PTMs) have been briefly discussed previously. Recently the use of large mass tolerances in database search, called open search or blind search, has received interest as a method to identify peptides with uncommon PTMs and Amino Acid substitutions. The opening up of the search space in such searches can significantly affect sensitivity of peptide identification, making it an ideal candidate to benefit from the combined searches workflow by merging a blind search and a typical low mass tolerant search.

A different type of integrative analysis discussed in this dissertation is that of combining data from multiple experiments to obtain a very deep coverage of the proteome. The large number of true positive identifications in such datasets shrinks the potential false positive targets space while the decoy space is left un-changed. This leads to a break-down of the assumption that the number of decoy hits at a given score threshold is an acceptable estimate of the number of false positive hits at the same threshold. Therefore, applying the classical target-decoy strategy for FDR estimation in such datasets leads to an over estimation of the FDR and consequently under estimation of the number of true positive protein identifications.

Chapter 4 discusses three different methods to address this issue: R-factor correction, the picked FDR approach, and MAYU (R-factor correction was developed by the author). Comparison of the three methods, on a large scale dataset, shows no significant differences between the three, in terms of adjustment to the number of decoys. However, it must be noted this comparison is limited by the fact that there are not many datasets of large enough scale to perform comprehensive comparisons. Further studies as more large datasets become available might shed more light on differences between them. An R-script implementing all three methods was developed and might be a useful tool for facilitating easy comparisons in future studies.

An important question that remains to be addressed in determining protein identifications is about how best to handle protein grouping. As demonstrated in the dissertation, the way proteins in an indistinguishable group are treated can make a clear impact on the results of the picked FDR approach. Based on results presented in this dissertation, it might need to be intermediate between fully expanding indistinguishable groups and fully collapsing them. However, closer study is needed before this question can be addressed.

Similarly, the right way to treat proteins in distinguishable groups is also not immediately obvious. Distinguishable groups might often involve proteins that have no unique high probability peptides, but do have high probability shared peptides. In our analyses, such proteins are treated as present if they have individual protein identification probabilities of 0.5 or higher. However that is a heuristic threshold and further studies might be able to provide a more sound theoretical grounding for the treatment of such proteins. Addressing these questions can have important consequences on the number of proteins identified in typical protein informatics pipelines. Leveraging orthogonal information, such as from RNAseq, could be one potential avenue to addressing these questions.

Thus, in this dissertation we have developed and demonstrated frameworks for the integration of orthogonal information into proteomics pipelines and discussed methods for accurate FDR estimation in large scale integration of proteomic datasets. We believe this type of integrative analyses provide great value toward improving peptide and protein identifications in current proteomic analysis pipelines. We hope that these frameworks are useful to the proteomics community and help simulate further research in this area.