Journal of proteome research

Article

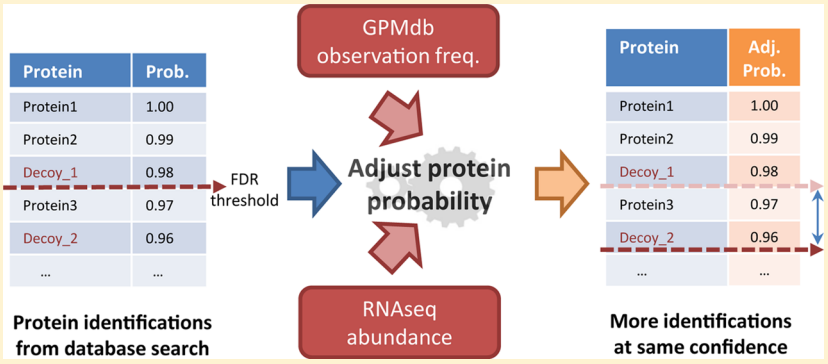
pubs.acs.org/jpr

Utility of RNA-seq and GPMDB Protein Observation Frequency for Improving the Sensitivity of Protein Identification by Tandem MS

Avinash K. Shanmugam,[†] Anastasia K. Yocum,[‡] and Alexey I. Nesvizhskii*[,†,‡]

[†]Department of Computational Medicine and Bioinformatics and [‡]Department of Pathology, University of Michigan, Ann Arbor, Michigan 48109, United States

Supporting Information



ABSTRACT: Tandem mass spectrometry (MS/MS) followed by database search is the method of choice for protein identification in proteomic studies. Database searching methods employ spectral matching algorithms and statistical models to identify and quantify proteins in a sample. In general, these methods do not utilize any information other than spectral data for protein identification. However, considering the wealth of external data available for many biological systems, analysis methods can incorporate such information to improve the sensitivity of protein identification. In this study, we present a method to utilize Global Proteome Machine Database identification frequencies and RNA-seq transcript abundances to adjust the confidence scores of protein identifications. The method described is particularly useful for samples with low-to-moderate proteome coverage (i.e., <2000−3000 proteins), where we observe up to an 8% improvement in the number of proteins identified at a 1% false discovery rate.

KEYWORDS: *Tandem mass spectrometry, RNA-seq, GPMDB, integrative analysis, probability adjustment, FDR, confidence threshold*

INTRODUCTION

The first step in a typical computational pipeline for protein identification through database searching is comparing MS/MS spectra to peptide sequences to identify the best matching peptide for each spectrum, referred to as peptide-to-spectrum matches (PSMs). These PSMs are then processed by protein inference algorithms[1] to produce a minimal list of proteins that would need to be present in the sample to explain the identified PSMs. The proteins in this list are also assigned a confidence score or protein probability calculated on the basis of several factors, including the number of PSMs supporting the protein and the match score of the PSMs. To determine the confident identifications from the results of protein inference, false discovery rates[2] (FDR) are estimated. Only proteins identified at or above a certain FDR threshold (typically 1 or 5%) are chosen as high-confidence identifications for further analyses.

Although sophisticated algorithms for spectral matching and analysis have been developed, protein identification 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.[3] 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,[4] pI,[5] or mass accuracy,[6] some of which are now a routine part of many proteomic analysis pipelines.[4] Studies have also investigated using matching MS^2 and MS^3 information,[7] match scores from multiple search engines,[8,9] and various other information sources to rescore or adjust protein identification probability.

Another category of methods have investigated utilizing external information (information from outside the MS/MS experiment) such as microarray data,[10] protein−protein interaction networks,[11] or gene functional networks[12] to improve protein identifications. A recent study by Wang et al.[13] described an approach to utilize RNA-seq abundance information to limit the size of protein sequence databases and thereby improve protein identification sensitivity.

Received: May 16, 2014
Published: July 15, 2014

   dx.doi.org/10.1021/pr500496p | *J. Proteome Res.* 2014, 13, 4113−4119

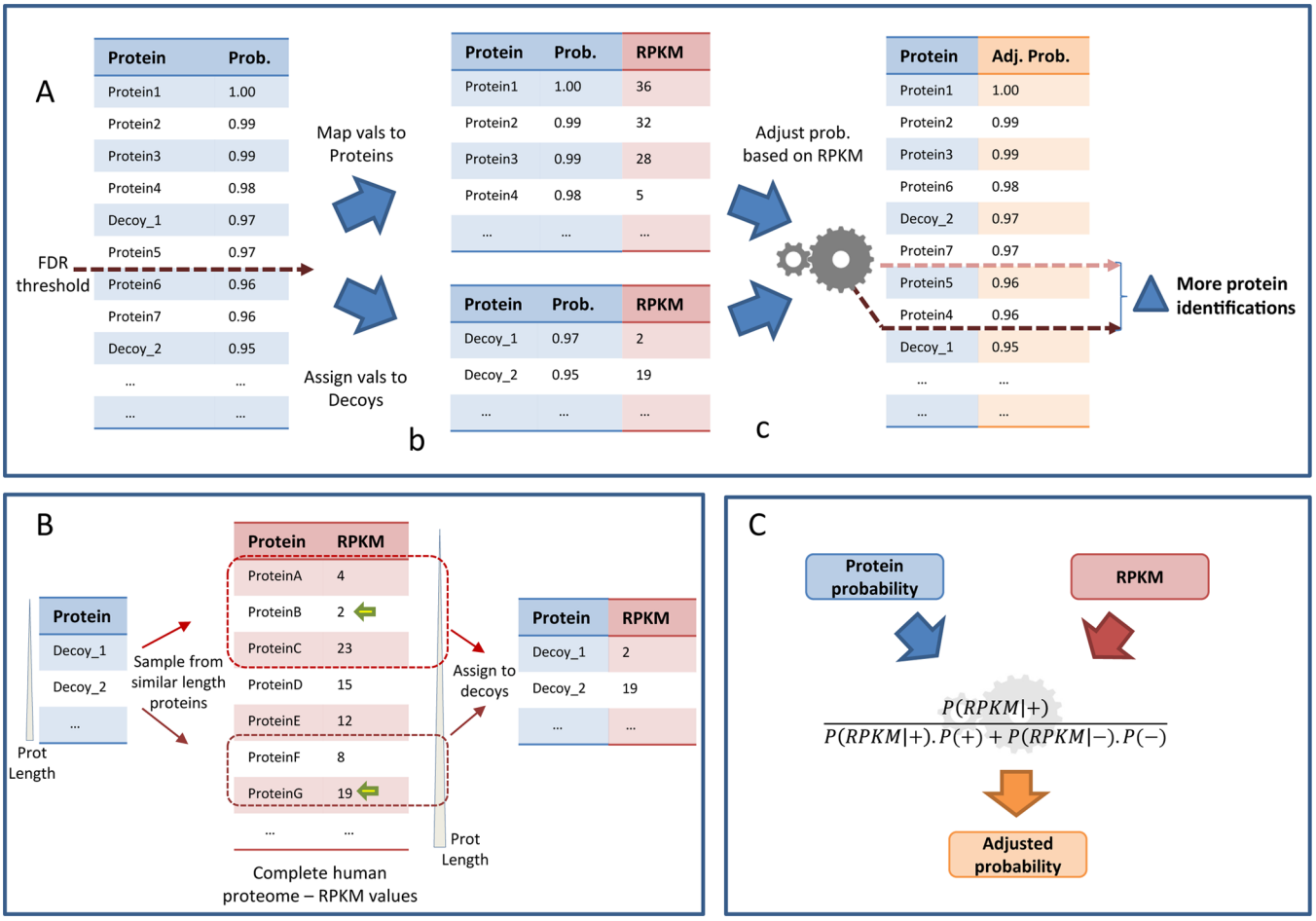


Figure 1. Overview of the method. (A) External information is added to protein identifications from the analysis pipeline. Protein probabilities are adjusted on the basis of transcript abundance (RPKM) or GPMDB identification frequency (GPMfreq). (B) Decoys sequences used to estimate FDR thresholds do not have native RPKM (or GPMfreq) values; they are assigned values by sampling from a set of all forward sequences with similar length (see Methods). (C) Protein identification probabilities are adjusted using Bayes' theorem.

In this article, we describe an alternate method for incorporating external information such as RNA-seq abundance and GPMDB identification frequency into proteomic analysis pipelines, through rescoring or adjustment of protein identification probabilities. RNA-seq uses short read sequencing technologies to sequence the RNA content (transcriptome profile) of a sample.[14] On the basis of the central dogma of molecular biology, it is a reasonable assumption that proteins corresponding to high-abundance transcripts are more likely to be found in a sample. The Global Proteome Machine Database (GPMDB)[15] is a repository storing the results of proteomics experiments. With the large volume of data aggregated in GPMDB, the frequency of identification of a protein in GPMDB can be used as a surrogate measure of a protein's propensity to be observed in a MS/MS experiment. In other words, we can reasonably assume that proteins with a high GPMDB identification frequency (GPMfreq) are more likely to be identified in an MS/MS experiment. In this study, we evaluate the utility of incorporating both of the above types of information into proteomics analysis pipelines.

METHODS

Data Sets

Data from VCaP,[16] a human prostate cancer cell line, and HEK293,[17] a cell line derived from human embryonic kidney cells, were used in this study. The MS/MS and RNA-seq data for the VCaP cell line were generated in parallel at the same lab. This RNA-seq data was also used as the control sample in a paper by Sam et al.[18] and is available for download from the NCBI short read archive, SRA (SRA accession nos. SRR090590 and SRR090591). For the HEK293 cell line, MS/MS data was obtained from control samples in a publication by Fonslow et al.,[19] whereas the RNA-seq data was downloaded from data generated by Sultan et al.[20] (SRA accession nos. SRR023583 and 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, MI). 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. Fifty milligram aliquots of total cellular 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.[21] Extracted peptides were

dx.doi.org/10.1021/pr500496p | J. Proteome Res. 2014, 13, 4113−4119

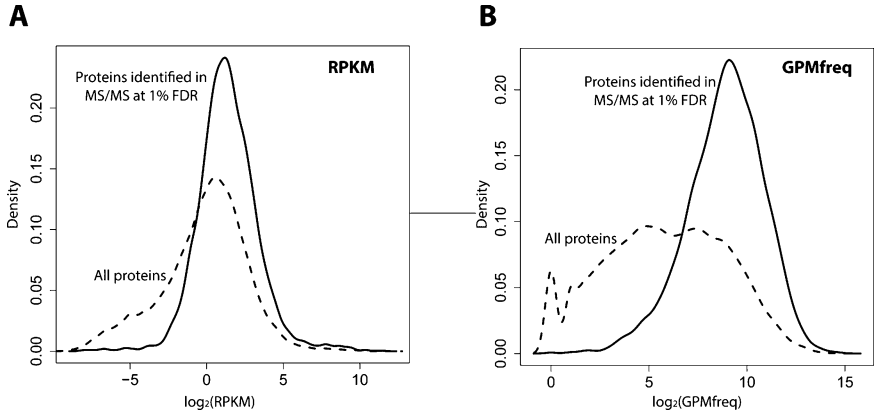


Figure 2. Density distributions of RPKM (A) and GPMfreq (B) values (log-scaled) for proteins identified at 1% FDR in the VCaP cell line and all proteins in Ensembl are plotted. The difference between the two distributions allows us to sample the "all proteins" distribution to assign values to decoys in an unbiased manner while still maintaining discrimination between true positive and decoy identifications.

reconstituted with mobile phase A prior to online reverse-phase nanoLC–MS/MS (LTQ-Velos with Proxeon nanoHPLC, ThermoFinnigan). Peptides were eluted online 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.[22]

Preparation of Data Sets

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 under 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, whereas the HEK293 data consisted of 60 mzXML files. The number of protein identifications in these subsets ranged between about 500 and 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)[23] search engine with a K-score plugin[24,25] 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 fixed and variable modifications, respectively. VCaP data was searched using a precursor mass error of −1 to +4 Da, whereas 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.[4] VCaP data was processed with +1 charge state ions set to be ignored and using a semisupervised model[26] 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[27] to select a representative protein for each protein group, according to heuristic filters built into the tool.

RNA-Seq Processing Pipeline

RNA-seq data was aligned to the Ensembl v.66 human genome (hg19 build 37) using the Tophat aligner (Tophat v.1.3.2).[28] 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)[29] (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[30] packages Rsamtools (v.1.6.3)[31] and GenomicFeatures (v.1.6.9).[32]

RESULTS

Overview of the Approach

Our approach to incorporating RNA-seq or GPMDB frequency information (Figure 1) is built upon a statistical adjustment[7,8] of the protein probability. The probability adjustment increases the identification confidence scores of proteins that have significant supporting evidence from external data (high transcript abundance in RNA-seq or high frequency of identification in GPMDB), 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 "gray zone" of identification confidence, can be promoted above the threshold when ranked by the adjusted probability. Therefore, we are able to obtain more protein identifications at the same FDR.

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. The number of identifications

dx.doi.org/10.1021/pr500496p | J. Proteome Res. 2014, 13, 4113–4119

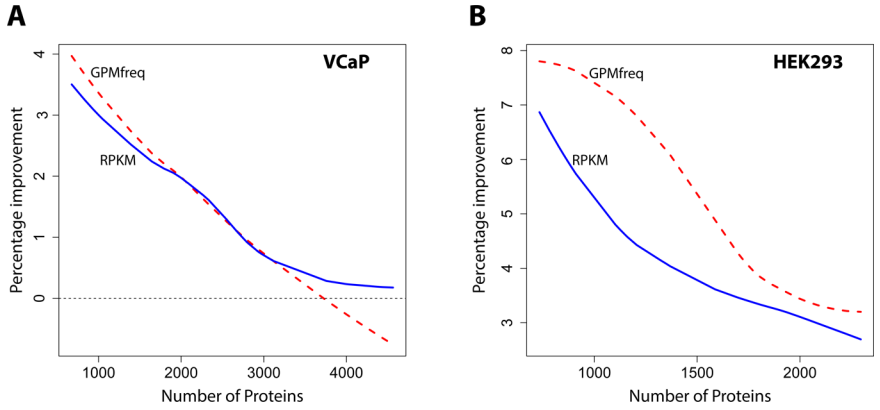


Figure 3. Percentage improvement due to the probability adjustments (RPKM and GPMfreq) for VCaP (A) and HEK293 (B) cell lines plotted at various depths of proteome coverage (no. of proteins). The adjustment is more effective for low- and medium-coverage data sets.

matched to decoy sequences is used to estimate the rate of random matches in the identifications mapped to forward sequences). Because 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 estimate FDR after probability adjustment in an unbiased manner, 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). On the basis of this observation, sampling from the "all proteins" distribution allows for the unbiased assignment of RPKM/GPMfreq values to decoy sequences while maintaining discrimination between decoy and forward identifications. Weak correlations among 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 Supporting Information methods for a more detailed description of the sampling process). In further analysis, it was observed that the improvement from probability could be slightly increased if the sampling for decoy values was weighted to prefer values from proteins not identified in the MS/MS experiment, instead of using a completely random sampling (Supporting Information Figure 5). However, only results from the more statistically rigorous approach of completely random sampling are reported here.

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 Supporting Information methods 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 identifications than the protein probability value (Supporting Information Figure 1), 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 (eq 1).

$$P(+|V) = \frac{P(V|+).P(+)}{P(V|+).P(+) + P(V|-).P(-)} \quad (1)$$

where V can be either an RPKM or GPMfreq value.

The prior probability terms $P(+)$ and $P(-)$ were substituted with pProt and 1 − pProt, respectively. To estimate the conditional probabilities of decoy or forward identifications having value V, $P(V|-)$ and $P(V|+)$, the density distribution of log-scaled RPKM/GPMfreq values was placed into bins of equal width (Supporting Information Figure 2). 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

Because the RPKM/GPMfreq values are assigned through random sampling, the assignment and probability adjustment (Figure 1A) are repeated multiple times to nullify any sampling artifacts and to obtain stable mean adjusted probability values. In our study, the mean values were typically seen to stabilize after about 200 iterations (Supporting Information Figure 3), but the process was repeated to 500 iterations for the results reported here. 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 3A,B. Loess smoothing was performed on the values to show trends clearly.

The probability adjustment results in improvements of almost 8% in the HEK293 cell line and up to 4% in VCaP (Figure 3). Notably, the amount of improvement observed is similar for both the RPKM and GPMfreq adjustments. Furthermore, it appears that using RNA-seq data generated in parallel to the MS/MS data (VCaP) or RNA-seq 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 gray zone of confidence of identification. To test this hypothesis, the entire analysis

dx.doi.org/10.1021/pr500496p | J. Proteome Res. 2014, 13, 4113−4119

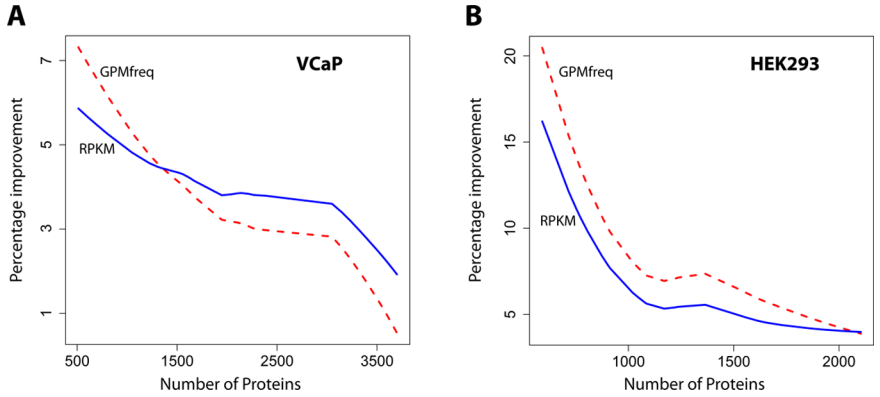


Figure 4. Percentage improvement at various depths of proteome coverage when probability adjustment is performed on a maximum hyperscore-based protein identification probability. As expected, the improvement is significantly greater when the suboptimal maximum hyperscore is used instead of maximum peptide probability (Figure 3).

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. Because maximum hyperscore is a suboptimal score compared to maximum peptide probability, the resulting protein identifications should have more proteins in the gray zone and therefore the probability adjustment on these identifications should provide increased improvement. As expected, Figure 4A,B shows that the percentage improvement is much greater (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 gray zone of the 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 data set) increases can be seen (Figure 3). With deeper coverage of the proteome, low abundance and rare proteins are increasingly identified. As per our assumptions, such proteins would have low RNA-seq 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 gray zone. On the basis of this, we believe that the observation of decreased improvement in deeper coverage data sets reflects the fact that in these data sets there are fewer proteins that would benefit from the probability adjustment.

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 data set (largest data set without any sub sampling) of that cell line. A promoted protein being found in the complete sample would suggest that the protein is indeed a true identification. It is possible that there was not sufficient MS/MS evidence in the smaller sampled data set 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.

Table 1

cell line	rescoring	no. of promoted proteins in sampled data set	no. identified in complete data set
VCaP	RPKM	55	43
	GPMfreq	52	41
HEK293	RPKM	82	55
	GPMfreq	88	63

The remaining 20−30% of promoted proteins, which were not observed in the complete data set, were seen to have high confidence scores in the same range as that of the validated proteins. In other words, unobserved proteins were not outliers (Supporting Information Figure 6). We believe that these unobserved proteins are also true positive protein identifications. It is possible that these proteins were not observed in the complete data set because, even with the increased amount of MS/MS evidence collected in the complete data set, 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.

DISCUSSION

The probability adjustment method described here allows us to utilize external data, such as RNA-seq abundance, or GPMDB identification frequency, to improve the sensitivity of protein identification through database searching. Although some studies generate RNA-seq data in parallel to proteomics data, large amounts of RNA-seq 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

dx.doi.org/10.1021/pr500496p | J. Proteome Res. 2014, 13, 4113−4119

Archive (SRA).[33] As we can see from Figure 3, whether RNA-seq data is generated in 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 RNA-seq 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. However, for commonly studied organisms of interest such as human or mouse, with numerous experiments in GPMDB, it can be a useful source of external information. Computing a combined adjusted probability from both RNA-seq and GPMDB information does not result in a marked improvement in protein identification over the individual probability adjustments (Supporting Information Figure 4), suggesting that RNA-seq and GPMDB capture similar types of information about a sample for the purposes of probability adjustment.

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 gray zone that would benefit from the probability adjustment and there are more rare and low-abundance proteins that could be penalized by it. This is an inherent upper limit to the amount of improvement that is possible by this method and must be taken into consideration when applying this probability adjustment to large samples. However, this method 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 obtained from older instruments or experiments where the amount of instrument time available was low. Compared to the customized database approach described by Wang et al.,[13] the probability adjustment method was seen to provide better improvement for low-to-medium-coverage samples, whereas the customized database approach performed better for deep-coverage samples (Supporting Information Figure 9), suggesting nonoverlapping scenarios of usage for the two methods.

Although the probability adjustment approach has been demonstrated using RNA-seq and GPMDB data, it does not include any assumptions that would limit it to only these two kinds of data. Therefore, this approach can be utilized to incorporate any external source of data (with a significant association with protein presence or abundance) into proteomic analysis pipelines to improve the sensitivity of protein identification.

■ ASSOCIATED CONTENT

Supporting Information

Additional methods: R factor correction, sampling for assigning decoy values, computing pProt values, computing combined adjusted probability, computing mean adjusted probability for decoys, and replicating the customized database approach. Figure 1: ROC curves from Protein Prophet probability and maximum peptide probability based ranking of protein identifications. Figure 2: Density distributions of RPKM and GPMfreq values for high-confidence forward identifications and decoys. Figure 3: Stabilization of adjusted probability values across iterations. Figure 4: Percentage improvement using combined (RPKM and GPMfreq) probability adjustment. Figure 5: Percentage improvement using weighted sampling for decoy value assignment. Figure 6: Probabilities histogram for proteins promoted above 1% FDR by probability adjustment. Figure 7: Density distribution of RPKM values for all protein coding transcripts in VCaP and HEK293. Figure 8: Percentage improvement from the customized database approach. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: nesvi@med.umich.edu. Tel: +1 734 764 3516.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We would like to acknowledge Dr. Hyungwon Choi and Dr. Damian Fermin for helpful discussions. We would also like to thank Mr. Dattatreya Mellacheruvu and Dr. Scott Walmsley for suggestions about the study and for valuable guidance in the preparation of the manuscript. This work was supported by U.S. National Institutes of Health grant no. 5R01GM94231

■ REFERENCES

(1) Nesvizhskii, A. I.; Aebersold, R. Interpretation of shotgun proteomic data: the protein inference problem. *Mol. Cell. Proteomics* **2005**, *4* , 1419−40.

(2) Choi, H.; Nesvizhskii, A. I. False discovery rates and related statistical concepts in mass spectrometry-based proteomics. *J. Proteome Res.* **2008**, *7*, 47−50.

(3) 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*, 2092−123.

(4) Deutsch, E. W.; Mendoza, L.; Shteynberg, D.; et al. A guided tour of the Trans-Proteomic Pipeline. *Proteomics* **2010**, *10*, 1150−9.

(5) Malmström, J.; Lee, H.; Nesvizhskii, A. I.; et al. Optimized peptide separation and identification for mass spectrometry based proteomics via free-flow electrophoresis. *J. Proteome Res.* **2006**, *5*, 2241−9.

(6) 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*, 1183−93.

(7) 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*, 71−87.

(8) Shteynberg, D.; Deutsch, E. W.; Lam, H.; et al. iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Mol. Cell. Proteomics* **2011**, *10*, M111.007690.

(9) 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*, 1494−502.

(10) Ramakrishnan, S. R.; Vogel, C.; Prince, J. T.; et al. Integrating shotgun proteomics and mRNA expression data to improve protein identification. *Bioinformatics.* **2009**, *25*, 1397−1403.

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

(12) 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*, 2955−2961.
(13) Wang, X.; Slebos, R. J. C.; Wang, D. Protein identification using customized protein sequence databases derived from RNA-Seq data. *J. Proteome Res.* **2012**, *11*, 1009−1017.
(14) Wang, Z.; Gerstein, M.; Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **2009**, *10*, 57−63.
(15) Craig, R.; Cortens, J. P.; Beavis, R. C. Open source system for analyzing, validating, and storing protein identification data. *J. Proteome Res.* **2004**, *3*, 1234−1242.
(16) Korenchuk, S.; Lehr, J. E.; MClean, L. VCaP, a cell-based model system of human prostate cancer. *In Vivo* **2001**, *15*, 163−168.
(17) 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*, 59−74.
(18) Sam, L. T.; Lipson, D.; Raz, T.; et al. A comparison of single molecule and amplification based sequencing of cancer transcriptomes. *PLoS One.* **2011**, *6*, e17305.
(19) Fonslow, B. R.; Stein, B. D.; Webb, K. J.; et al. Digestion and depletion of abundant proteins improves proteomic coverage. *Nat. Methods* **2013**, *10*, 54−56.
(20) Sultan, M.; Schulz, M. H.; Richard, H.; et al. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* **2008**, *321*, 956−960.
(21) 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*, 3506−3514.
(22) Pedrioli, P. G. A. Trans-proteomic pipeline: a pipeline for proteomic analysis. *Methods Mol. Biol.* **2010**, *604*, 213−238.
(23) Craig, R.; Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20*, 1466−1467.
(24) 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.
(25) 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*, 2830−2832.
(26) Choi, H.; Nesvizhskii, A. I. Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics. *J. Proteome Res.* **2008**, *7*, 254−265.
(27) 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*, 1340−1345.
(28) Trapnell, C.; Pachter, L.; Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **2009**, *25*, 1105−1111.
(29) Mortazavi, A.; Williams, B. A.; McCue, K.; Schaeffer, L.; Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **2008**, *5*, 621−628.
(30) Gentleman, R. C.; Carey, V. J.; Bates, D. M.; et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **2004**, *5*, R80.
(31) Morgan, M.; Pagès, H.; Obenchain, V. Rsamtools: binary alignment (BAM), variant call (BCF), or tabix file import; http://bioconductor.org/packages/release/bioc/html/Rsamtools.html.
(32) Lawrence, M.; Huber, W.; Pagès, H.; Aboyoun, P.; Carlson, M.; Gentleman, R.; Morgan, M.; Carey, V. Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **2013**, *9*, e1003118.
(33) Leinonen, R.; Sugawara, H.; Shumway, M. The sequence read archive. *Nucleic Acids Res.* **2011**, *39*, D19−D21.