Published in final edited form as:

J Proteome Res. 2012 February 3; 11(2): 1009–1017. doi:10.1021/pr200766z.

Protein identification using customized protein sequence databases derived from RNA-Seq data

Xiaojing Wang¹, Robbert J. C. Slebos^{2,3,4}, Dong Wang¹, Patrick J. Halvey^{3,5}, David L. Tabb^{1,2,5}, Daniel C. Liebler^{1,2,3,5}, and Bing Zhang^{1,2,4,*}

¹Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, TN 37232

²Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, TN 37232

³Jim Ayers Institute for Precancer Detection and Diagnosis, Vanderbilt University School of Medicine, Nashville, TN 37232

⁴Department of Cancer Biology, Vanderbilt University School of Medicine, Nashville, TN 37232

⁵Department of Biochemistry, Vanderbilt University School of Medicine, Nashville, TN 37232

Abstract

The standard shotgun proteomics data analysis strategy relies on searching MS/MS spectra against a context-independent protein sequence database derived from the complete genome sequence of an organism. Because transcriptome sequence analysis (RNA-Seq) promises an unbiased and comprehensive picture of the transcriptome, we reason that a sample-specific protein database derived from RNA-Seq data can better approximate the real protein pool in the sample and thus improve protein identification. In this study, we have developed a two-step strategy for building sample-specific protein databases from RNA-Seq data. First, the database size is reduced by eliminating unexpressed or lowly expressed genes according to transcript quantification. Secondly, high-quality nonsynonymous coding single nucleotide variations (SNVs) are identified based on RNA-Seq data, and corresponding protein variants are added to the database. Using RNA-Seq and shotgun proteomics data from two colorectal cancer cell lines SW480 and RKO, we demonstrated that customized protein sequence databases could significantly increase the sensitivity of peptide identification, reduce ambiguity in protein assembly, and enable the detection of known and novel peptide variants. Thus, sample-specific databases from RNA-Seq data can enable more sensitive and comprehensive protein discovery in shotgun proteomics studies.

Keywords

RNA-Seq; Shotgun Proteomics; Peptide Identification; Single Nucleotide Variations; data integration

Introduction

Among different strategies for proteome profiling, the tandem mass spectrometry (MS/MS)-based shotgun proteomics technology is the most effective option for large-scale protein identification in complex samples¹. In a typical shotgun proteomics experiment, proteins are enzymatically digested into peptides, separated by liquid chromatography (LC), and

*Correspondence should be addressed to B.Z. bing.zhang@vanderbilt.edu.

identified by MS/MS. Next, MS/MS data acquired from the analyses are processed to identify peptides that gave rise to observed spectra, and proteins are inferred based on identified peptides. Protein sequence database searching is the most commonly used technique for the identification of peptides following the acquisition of MS/MS spectra. Database search engines have been developed for this purpose, including Sequest², Mascot³, X!Tandem⁴, and MyriMatch⁵, among many others. Protein sequence databases from public resources such as RefSeq, Uniprot, IPI and ENSEMBL are usually used as reference databases for a search.

Although convenient for routine use, these public databases are collections of all known and predicted proteins in a species and may not closely represent the real protein pool in a specific sample. According to the current annotation in ENSEMBL, the number of proteins encoded by the human genome is more than 45,000. Presumably, only a fraction of these proteins are expressed in a specific sample, and the number of expressed proteins may vary substantially in different samples. Larger databases yield more distraction, lower signal to noise ratio, and reduced sensitivity under search criteria needed to maintain a low false discovery rate⁶. Ramakrishnan *et al.*⁷ incorporated mRNA concentration measured by microarray as prior knowledge of protein presence to improve protein identification in shotgun proteomics experiments. Although promising, this method makes an assumption of high correlation between mRNA and protein abundance, which might not be realistic⁸.

Meanwhile, databases employed in proteomics searches are usually incomplete with respect to sequence variation information, such as Single Nucleotide Variations (SNVs) and RNA-splice and -editing variants. These genomic variations are closely related to phenotypic variations or pathogenesis. Without taking them into account, proteomic studies may fail to detect novel, important protein forms. Efforts have been made to enable the identification of protein sequence variations through incorporating genomic variation information from databases such as dbSNP and COSMIC^{9–12}, ENREF_3 and inferring alternative splice variants from the Expressed Sequence Tags (EST) database^{13–15}. These approaches provide an opportunity to detect protein variants. However, they may lead to significantly increased database size and higher risk of false positive identifications¹⁰.

In order to gain a complete understanding of cellular systems at both genomic and proteomic levels, many research groups have started to apply RNA and protein profiling technologies in parallel to the same samples^{16–18}. Although protein abundance can only be partially explained by mRNA concentration⁸, mRNA expression is clearly a prerequisite for protein expression. The emerging high-throughput RNA sequencing (RNA-Seq) technology offers an opportunity to obtain transcript expression levels and sequence variations simultaneously. Because RNA-Seq promises an unbiased and comprehensive picture of the transcriptome, we reason that a sample-specific protein sequence database derived from RNA-Seq data can better represent the real protein pool in a sample and thus improve protein identification in shotgun proteomics studies.

Despite their obvious potential, RNA-Seq-derived customized databases have seen limited use in proteomics searching. There are several important issues to be addressed before this promising approach can be fully adopted. First, how can we effectively reduce the database size by incorporating RNA-Seq data? Secondly, how can we sift through millions of candidate SNVs to find reliable non-synonymous coding SNVs? Thirdly, how much performance-gain can we achieve using this approach? In addition to benchmarking the performance-gain, we need to detect the extent of protein loss associated with this technique.

In this study, we investigated the above questions using matched RNA-Seq and shotgun proteomics data from two colorectal cancer cell lines: SW480 and RKO. We demonstrated

that customized protein sequence databases could significantly increase the sensitivity in peptide identification, reduce ambiguity in protein assembly, and enable the detection of known and novel peptide variants. Based on the results, we have proposed a workflow for constructing sample-specific protein sequence databases from RNA-Seq data for more sensitive and sequence variant-inclusive proteomic studies.

Methods

Data sets

Two colorectal cancer cell lines, SW480 and RKO, were used in this study. Proteomic analyses were conducted in the Ayers Institute at the Vanderbilt University. Detailed information about the cell lines and the proteomics experiments have been described in our previous paper¹¹. mzML files for this data set were downloaded from ProteomeCommons.org Tranche using the following hash:

```
WUIXD5E3faTDkXevWvEAGfMTfHsoM
+ghM1egXBiwvmnOSv0PgDXGzKDY6gk1KbIb2AihSfUX9CWH/
WTQAmKhfLzn1e0AAAAAAAJJg==
```

Single read RNA-Seq data were generated for SW480 and RKO using the Illumina Genome Analyzer IIx (Illumina GAIIx) platform by the Vanderbilt Genome Sciences Resource. The read length was 43 bp. Total RNA was isolated from the cells using the RNeasy minikit from Qiagen. mRNA was purified from total RNA using the Oligotex mRNA Mini Kit. The mRNA samples were run on the Agilent 2100 Bioanalyzer using the Pico Chip to assess ribosomal contamination. An input of 100 ng was used following the Illumina mRNA-Seq protocol. The RNA was chemically fragmented and converted to ds-cDNA. The ds-cDNA went through end-repair, 3' dA-tailing, and adapter ligation. Prior to PCR enrichment, samples were separated on a 2% agarose E-gel at 120V for 10min. A gel region was excised at ~200bp +/- 25bp and purified using the QIAGEN Gel Extraction Kit. The purified cDNA templates were enriched by PCR using 15 cycles, purified using QIA quick PCR Kit, eluting in 30ul EB. The library was validated for size and concentration by running on the Agilent Bioanalyzer DNA 1000 chip. The libraries were normalized to 10 nM, denatured, and 8 pM was loaded into the Illumina Cluster Station for cluster generation. The flow cell was loaded onto the Illumina GAIIx and the SR-36 recipe was used.

Microarray gene expression data for SW480 and RKO were downloaded from the Gene Expression Omnibus (GEO) database with the GEO accession number GSE10843 (<http://www.be-md.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi?acc=GSE10843>). This data set was generated using the Affymetrix Human Genome U133 Plus 2.0 Array and contains tens of cancer cell lines, including duplicates for both SW480 and RKO cell lines.

RNA-Seq data analysis

We used the Tophat software¹⁹ to align reads to the human reference genome (hg18, ENSEMBL v54) in a spliced mode. Using the default settings in Tophat, up to 2 mismatches were allowed in the alignment. Among the 29,553,736 reads in the SW480 data set, 86% were mapped to the genome. Among the 28,968,492 reads in the RKO data set, 87% were mapped to the genome. SAMtools²⁰ were used to convert the resulting SAM files to binary format (bam), which were then imported to the R software environment (<http://www.r-project.org/>) using the Rsamtools package (<http://www.bioconductor.org/packages/2.6/bioc/html/Rsamtools.html>). RPKM (Reads Per Kilo base per Million mapped reads)²¹ was calculated to represent the expression level for each transcript. `_ENREF_4` The pileup algorithm in SAMtools was used to generate a list of candidate SNVs based on bam files. We set additional filters (coverage, mapping quality and consensus quality) to control the

SNV quality and retained only nonsynonymous protein coding SNVs. In this process, we used the biomaRt package²² in R to retrieve transcript and protein information from the ENSEMBL v54 database. The Ensemble BioMart database (<http://www.ensembl.org/biomart/martview>) was used to match accession numbers in RNA-Seq, microarray, and proteomic data.

Microarray data analysis

We applied the Affymetrix MAS5 algorithm to obtain a present/absent call for each probe set. For each cell line, probe sets with consistent present or absent calls in both replicates were identified and related to RPKM values for their corresponding transcripts.

Proteomics database search

Three database search engines, MyriMatch⁵ (v1.6.63), Sequest² (TurboSEQUEST v.27 (rev. 12)), and X!Tandem⁴ (CYCLONE v2010.12.01.1) were used in this study. All cysteines were assumed to be carbamidomethylated and methionines were allowed to be oxidized. We used the semi-tryptic mode for database search. One missed cleavage was permitted. The configurations for all search engines are provided in Supplemental File 1. The IDPicker software (v2.6.142.0)²³ _ENREF_10 was used for protein assembly. IDPicker applies parsimony analysis in protein assembly to derive a minimum list of proteins that could account for all observed peptides, and proteins that could not be distinguished due to shared peptide sequences were grouped together as an indiscernible protein group²⁴. Ambiguous identifications that mapped to three or more peptide sequences with equal scores were excluded. Minimum peptide length was set to 6. Peptide identifications were filtered to achieve a False Discovery Rate of 5% for peptide-spectrum matches (PSMs). We only retained proteins with at least two distinct peptides. The msConvert²⁵ tool was used to generate an mzML file for all unassigned spectra after searching against the protein sequence database derived from the RNA-Seq data.

Results

Improved peptide identification using reduced protein databases derived from RNA-Seq data

In any cell, only a subset of genes is transcribed. Moreover, genes that are transcribed at a very low level produce limited amounts of proteins, which thus are likely to be undetectable by shotgun proteomics. RNA-Seq analysis provides a good estimate for the absolute expression level of transcripts through the RPKM measurement²¹, a metric that may be used to eliminate unexpressed or lowly expressed genes from a protein sequence database. Thus, an initial task is to identify an RPKM threshold indicating sufficient transcript abundance to predict probable detection of corresponding protein products by shotgun proteomics techniques.

We first investigated RPKM threshold selection in the colorectal cancer cell line SW480 (Figure 1A). Specifically, we compared the distributions of the logarithmic scale RPKM values for all detected transcripts by RNA-Seq, genes analyzed by microarray, and all identified proteins when searching MS/MS spectra against the ENSEMBL v54 human protein database. As shown in Figure 1A, RPKMs for all RNA-Seq detected transcripts followed a bimodal distribution (green curve), which was likely caused by two distinct transcript populations: unexpressed transcripts (background noise) and expressed transcripts in the sample (signal). To further explore this possibility, we overlaid the distributions for “present” and “absent” transcripts as determined by microarray analysis on the RNA-Seq distribution. The “present” transcripts showed a unimodal distribution (solid blue curve), which correlated well with the major peak in RNA-Seq. 95% of the “present” transcripts had

a RPKM value greater than 0.616 (dashed black line). This RPKM value seemed to provide a reasonably good separation between the two peaks in the RNA-Seq distribution and could serve as an optional threshold to distinguish expressed transcripts from unexpressed ones. Interestingly, the “absent” transcripts showed a bimodal distribution (dashed blue curve) similar to RNA-Seq, suggesting that many of these transcripts were actually expressed but incorrectly classified as “absent” by microarray.

For transcripts that encode proteins detected by shotgun proteomics, the distribution was very similar to that of the microarray “present” transcripts, except for a small shift to larger RPKM values, suggesting that highly expressed transcripts are more likely to be detected by shotgun proteomics. The distribution also showed a sharp increase in protein identification beyond RPKM 2 (dashed red line), which led us to empirically choose this value as an optional threshold to eliminate low-abundance transcripts that are likely to be undetectable by shotgun proteomics.

Using the above two threshold options (Microarray95 and RPKM2), we constructed reduced protein sequence databases by eliminating proteins whose corresponding transcript expression was below the selected threshold. The regular ENSEMBL protein database contains 47,509 proteins. Applying the Microarray95 and RPKM2 thresholds from the SW480 data set reduced databases to 31,124 and 24,951 proteins, respectively.

To benchmark the performance advantage of the reduced databases, we compared the search results from three databases (ENSEMBL v54, with Microarray95 and RPKM2 thresholds) for the SW480 data set. As shown in Table 1, both reduced databases performed better than the regular database, and the database corresponding to the RPKM2 cutoff had the best performance. As compared to the regular database, the RPKM2 database for SW480 increased the number of identifiable spectra by 896 (5.0%), the number of identified peptides by 397 (6.1%), and the number of indiscernible protein groups by 81 (5.9%).

To test whether the findings could be generalized to other data sets, we applied the same analysis procedure to the RKO data set. The RPKM distribution plot for RNA-Seq, microarray, and shotgun proteomics showed similar patterns as observed for SW480 (Figure 1B). Reduced databases for RKO contained 28,756 and 23,172 protein sequences using the Microarray95 and RPKM2 thresholds, respectively. We searched these databases as well as the regular database with the proteomics data. Consistent with above observations, the RPKM2 database performed the best among the three databases (Table 1), with the total number of identifiable spectra increased by 1055 (4.9%), the number of identified peptides increased by 533 (5.8%), and the number of indiscernible protein groups increased by 84 (4.1%) as compared to the regular database. Application of an RPKM threshold for transcript abundance thus reduces the size of the reference protein sequence database and improves the sensitivity of peptide identification in shotgun proteomics.

Reduced ambiguity in protein assembly using reduced protein database

Although the reduced RPKM2 database increased the number of identified indiscernible protein groups by about 5% in both cell lines, the total number of all proteins in these groups did not increase proportionally (Table 1). Increased sensitivity in peptide identification may lead to better discrimination power in protein assembly and thus may remove suspicious proteins for more parsimonious protein reporting.

Figure 2A uses a real example to illustrate this possibility. Using the regular database, an indiscernible protein group with four proteins was identified with four supporting peptides. With the reduced RPKM2 database, we identified an additional peptide uniquely mapped to

one of the four proteins. The parsimony algorithm implemented in IDPicker²⁴ eliminated the other three proteins and reduced the size of this protein group to one.

To examine this effect at a global level, we compared the sizes of the indiscernible protein groups identified with the regular database and the RPKM2 database. While the numbers of smaller protein groups (1~5 proteins per group) increased in both cell lines, the numbers of protein groups with more than 5 proteins decreased considerably in both cell lines (Figure 2B). These results demonstrate that the reduced database can reduce ambiguity in protein assembly.

Variant peptide identification based on SNVs derived from RNA-Seq

Detection of variant peptides is an important potential advantage of RNA-Seq-derived databases. To evaluate this possibility, we generated customized variation-inclusive protein sequence databases for proteomics search. Initial RNA-Seq data analysis identified over a million candidate SNVs in both cell lines. Based on parameter settings reported in previous studies²⁶, we set up a pipeline to filter for high-quality nonsynonymous SNVs in protein coding regions with an RPKM greater than 2. Figure 3 illustrates the pipeline filtering process, the yields of SNV identifications and reduction ratios at each step with data from both cell lines (Figure 3).

Using the pipeline, we detected 3,501 nonsynonymous SNVs in SW480 and 3,995 in RKO. Of these, 2,032 (58.0%) SNVs in SW480 and 2,091 (52.3%) SNVs in RKO were found in dbSNP54. The nonsynonymous SNVs were mapped to 5,437 proteins in RKO and 4,838 proteins in SW480. These protein sequences with SNVs were added to the RPKM2 protein database to create customized databases for SW480 and RKO, respectively.

The customized databases comprised 29,789 entries in SW480 and 28,609 in RKO. Proteomics data sets were then searched against these databases. The numbers of identified spectra, peptides and protein groups were 18,760, 6,958 and 1,467, respectively, in SW480. In RKO, the numbers were 22,623 spectra, 9,728 peptides and 2,129 protein groups (Table 1). A total of 33 unique variant peptides were identified in SW480, of which 23 (69.7%) were supported by dbSNP54. In RKO cells, 43 unique peptides were detected, of which 30 (69.8%) were supported by dbSNP54. A complete list of variant peptides and related information can be found in Supplementary Files 2 and 3. The high dbSNP support rate for the identified variant peptides suggests high reliability of our variant peptide detection pipeline.

As somatic mutations play important roles in cancer, we further searched for candidate somatic mutations in these colorectal cancer cell lines by removing known variations listed in the dbSNP database. This identified 10 SNV-containing peptides in SW480 and 14 in RKO (Table 2). Among variations in these peptides, the *TP53*^{P309S} mutation is catalogued in the COSMIC database and has been reported as a gain of function mutation that contributes to increased cell proliferation and resistance to anticancer drugs in SW480^{27, 28}. The *NLE1*^{Q319K} mutation found in RKO is also catalogued in COSMIC but has not been reported in the RKO cell line.

To investigate the potential functional impact of the newly identified candidate somatic mutations, we used the MutationAssessor²⁹ software to calculate a functional impact score (FIS) for each amino acid substitution. The calculation is based on evolutionary conservation of the affected amino acid in protein homologs. About half of the candidate mutations that could be analyzed by MutationAssessor are of medium (FIS > 1) or high (FIS > 2) impact to proteins, while the average FIS for all SNPs detected in this study was about 0. The *HSP90AA1*^{D393N} mutation in RKO is of particular interest. The mutation had a very

high FIS of 3.24 and the protein was highly expressed in RKO, as indicated by a total spectral count of 153. HSP90AA1 is critical for correct conformation and stability of key oncogenic proteins involved in signal transduction pathways leading to cell proliferation, apoptosis, invasion, angiogenesis and metastasis^{30, 31}_ENREF_28. Over-expression of HSP90AA1 has been associated with the progression, invasion and metastasis of colorectal cancer^{32, 33}_ENREF_28. In addition to D393N, HSP90AA1 harbored another mutation N504D, which also had a relatively high FIS (1.64). Besides TP53 in SW480 and HSP90AA1 in RKO, many other genes listed in Table 2, including those harboring mutations with medium or high functional impact, have known roles in cancer.

Rescue of unassigned spectra

The above results clearly demonstrated that customized databases derived from RNA-Seq significantly improve protein identification in shotgun proteomics. However, an equally important question is whether we have missed certain proteins by applying this procedure. To identify proteins that might have been missed due to the empirical cutoff selection, we re-searched unassigned spectra for each cell line against all ENSEMBL proteins with corresponding RPKM values less than 2 in the cell line. The databases contained 22,558 sequences for SW480 and 24,337 for RKO. The search results are shown in Supplementary File 4. We rescued 74 peptides in SW480 and 49 in RKO, corresponding to 23 indiscernible protein groups in both cell lines. They corresponded to 0.7% of the identifiable spectra in SW480 and 0.3% in RKO with a regular database search.

Interestingly, half of the rescued proteins in SW480 (30 out of 59) and one third in RKO (18 out of 66) were histones. This may relate to the use of the polyA RNA enrichment Oligotex mRNA Kits in mRNA preparation for RNA-Seq analysis. Histone mRNAs are known to lack polyA tails and their 3' ends have a stem-loop structure, hence the commonly used RNA-Seq protocol cannot capture this class of genes^{34, 35}. According to the IDPicker report, except for the histones, all other missed proteins had less than 3 spectral counts, which may be explained by either low-abundance or spurious identification. Therefore, adding histones to a customized database should address the primary limitation of the RNA-Seq based database search approach.

Workflow for deriving customized databases from RNA-Seq data

Based on above results, we proposed a workflow for deriving customized protein sequence databases from RNA-Seq data (Figure 4). A customized database is generated in two major steps. First, an RPKM threshold is selected to eliminate proteins with insufficient corresponding transcript abundance, because they are likely to be undetectable by shotgun proteomics. The threshold can be empirically selected by plotting the RPKM distribution for all identified proteins when searching MS/MS spectra against the regular protein database and then identifying an RPKM value above which a sharp increase in protein identification is observed. Secondly, high-quality nonsynonymous coding SNVs are identified based on RNA-Seq data and corresponding proteins are added to the customized database. The easiest way to ensure compatibility with different search engines is to keep the entire variant containing protein sequence as an entry. Related variation information, such as variation position, change status and corresponding dbSNP ID (if available), should be included in the sequence header for easy interpretation of the search result. Finally, because RNA-Seq usually fails to detect histones, these proteins should be added to the customized databases. The database should also append reverse sequences as decoy sequences for FDR estimation.

Applying this workflow resulted in a customized database for SW480 with 29,894 proteins and one for RKO with 28,713 proteins, each including 105 histone proteins. Searching shotgun proteomics data from these two cell lines against the customized databases led to

the identification of 6,984 peptides in SW480 and 9,730 peptides in RKO, which corresponded to 7.1% and 6.0% increase respectively as compared to search results from the regular database. Finally, we tested the customized databases with two other popular search tools, X!tandem and Sequest, and observed similar level of improvement (Table 3).

Discussion

Numerous studies have reported that RNA-Seq is superior to microarrays to characterize transcriptomes^{36, 37}. This sequencing technology not only can accurately measure the abundance of transcripts, but also can detect sequence variations at the same time³⁸, thus providing a more complete view of the transcriptome. Meanwhile, it is increasingly clear that mRNA and protein expression data are complementary. Concurrent measurement of both provides a better understanding of complex biological systems^{18, 39–42} and provides a strong rationale to integrate RNA-Seq data into proteomics studies.

Previous integrative analyses focused primarily on the correlation between steady state mRNA and protein abundance or between induced changes in mRNA and protein expression^{8, 18, 39–42}. Although technology advancements have made these comparisons possible on an increasingly large scale, protein identification and quantification remains the limiting factor in such studies¹⁸. We addressed this challenge by asking how protein identification can be improved through integrating transcriptomic data. To this end, we have developed a workflow that creates sample-specific protein sequence databases from RNA-Seq data to facilitate protein identification in shotgun proteomics.

We first demonstrated that use of an RPKM threshold to eliminate entries for low-abundance and unexpressed genes from a protein sequence database could improve the sensitivity of peptide identification and reduce ambiguity in protein assembly. Besides protein identification, spectral count data derived from shotgun proteomics has provided a valuable means for the quantification of protein abundance and detection of differentially expression proteins^{43, 44}. Because a reduced database could increase the number of identifiable spectra by more than 5%, it is expected to have impact on spectral counting based quantitative analysis, especially for lowly expressed proteins. This aspect will be investigated in a future study.

A reduced database relies on the selection of an RPKM threshold. Various approaches have been proposed for setting the RPKM level corresponding to detectable mRNA expression. Ramsköld et al. used the expression level of intergenic region as the background to choose a threshold to define expression⁴⁵. Gan et al. set a RPKM threshold that allows the detection of 99% of the “present” transcripts as determined by matched microarray study⁴⁶. Because shotgun proteomics is obviously less sensitive than RNA-Seq, we selected a higher RPKM threshold that would allow detectable protein expression. A threshold determined from the RPKM distribution for all proteins identified through searching against the regular protein database could maximize the numbers of identifiable spectra and peptide identifications. Although easy to implement, dichotomizing data using a selected threshold usually results in a loss of information. Future work may directly incorporate RPKM data as Bayesian prior information in database search tools. Thus, a lower prior likelihood can be given to peptides derived from transcripts with lower RPKM values.

We also showed that incorporating SNVs identified in RNA-Seq into a protein sequence database could effectively enable variant peptide identification. In a recently published study, we tackled the same problem using a different approach, namely integrating all known variations from dbSNP, COSMIC, and other genomic mutation databases to the regular protein sequence database¹¹. Because the same proteomics data sets were used for

both studies, we were able to perform a direct comparison between the two approaches. The current approach identified 33 and 43 variant peptides in SW480 and RKO respectively. In comparison, the previous approach only identified 20 and 27 variant peptides in SW480 and RKO respectively. Among the variant peptides identified in the previous study, 6 and 15 were missed in the current study. A closer look at the data found that these could be explained by the elimination of peptides from proteins supported by a single peptide in this study, by the stringent filter used for SNV calling, by a lack of read coverage for the position, and by a few possible false discoveries in the previous study. This result suggested that, although the sample-specific database identified significantly more variant peptides, its performance could be further improved with increased read depth in RNA-Seq experiments and refined SNV calling algorithms. On the other hand, the previous approach remains an effective option when matched RNA-Seq data is not available.

Detecting variants at the peptide level may seem redundant when they are already found at the transcript level. However, protein-level data can provide complementary information on the functional consequence of the mutations. In order to test whether nonsynonymous coding SNVs may affect protein stability at a large scale, we examined the detectable ratio for normal and variant peptides. To estimate the total number of peptides, the protein sequences were *in silico* trypsin digested without allowing for missed enzyme cleavages and resultant peptides with less than 6 amino acids were excluded. In SW480, 33 out of 3041 variant peptides were detected, resulting in a ratio of 1.1%. The number in RKO was 1.2% (43 out of 3503). In comparison, the detectable ratio for normal peptides based on the same estimation was 2.0% for SW480 and 3.0% for RKO. This result showed that variations indeed led to lower detection ratio, possibly due to reduced protein stability.

However, not all variant proteins have accelerated degradation; some variants may have increased protein stability and expression levels. This is particularly true for gain of function mutations in cancer, because protein expression is required for performing the newly acquired function. In this case, proteomics data can serve as an efficient functional filter for the biologically important variations, such as the well-known *TP53*^{P309S} mutation and the newly identified *HSP90AA*^{D393N} mutation in this study. One great challenge presented by next generation sequencing technology is how to sift through tens of thousands of SNVs to identify a manageable number for further experimental investigation. One obvious approach is to focus on the nonsynonymous coding SNVs. However, there are still thousands, as shown in this study. We believe that protein level identification by shotgun proteomics, in combination with functional impact estimation, can serve as a useful approach for prioritizing important SNVs for targeted proteomics assay and functional analysis. From a clinical point of view, proteins with newly acquired oncogenic properties due to sequence variations represent important candidate therapeutic targets because their new functions may be essential in cancer cells, but non-essential in normal cells.

This framework generates sample-specific databases from RNA-Seq data to enable more sensitive and variation-aware protein discovery. Although here we focused on SNVs, RNA-Seq data could also facilitate the identification of novel alternative splice forms or fusion genes. A recent study suggests that novel alternative splice forms identified in RNA-Seq can be identified in proteomics data, although at a very limited scale⁴⁷. With the improved sensitivity in proteomics and the availability of RNA-Seq data with paired-end reads and longer read length, novel transcripts derived from RNA-Seq data might become a useful addition in our framework for more comprehensive customized protein sequence databases.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the National Institutes of Health (NIH)/National Cancer Institute (NCI) through grant R01 CA126218, the NIH/National Institute of General Medical Sciences (NIGMS) through grant R01 GM088822, and the NCI Clinical Proteomic Technologies Assessment for Cancer (CPTAC) program through grant U24 CA126479. This work was conducted in part using the resources of the Advanced Computing Center for Research and Education at Vanderbilt University, Nashville, TN.

References

1. Gstaiger M, Aebersold R. Applying mass spectrometry-based proteomics to genetics, genomics and network biology. *Nat Rev Genet.* 2009; 10(9):617–27. [PubMed: 19687803]
2. Eng JK, McCormack AL, Yates JR 3rd. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry.* 1994; 5(11):976–989.
3. Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis.* 1999; 20(18):3551–67. [PubMed: 10612281]
4. Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics.* 2004; 20(9):1466–7. [PubMed: 14976030]
5. Tabb DL, Fernando CG, Chambers MC. MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J Proteome Res.* 2007; 6(2):654–61. [PubMed: 17269722]
6. 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):2092–123. [PubMed: 20816881]
7. Ramakrishnan SR, Vogel C, Prince JT, Li Z, Penalva LO, Myers M, Marcotte EM, Miranker DP, Wang R. Integrating shotgun proteomics and mRNA expression data to improve protein identification. *Bioinformatics.* 2009; 25(11):1397–403. [PubMed: 19318424]
8. Vogel C, Abreu Rde S, Ko D, Le SY, Shapiro BA, Burns SC, Sandhu D, Boutz DR, Marcotte EM, Penalva LO. Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Molecular systems biology.* 2010; 6:400. [PubMed: 20739923]
9. Alves G, Ogurtsov AY, Yu YK. RAId_DbS: mass-spectrometry based peptide identification web server with knowledge integration. *BMC Genomics.* 2008; 9:505. [PubMed: 18954448]
10. Bunger MK, Cargile BJ, Sevinsky JR, Deyanova E, Yates NA, Hendrickson RC, Stephenson JL Jr. Detection and validation of non-synonymous coding SNPs from orthogonal analysis of shotgun proteomics data. *J Proteome Res.* 2007; 6(6):2331–40. [PubMed: 17488105]
11. Li J, Su Z, Ma ZQ, Slebos RJ, Halvey P, Tabb DL, Liebler DC, Pao W, Zhang B. A bioinformatics workflow for variant Peptide detection in shotgun proteomics. *Mol Cell Proteomics.* 2011; 10(5) M110006536.
12. Schandorff S, Olsen JV, Bunkenborg J, Blagoev B, Zhang Y, Andersen JS, Mann M. A mass spectrometry-friendly database for cSNP identification. *Nat Methods.* 2007; 4(6):465–6. [PubMed: 17538625]
13. Chang KY, Georgianna DR, Heber S, Payne GA, Muddiman DC. Detection of alternative splice variants at the proteome level in *Aspergillus flavus*. *Journal of proteome research.* 2010; 9(3): 1209–17. [PubMed: 20047314]
14. Edwards NJ. Novel peptide identification from tandem mass spectra using ESTs and sequence database compression. *Mol Syst Biol.* 2007; 3:102. [PubMed: 17437027]
15. Fermin D, Allen BB, Blackwell TW, Menon R, Adamski M, Xu Y, Ulintz P, Omenn GS, States DJ. Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics. *Genome biology.* 2006; 7(4):R35. [PubMed: 16646984]
16. Desgagne-Penix I, Khan MF, Schriemer DC, Cram D, Nowak J, Facchini PJ. Integration of deep transcriptome and proteome analyses reveals the components of alkaloid metabolism in opium poppy cell cultures. *BMC Plant Biol.* 2010; 10:252. [PubMed: 21083930]

17. Adamidi C, Wang Y, Gruen D, Mastrobuoni G, You X, Tolle D, Dodt M, Mackowiak SD, Gogol-Doering A, Oenal P, Rybak A, Ross E, Alvarado AS, Kempa S, Dieterich C, Rajewsky N, Chen W. De novo assembly and validation of planaria transcriptome by massive parallel sequencing and shotgun proteomics. *Genome Res.* 2011
18. Lundberg E, Fagerberg L, Klevebring D, Matic I, Geiger T, Cox J, Algenas C, Lundberg J, Mann M, Uhlen M. Defining the transcriptome and proteome in three functionally different human cell lines. *Mol Syst Biol.* 2010; 6:450. [PubMed: 21179022]
19. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics.* 2009; 25(9):1105–11. [PubMed: 19289445]
20. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* 2009; 25(16):2078–9. [PubMed: 19505943]
21. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods.* 2008; 5(7):621–8. [PubMed: 18516045]
22. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc.* 2009; 4(8):1184–91. [PubMed: 19617889]
23. Ma ZQ, Dasari S, Chambers MC, Litton MD, Sobecki SM, Zimmerman LJ, Halvey PJ, Schilling B, Drake PM, Gibson BW, Tabb DL. IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. *J Proteome Res.* 2009; 8(8):3872–81. [PubMed: 19522537]
24. Zhang B, Chambers MC, Tabb DL. Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. *Journal of proteome research.* 2007; 6(9):3549–57. [PubMed: 17676885]
25. Kessner D, Chambers M, Burke R, Agus D, Mallick P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics.* 2008; 24(21):2534–6. [PubMed: 18606607]
26. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* 2008; 18(11):1851–8. [PubMed: 18714091]
27. Bossi G, Lapi E, Strano S, Rinaldo C, Blandino G, Sacchi A. Mutant p53 gain of function: reduction of tumor malignancy of human cancer cell lines through abrogation of mutant p53 expression. *Oncogene.* 2006; 25(2):304–9. [PubMed: 16170357]
28. Yan W, Liu G, Scoumanne A, Chen X. Suppression of inhibitor of differentiation 2, a target of mutant p53, is required for gain-of-function mutations. *Cancer Res.* 2008; 68(16):6789–96. [PubMed: 18701504]
29. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* 2011
30. Maloney A, Workman P. HSP90 as a new therapeutic target for cancer therapy: the story unfolds. *Expert opinion on biological therapy.* 2002; 2(1):3–24. [PubMed: 11772336]
31. Pearl LH, Prodromou C. Structure and mechanism of the Hsp90 molecular chaperone machinery. *Annu Rev Biochem.* 2006; 75:271–94. [PubMed: 16756493]
32. Milicevic Z, Bogojevic D, Mihailovic M, Petrovic M, Krivokapic Z. Molecular characterization of hsp90 isoforms in colorectal cancer cells and its association with tumour progression. *International journal of oncology.* 2008; 32(6):1169–78. [PubMed: 18497978]
33. Park KA, Byun HS, Won M, Yang KJ, Shin S, Piao L, Kim JM, Yoon WH, Junn E, Park J, Seok JH, Hur GM. Sustained activation of protein kinase C downregulates nuclear factor-kappaB signaling by dissociation of IKK-gamma and Hsp90 complex in human colonic epithelial cells. *Carcinogenesis.* 2007; 28(1):71–80. [PubMed: 16774932]
34. Yang L, Duff MO, Graveley BR, Carmichael GG, Chen LL. Genomewide characterization of non-polyadenylated RNAs. *Genome Biol.* 2011; 12(2):R16. [PubMed: 21324177]
35. Marzluff WF, Wagner EJ, Duronio RJ. Metabolism and regulation of canonical histone mRNAs: life without a poly(A) tail. *Nat Rev Genet.* 2008; 9(11):843–54. [PubMed: 18927579]
36. Fu X, Fu N, Guo S, Yan Z, Xu Y, Hu H, Menzel C, Chen W, Li Y, Zeng R, Khaitovich P. Estimating accuracy of RNA-Seq and microarrays with proteomics. *BMC Genomics.* 2009; 10:161. [PubMed: 19371429]

37. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* 2008; 18(9):1509–17. [PubMed: 18550803]
38. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet.* 2009; 10(1):57–63. [PubMed: 19015660]
39. Greenbaum D, Colangelo C, Williams K, Gerstein M. Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome biology.* 2003; 4(9):117. [PubMed: 12952525]
40. Griffin TJ, Gygi SP, Ideker T, Rist B, Eng J, Hood L, Aebersold R. Complementary profiling of gene expression at the transcriptome and proteome levels in *Saccharomyces cerevisiae*. *Molecular & cellular proteomics : MCP.* 2002; 1(4):323–33. [PubMed: 12096114]
41. Tian Q, Stepaniants SB, Mao M, Weng L, Feetham MC, Doyle MJ, Yi EC, Dai H, Thorsson V, Eng J, Goodlett D, Berger JP, Gunter B, Linseley PS, Stoughton RB, Aebersold R, Collins SJ, Hanlon WA, Hood LE. Integrated genomic and proteomic analyses of gene expression in Mammalian cells. *Molecular & cellular proteomics : MCP.* 2004; 3(10):960–9. [PubMed: 15238602]
42. Washburn MP, Koller A, Oshiro G, Ulaszek RR, Plouffe D, Deciu C, Winzeler E, Yates JR 3rd. Protein pathway and complex clustering of correlated mRNA and protein expression analyses in *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences of the United States of America.* 2003; 100(6):3107–12. [PubMed: 12626741]
43. Zhang B, VerBerkmoes NC, Langston MA, Uberbacher E, Hettich RL, Samatova NF. Detecting differential and correlated protein expression in label-free shotgun proteomics. *Journal of proteome research.* 2006; 5(11):2909–18. [PubMed: 17081042]
44. Liu H, Sadygov RG, Yates JR 3rd. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Analytical chemistry.* 2004; 76(14):4193–201. [PubMed: 15253663]
45. Rasmakold D, Wang ET, Burge CB, Sandberg R. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput Biol.* 2009; 5(12):e1000598. [PubMed: 20011106]
46. Gan Q, Schones DE, HoEun S, Wei G, Cui K, Zhao K, Chen X. Monovalent and unpoised status of most genes in undifferentiated cell-enriched *Drosophila* testis. *Genome biology.* 2010; 11(4):R42. [PubMed: 20398323]
47. Ning K, Nesvizhskii AI. The utility of mass spectrometry-based proteomic data for validation of novel alternative splice forms reconstructed from RNA-Seq data: a preliminary assessment. *BMC Bioinformatics.* 2010; 11(Suppl 11):S14. [PubMed: 21172049]

Figure 1.

The distribution of log₂RPKM for three technologies in SW480 and RKO. The green line represents the distribution for all transcripts detected by RNA-Seq. The blue lines represent microarray data, solid for transcripts with a present call and dashed for absent call. The red line is the distribution for all proteins detected by shotgun proteomics using regular database searching. The black vertical dashed line corresponds to the 95% quantile of microarray present call and the red vertical dashed line corresponds to RPKM2.

Figure 2. Reduction of ambiguity in protein identification. A) A specific example showing how protein group becomes smaller. B) Barplot shows that the number of small groups increases while that of big groups decreases in comparison with the regular database, suggesting that the customized database can reduce ambiguity in protein assembly.

Figure 3. SNV detection pipeline. The SNV detection pipeline set up in this study. Results at each step for the two cell lines are also presented. #PROs: number of proteins.

Figure 4.

Workflow of database generation based on RNA-Seq data. The whole pipeline is divided into two components, including the identification of expressed proteins and the identification of variations. Then they are combined to build a comprehensive sample specific database for proteomics search. Genes known to be missed in RNA-Seq, such as the histones should be manually added to the database.

Table 1

Summary of identifications for SW480 and RKO cell lines.

Cell line	Database	Database size	Identifiable spectra	Identified peptides	Indiscernible protein groups	All possible proteins
SW480	Regular database	47,509	17,791	6,524	1,381	2,856
	Microarray95	31,124	18,448	6,813	1,444	2,974
	RPKM2	24,951	18,687	6,921	1,462	2,959
	RPKM2 + SNV s	24,951+4,838	18,760	6,958	1,467	–
	Regular database	47,509	21,537	9,178	2,040	4,133
RKO	Microarray95	28,756	22,184	9,522	2,107	4,210
	RPKM2	23,172	22,592	9,711	2,124	4,211
	RPKM2 + SNV s	23,172+5,436	22,623	9,728	2,129	–
	Regular database	47,509	21,537	9,178	2,040	4,133

Table 2

Identified variant peptides that are not present in dbSNP54 for the two cell lines

No.	Peptide ^a	Gene Symbol	Protein	Variation ^b	Coverage	Peptide Spectra Count	Protein Spectra Count	FI Score ^c
SW480 cell line								
1	ALSNNTSSSPQPKK	TP53	ENSP00000269305 ENSP00000352610	P309S	5% 6%	5	8	1.45
2	DDYPSRGYSDR	RBMXL1	ENSP00000318415	G261S	22%	1	36	-0.68
3	EYNLNFSGSSAIQEVK	FAF1	ENSP00000379457	T216A	5%	1	3	1.47
4	GQGITGVEDK	TKT	ENSP00000296289	R246Q	9%	2	12	1.13
5	GSPLSAAPYGIESMSQDTEVR	PTPN1	ENSP00000360683	N355S	8%	4	5	0.32
6	ILNHFDAQLPESMEDDGLSER	UTP20	ENSP00000261637	V658A	3%	1	10	-0.41
7	ISQYSEIAIQMMMHMHQPK EICALVGFCDEVK	PSAP	ENSP00000378392	L260H	15%	1	24	0.31
8	SAEGAAKEEPK	NULL	ENSP00000368252	S17P	50%	1	35	-
9	SSAEGAAKEEPK	NULL	ENSP00000368252	S17P	50%	2	35	-
10	VSSAEGAAKEEPK	NULL	ENSP00000368252	S17P	50%	2	35	-
RKO cell line								
1	ALEQLFQDEVAK	PLEC	ENSP00000323856	K2652E	19%	3	165	0.89
2	DQVADSAFVER	HSP90AA1	ENSP00000216281 ENSP00000335153	N504D N626D	38% 33%	5	153	1.64
3	DSGSQEVLSLR	UTP14C	ENSP00000367361 ENSP00000367364	S450G	6%	4	10	0.03
4	EIEGSQVTFDVALSPK	CTBS	ENSP00000359659 ENSP00000359664	W91L W182L	9% 7%	1	2	1.08
5	ERWLDPLK	SMC5	ENSP00000354957	N930D	2%	1	2	0.86
6	GAMGGGVGASSGGGPGGS GGGGSGGPGAGTSFPPPGK	GSK3A	ENSP00000222330	V91G	15%	1	2	0.35
7	GVVDSENLPLNISR	HSP90AA1	ENSP00000216281 ENSP00000335153	D393N D515N	38% 33%	4	153	3.24
8	LESLNIQCEK	EIF3A	ENSP00000358140	R589C	16%	1	61	1.66
9	QTITWMDNMGRAGEMELK	TOP2A	ENSP00000350217	A193T	6%	1	16	-
10	SRGQLDTYTIASPPER	DOCK7	ENSP00000340742 ENSP00000360182	G1388D	1%	1	3	0.42
11	TGAFEPAEASVNPQDLK	NLE1	ENSP00000262335 ENSP00000354075	Q319K Q321K	6%	1	4	-0.12
12	VLEAELNDLM	CTDP1	ENSP00000299543	A952V	2%	1	2	1.04
13	VNPHSLPDPEVNER	POP1	ENSP00000339529 ENSP00000385787	Q74R	7%	1	13	-
14	LAVLALASSENSSTPEECEET SEKPK	NOP56	ENSP00000370589	A456V	19%	1	24	0.90

^a Amino acid sequence of peptides with variation residue in red.^b Variations in blue could be found in COSMIC53.^c Functional impact score generated by MutationAssessor. < 0 is neutral; 0~1 is low; 1~2 is medium; >2 is high. Scores represent medium or high impact are colored in red.

Table 3

Improvement of peptide identifications for three search engines using our customized database in two cell lines.

	SW480		RKO	
	Regular	Customized	Regular	Customized
MyriMatch	6,524	6,974 (7.1%)	9,178	9,730 (6.0%)
X!tandem	6,383	6,825 (6.9%)	9,833	10,452 (6.3%)
Sequest	4,811	5,148 (7.0%)	7,748	8,162 (5.3%)