Published in final edited form as:
J Proteome Res. 2015 December 4; 14(12): 5169–5178. doi:10.1021/acs.jproteome.5b00504.

Effective leveraging of targeted search spaces for improving peptide identification in MS/MS based proteomics

Avinash K. Shanmugam[1] and **Alexey I. Nesvizhski**[1,2,*]

[1]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109

[2]Department of Pathology, University of Michigan, Ann Arbor, MI 48109

Abstract

In shotgun proteomics, peptides are typically identified using database searching which involves scoring acquired tandem mass spectra against peptides derived from standard protein sequence databases such as Uniprot, Refseq, or Ensembl. In this strategy, the sensitivity of peptide identification is known to be affected by the size of the search space. Therefore, creating a targeted sequence database containing only peptides likely to be present in the analyzed sample can be a useful technique for improving the sensitivity of peptide identification. In this study we describe how targeted peptide databases can be created based on the frequency of identification in GPMDB – the largest publicly available repository of peptide and protein identification data. We demonstrate that targeted peptide databases can be easily integrated into existing proteome analysis workflows, and describe a computational strategy for minimizing any loss of peptide identifications arising from potential search space incompleteness in the targeted search spaces. We demonstrate the performance of our workflow using several datasets of varying size and sample complexity.

Keywords

Tandem Mass Spectrometry; GPMDB; RNA-Seq; Integrative analysis; Search space restriction; Targeted databases; combined searches

Corresponding author: Alexey I. Nesvizhskii, Department of Pathology, University of Michigan, 4237 Medical Science I, Ann Arbor, MI, 48109, nesvi@med.umich.edu, Tel: +1 734 764 3516.

SUPPORTING INFORMATION AVAILABLE
Supplementary Table: Dataset details.
Supplementary Figures
Figure 1: Comparison of the sizes of the targeted databases to full protein database in terms of number of peptides.
Figure 2: Comparison of GPMDB based search space restriction with random search space restriction.
Figure 3: Results from RNA-Seq based targeted database, using 30th percentile threshold.
Figure 4: Source of missed peptides.
Figure 5: iProphet probability accuracy
Figure 6: Effect of other models on the iProphet results
Figure 7: Schematic overview of the Peptide supplemented workflow, an alternative approach for minimizing missed peptides.
Figure 8: Results from the peptide supplemented workflow.
Figure 9: Completeness plot (percentage of peptides retained in each targeted database) for all datasets.
Figure 10: Percentage improvement from basic targeted workflow as a function of percentage of peptides retained.
This material is available free of charge via the internet at http://pubs.acs.org

INTRODUCTION

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). In other words, we can reasonably assume, with certain caveats to be discussed later, that a protein frequently identified in GPMDB is more likely to be identified in a new MS/MS experiment than a protein that was never or rarely observed previously. Results from our previous study[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).

GPMDB peptide identification frequencies for the search space restriction were retrieved on August 10th 2014, using a MySQL database dump of GPMDB data provided on their FTP site. Application of the workflows to RNA-Seq based search space restriction was tested using RNA-Seq data generated from a 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 the specific data files used are available in Supplementary table 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 way all peptide identifications were 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.

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 and not a protein like in a typical protein sequence database). Twelve different targeted databases, at different levels of search space reduction, were created for this study by setting 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 Supplementary Figure 1). Decoy sequences for these targeted peptide sequence databases were generated by keeping tryptic sites on the sequence fixed and reversing only the sequences between them, as described above.

Randomly search space restricted databases were also created by sampling from the space of all possible tryptic peptides (up to 1 missed cleavage, peptide length >= 7AA) in the Ensembl human proteome. Searches against these databases were used as a control to validate the efficacy of the GPMDB based targeted method of search space restriction (See Supplementary Figure 2).

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 provided to improve alignments and all other Tophat options were set to their 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 protein sequence database to only retain proteins with corresponding transcript abundances at or above a set threshold as described by Wang et al[2]. While a default threshold of the 30th percentile was suggested by Wang et al[23] for this filtering, for the data used in this study this was seen to be a too stringent threshold (See Supplementary Figure 3), and the 20th percentile was selected as the threshold instead.

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 1. 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).

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 2), 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, leading to a decrease in the overall performance. In these data, the 90th percentile database represents the level of search space restriction at which peptide loss due to search space incompleteness begins to outweigh the gain in the number of 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 database searches) are also included in the figure to provide a clearer picture of these trends. Further analysis of the missed peptides was also performed to confirm that the primary source of peptide loss was indeed search space incompleteness (See Supplementary Figure 4).

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, including combining the results from multiple searches. While iProphet has been used in previous studies to combine the results from multiple different databases search tools, in this study we applied it to combine results from searches against different databases (Figure 3). This computational strategy of performing two separate searches against the full and targeted databases, followed by combining the results using iProphet effectively retains the increased sensitivity advantages of using targeted databases while mitigating the potential negative impact of their incompleteness.

The NSS (number of sibling searches) model of iProphet, boosts a PSM score if the MS/MS spectrum was matched to the same peptide in other searches too, and penalizes it if the spectrum was matched to a different peptide(s), or to the same peptide but with a low score, in other searches. In the context of this work, applying the NSS model has the effect of promoting frequently observed peptides, which are likely to be present in both the full and targeted databases, and penalizing, and hence indirectly setting a more stringent threshold for the identification, of rarer peptides which are less likely to be present in the targeted databases. The NSS model was included when running iProphet in the combined searches workflow since it was seen to provide a slight improvement in discriminatory power.

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 4. 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 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.

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 4 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 Supplementary Figure 5). While turning on other models did not provide much further improvement in our data, in regular analysis pipelines it may be advisable to turn on the other iProphet models and also employ other strategies recommended with iProphet, such as combining multiple search engine results[27], to take full advantage of any potential improvements.

Accuracy of peptide probabilities from iProphet after the NSS model based rescoring can vary depending on the search spaces of the searches being combined. But, since the target-decoy approach, used in this workflow to estimate FDR and filter peptide identifications, only uses the peptide probabilities as a ranking score, the actual peptide probability values are not important as long as the relative order of the peptides is accurate. However, if the accuracy of the peptide probabilities is considered important for downstream analyses, the combining of searches using iProphet can be performed without using the NSS model. Such a method of combining results in more accurate peptide probabilities with only a slight reduction in the amount of improvement achieved (Supplementary Figure 6).

In addition to the two workflows described above, an additional workflow (the peptide supplemented workflow) was also designed and tested. While it wasn't seen to be as effective as the other two workflows, in the interest of potential future improvements to it, a description of the workflow and results from it are provided as supporting information (See Supplementary Figures 7 and 8).

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 5 shows that the overall trends are largely similar across all datasets. In the DIA pseudo MS/MS data (Figure 5A), the improvement in peptide identification is even higher than that seen earlier in the corresponding conventional DDA data (compare with Figure 2 & Figure 4), 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 5B) 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 5C 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 5C 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 5D 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.

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 6A). 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 5B). We also note 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.

Selecting targeted database percentile thresholds

As can be seen from results above, the choice of percentile threshold for creating the targeted database affects the degree of improvement achieved. While in this study multiple thresholds were tested to identify the point of maximum improvement, such an approach would be too time-consuming as part of a routine proteomic analysis pipeline. As an alternative, a heuristic to quickly select an optimal threshold for the targeted database using the ‘percentage of peptides retained’ was developed.

The ‘percentage of peptides retained’ is defined as the percentage of high confidence peptides (identified at 1% FDR) from the full database search that are retained in a particular targeted database. This can be quickly estimated from the results of the full database search by determining the number of peptides in the results which pass the percentile frequency threshold for each targeted database. In our analysis for the combined searches workflow, it was seen that the points of maximum improvement were consistently in the 90–97% range of percentage of peptides retained (See Figure 7).

Based on this, we suggest that selecting a percentile threshold corresponding to 90% of the high confidence peptides retained would serve as a good empirical threshold. In the datasets used in the study, this corresponds to 92^{nd} percentile in the DDA and DIA datasets, 80^{th} percentile in the AP-MS dataset and 70^{th} percentile in the deep coverage dataset (See Supplementary Figure 9), which would provide close to the maximum improvement in each dataset. Further, since the combined searches workflow already requires performing a separate search with the full database, determining ‘percentage of peptides retained’ would not add any significant processing time to the analysis pipeline. A similar analysis with results from the basic targeted pipeline is described in the supplementary information (See Supplementary Figure 10)

CONCLUSIONS

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.

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 e.g. 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 performed periodically as the GPMDB database continues growing in size.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to acknowledge Chih-Chiang Tsou for help running the DIA-Umpire software and Andy Kong for help investigating X! Tandem’s E-value modeling. We would also like to thank Dmitriy Avtonomov and Dattatreya

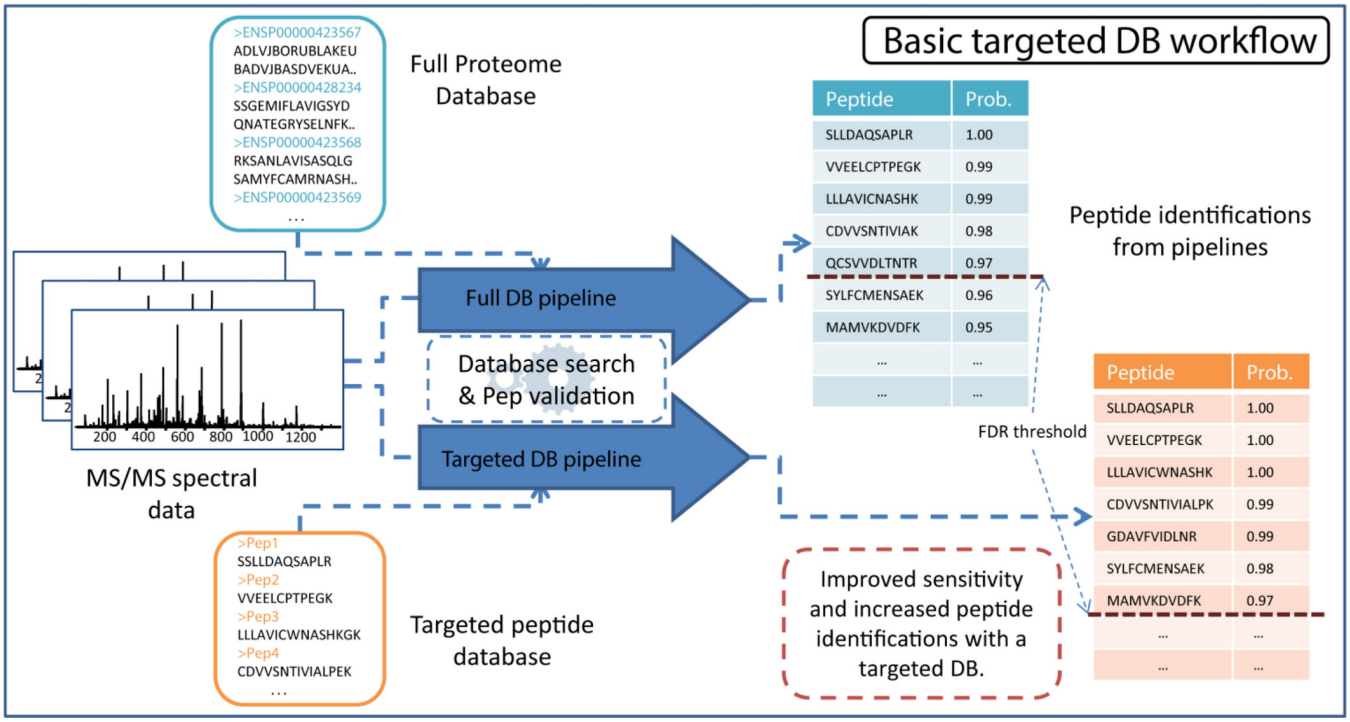
Mellachuruvu for helpful discussions. This work was supported in part by U.S. National Institutes of Health grant no. 5R01GM94231.

REFERENCES

1. Nesvizhskii AI. 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. [PubMed: 20816881]
2. Wang X, Slebos RJC, Wang D, Halvey PJ, Tabb DL, Liebler DC, Zhang B. Protein identification using customized protein sequence databases derived from RNA-Seq data. J. Proteome Res. 2012; 11(2):1009–1017. [PubMed: 22103967]
3. Craig R, Cortens JP, Beavis RC. Open source system for analyzing, validating, and storing protein identification data. J. Proteome Res. 3(6):1234–1242. [PubMed: 15595733]
4. Desiere F, Deutsch EW, King NL, Nesvizhskii AI, Mallick P, Eng J, Chen S, Eddes J, Loevenich SN, Aebersold R. The PeptideAtlas project. Nucleic Acids Res. 2006; 34(Database issue):D655–D658. [PubMed: 16381952]
5. Shanmugam AK, Yocum AK, Nesvizhskii AI. 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. [PubMed: 25026199]
6. Craig R, Cortens JP, Beavis RC. The use of proteotypic peptide libraries for protein identification. Rapid Commun. Mass Spectrom. 2005; 19(13):1844–1850. [PubMed: 15945033]
7. Mallick P, Schirle M, Chen SS, Flory MR, 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. [PubMed: 17195840]
8. Tsou C-C, Avtonomov D, Larsen B, Tucholska M, Choi H, Gingras A-C, Nesvizhskii AI. DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. Nat. Methods. 2015; 12(3):258–264. [PubMed: 25599550]
9. Mellacheruvu D, Wright Z, Couzens AL, Lambert J-P, St-Denis NA, Li T, Miteva YV, Hauri S, Sardiu ME, Low TY, et al. The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. Nat. Methods. 2013; 10(8):730–736. [PubMed: 23921808]
10. Scheltema RA, 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. [PubMed: 25360005]
11. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. 2011; 25(18):1915–1927. [PubMed: 21890647]
12. Flicek P, Amode MR, 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. [PubMed: 24316576]
13. Kim S, Pevzner PA. MS-GF+ makes progress towards a universal database search tool for proteomics. Nat. Commun. 2014; 5:5277. [PubMed: 25358478]
14. Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. Bioinformatics. 2004; 20(9):1466–1467. [PubMed: 14976030]
15. Deutsch EW, 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. [PubMed: 20101611]
16. Keller A, Nesvizhskii AI, 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. [PubMed: 12403597]
17. Choi H, Nesvizhskii AI. Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics. J. Proteome Res. 2008; 7(1):254–265. [PubMed: 18159924]
18. Shteynberg D, Deutsch EW, Lam H, Eng JK, Sun Z, Tasman N, Mendoza L, Moritz RL, Aebersold R, Nesvizhskii AI. 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. [PubMed: 21876204]

19. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009; 25(9):1105–1111. [PubMed: 19289445]
20. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat. Methods. 2012; 9(4): 357–359. [PubMed: 22388286]
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. [PubMed: 24058055]
24. Tharakan R, Edwards N, Graham DRM. Data maximization by multipass analysis of protein mass spectra. Proteomics. 2010; 10(6):1160–1171. [PubMed: 20082346]
25. Nesvizhskii AI, Aebersold R. Interpretation of shotgun proteomic data: the protein inference problem. Mol. Cell. Proteomics. 2005; 4(10):1419–1440. [PubMed: 16009968]
26. Ning K, Fermin D, Nesvizhskii AI. 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. [PubMed: 22329341]
27. Shteynberg D, Nesvizhskii AI, Moritz RL, Deutsch EW. Combining results of multiple search engines in proteomics. Mol. Cell. Proteomics. 2013; 12(9):2383–2393. [PubMed: 23720762]
28. Eng JK, Jahan TA, Hoopmann MR. Comet: an open-source MS/MS sequence database search tool. Proteomics. 2013; 13(1):22–24. [PubMed: 23148064]
29. Howbert JJ, Noble WS. Computing exact p-values for a cross-correlation shotgun proteomics score function. Mol. Cell. Proteomics. 2014; 13(9):2467–2479. [PubMed: 24895379]
30. Li, S.; Arnold, RJ.; Tang, H.; Radivojac, P. Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics - BCB'13. ACM Press; New York, New York, USA: 2007. Improving phosphopeptide identification in shotgun proteomics by supervised filtering of peptide-spectrum matches; p. 316-323.
31. Nesvizhskii AI. Proteogenomics: concepts, applications and computational strategies. Nat. Methods. 2014; 11(11):1114–1125. [PubMed: 25357241]
32. Blakeley P, Overton IM, Hubbard SJ. Addressing statistical biases in nucleotide-derived protein databases for proteogenomic search strategies. J. Proteome Res. 2012; 11(11):5221–5234. [PubMed: 23025403]
33. Krug K, Carpy A, Behrends G, Matic K, Soares NC, 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. [PubMed: 23908556]
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. [PubMed: 21298786]



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 1. Basic targeted database search 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.

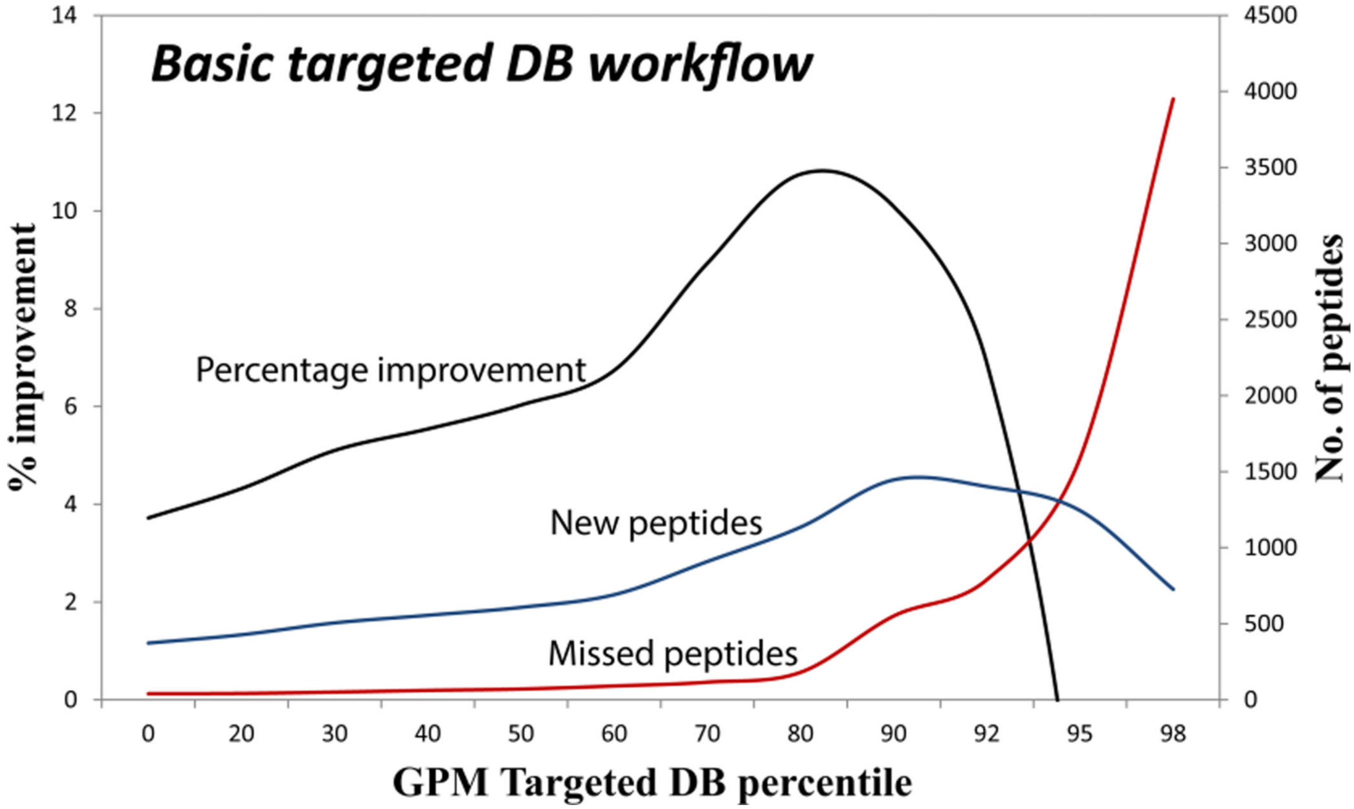
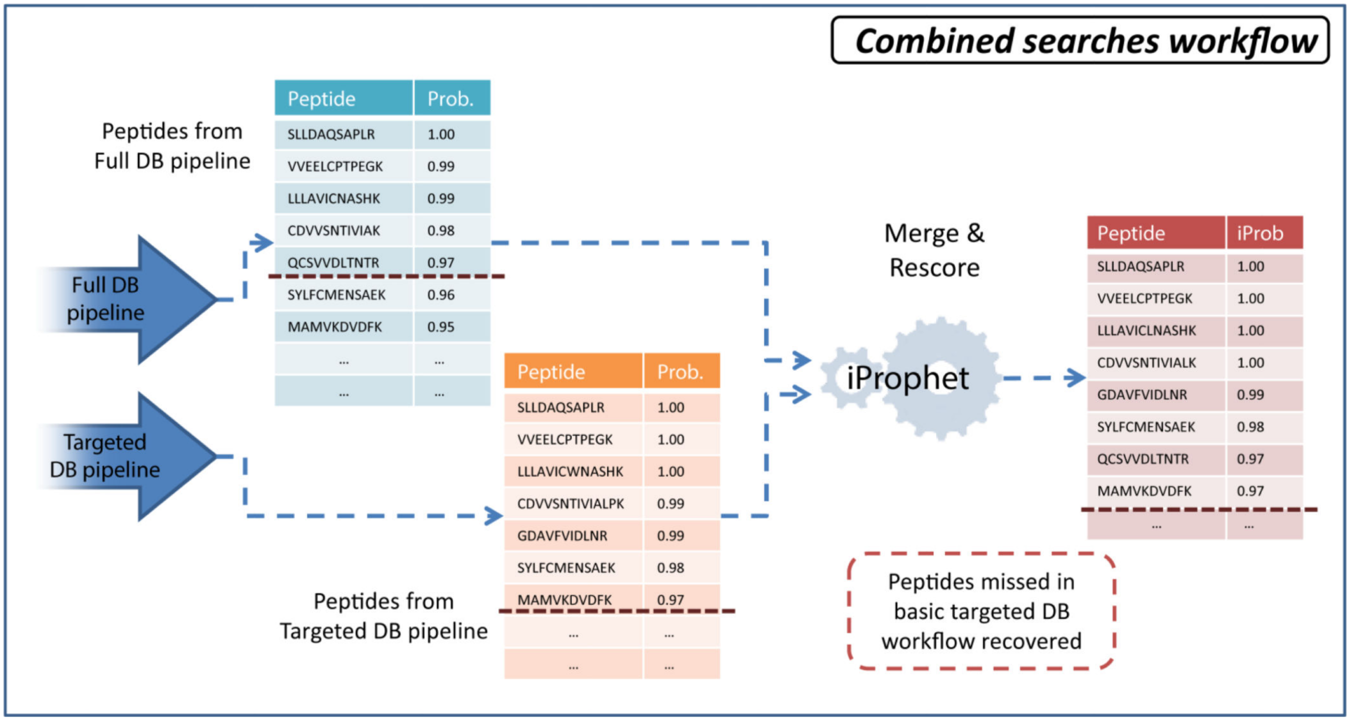


Figure 2.

Percentage improvement from the basic targeted database workflow is plotted for the varying levels of search space restriction, using data from the K562 lysate. The number of peptides missed compared to the full database search and the number of new peptides identified is also plotted for each targeted database percentile. Maximum percentage improvement is obtained at the balance between getting the maximum number of new peptides while still keeping the number of missed peptides low.


Combined searches workflow
Peptides from Full DB pipeline
Full DB pipeline
Targeted DB pipeline
Merge & Rescore
iProphet
Peptides from Targeted DB pipeline
Peptides missed in basic targeted DB workflow recovered

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. 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.

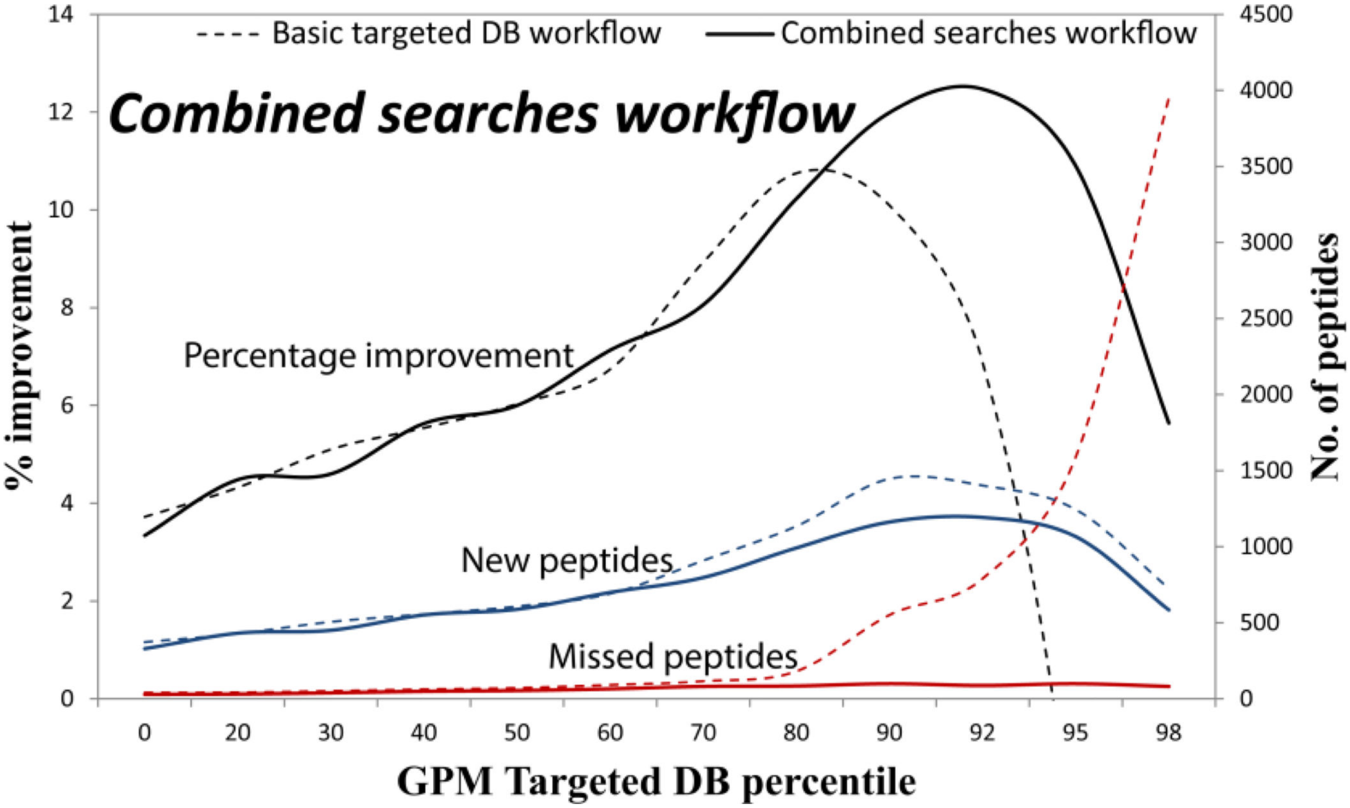


Figure 4.

Percentage improvement from applying the combined searches workflow to K562 lysate MS/MS data, at varying levels of search space restriction is plotted. The combined searches workflow outperforms the basic targeted database workflow in terms of the maximum improvement. Improved performance is achieved by minimizing the number of missed peptides.

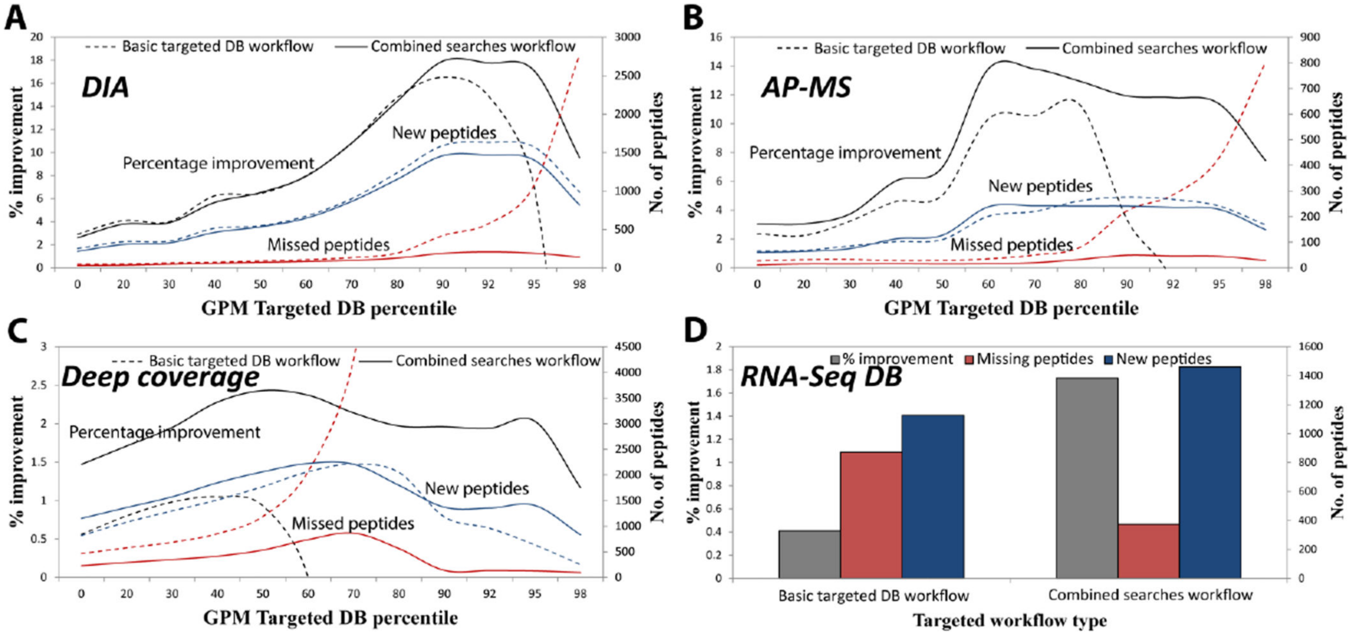


Figure 5.

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

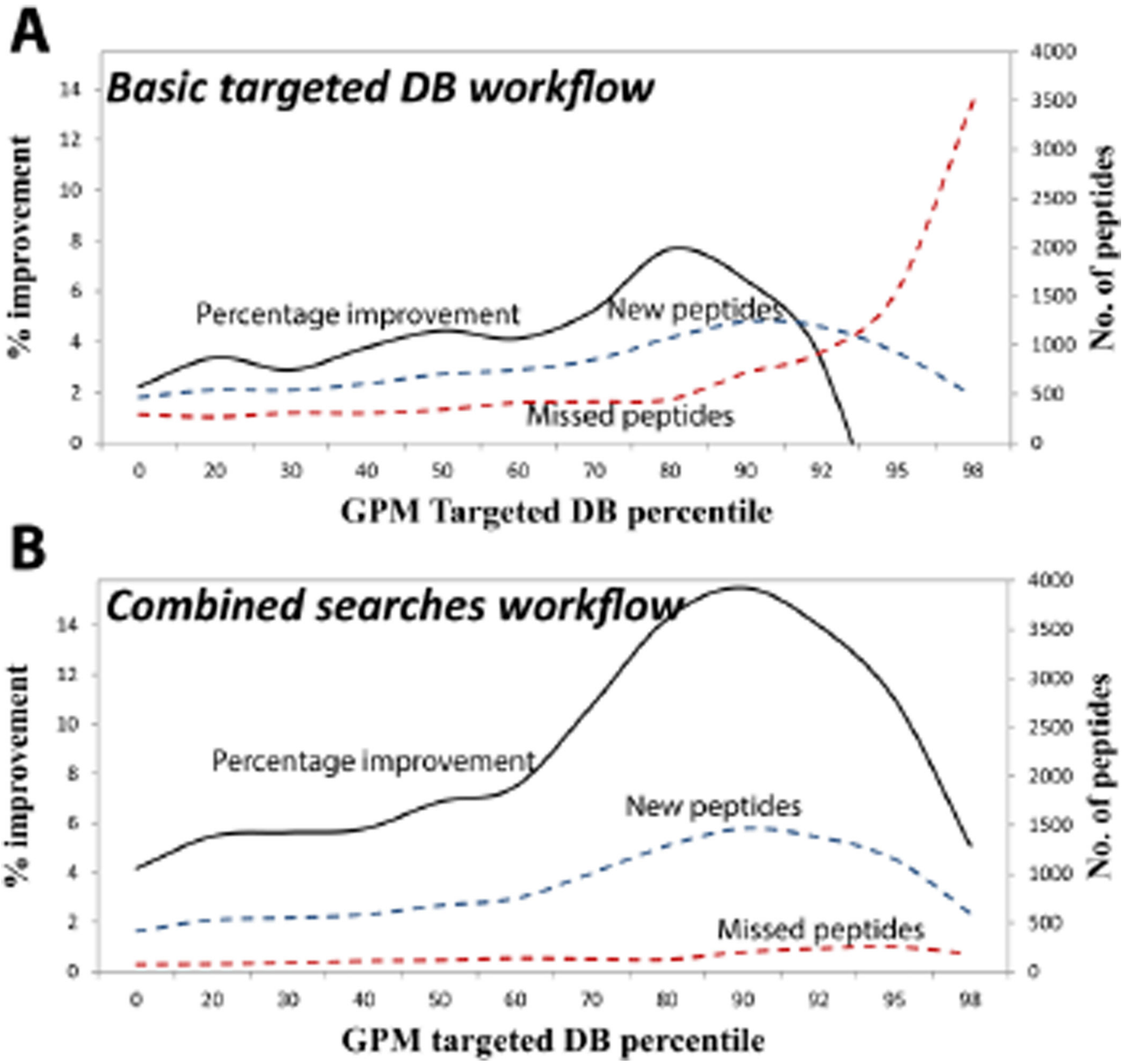


Figure 6.

Results from applying the targeted database search workflows in combination with the X! Tandem database search engine on data from K562 cell lysate using (A) the basic targeted database workflow; (B) Combined searches workflow.

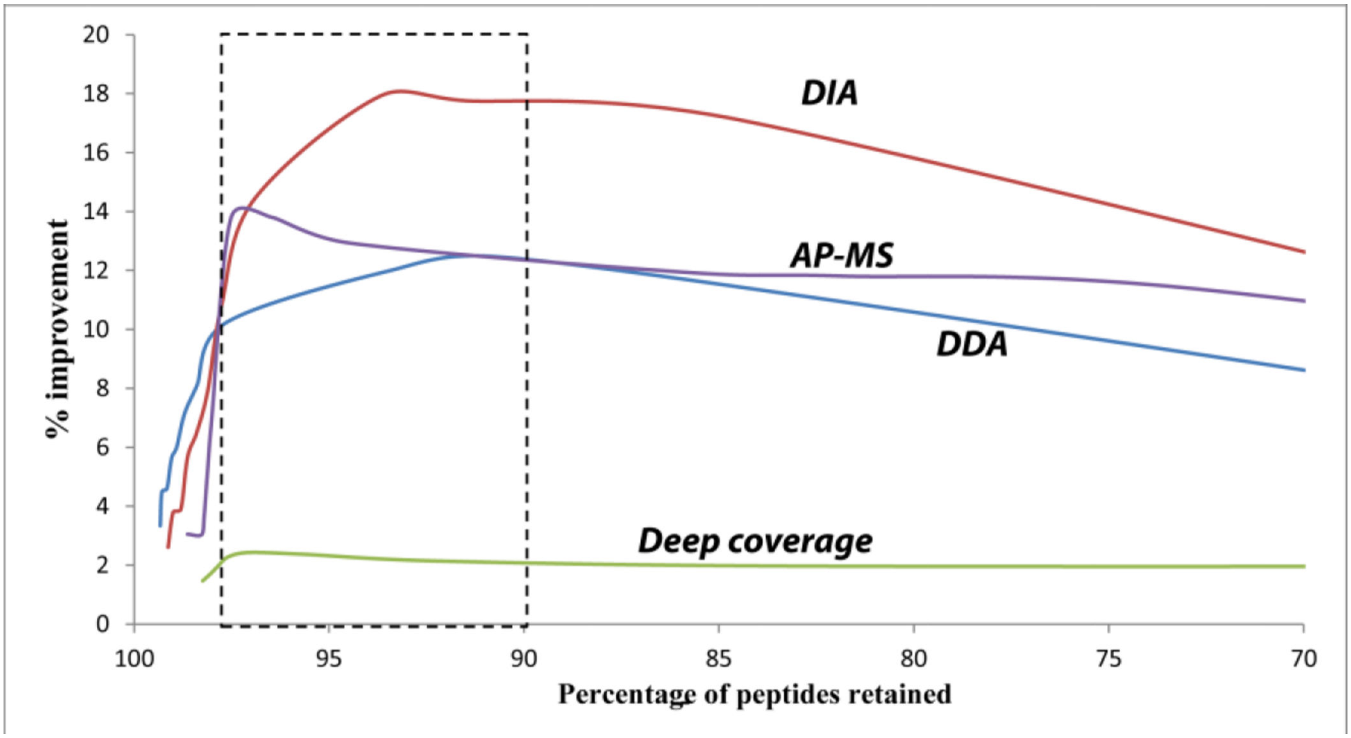


Figure 7.

Percentage improvement for the various datasets from the combined searches workflow plotted against the percentage of high confidence (1% FDR) peptides, from the full proteome database search, retained in the targeted databases. The point of peak improvement is seen to occur in the 90%–97% range for all the datasets.