Journal of proteome research

ARTICLE

pubs.acs.org/jpr

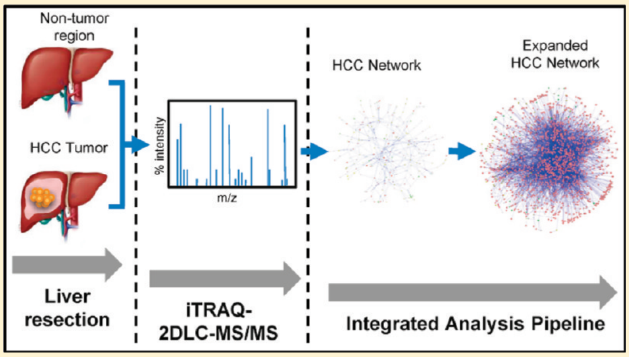
Network-Based Pipeline for Analyzing MS Data: An Application toward Liver Cancer

Wilson Wen Bin Goh,[†,‡,§] Yie Hou Lee,[†,‖] Ramdzan M. Zubaidah,[⊥] Jingjing Jin,[§] Difeng Dong,[§] Qingsong Lin,[#] Maxey C. M. Chung,[#,¶] and Limsoon Wong[*,§,◑]

[‡]Department of Computing, Imperial College London, South Kensington, London, United Kingdom
[§]School of Computing, National University of Singapore, Singapore
[‖]Singapore-MIT Alliance for Research and Technology, Singapore
[⊥]Rosalind and Morris Goodman Cancer Centre, McGill University, Canada
[#]Department of Biological Sciences, National University of Singapore, Singapore
[¶]Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
[◑]Department of Pathology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

S *Supporting Information*

ABSTRACT: Current limitations in proteome analysis by high-throughput mass spectrometry (MS) approaches have sometimes led to incomplete (or inconclusive) data sets being published or unpublished. In this work, we used an iTRAQ reference data on hepatocellular carcinoma (HCC) to design a two-stage functional analysis pipeline to widen and improve the proteome coverage and, subsequently, to unveil the molecular changes that occur during HCC progression in human tumorous tissue. The first involved functional cluster analysis by incorporating an expansion step on a cleaned integrated network. The second used an in-house developed pathway database where recovery of shared neighbors was followed by pathway enrichment analysis. In the original MS data set, over 500 proteins were detected from the tumors of 12 male patients, but in this paper we reported an additional 1000 proteins after application of our bioinformatics pipeline. Through an integrative effort of network cleaning, community finding methods, and network analysis, we also uncovered several biologically interesting clusters implicated in HCC. We established that HCC transition from a moderate to poor stage involved densely connected clusters that comprised of PCNA, XRCC5, XRCC6, PARP1, PRKDC, and WRN. From our pathway enrichment analyses, it appeared that the HCC moderate stage, unlike the poor stage, is enriched in proteins involved in immune responses, thus suggesting the acquisition of immuno-evasion. Our strategy illustrates how an original oncoproteome could be expanded to one of a larger dynamic range where current technology limitations prevent/limit comprehensive proteome characterization.



KEYWORDS: HCC (hepatocellular carcinoma), proteomics, protein networks, liver cancer, bioinformatics, systems biology

■ INTRODUCTION

HCC is the third most common cause of cancer-related mortality.[4] It is usually distinguished against other forms of cancers in that its etiology is primarily due to tissue damage. Risk factors of HCC include (i) chronic infection via hepatitis B or C virus (HBV or HCV), (ii) germline mutations, (iii) cirrhosis, (iv) alcoholic liver disease, (v) hemochromatosis, and (vi) other liver diseases.[5] Such diverse etiologies imply high variability in the initiation mechanisms leading to HCC. HCC can be categorized histologically into three progressive stages: well, moderately and poorly differentiated. While histological distinguishing aids stratification, HCC differentiation is notoriously heterogeneous—it is difficult to divide borderline cases, histological criteria variability exists, and they do not necessarily correlate well with clinical outcome such as prognosis and survival. Hence, the ability to comprehensively characterize and quantify the changes in protein expression at the molecular level may better distinguish HCC progression, enhance our understanding of cancer pathogenesis, and also yield molecular targets for the treatment of HCC and other cancer types.

To systematically quantify differences between paired human cancerous and the noncancerous HCC tissues, we used Isobaric Tag for Relative and Absolute Quantitation (iTRAQ) in combination with 2D-LC−MS/MS. iTRAQ has gained popularity for its ability to perform concurrent identification and relative quantification of hundreds of proteins for up to 8 biological samples in a single experiment, and has been employed in several

Received: October 27, 2010
Published: March 16, 2011

oncoproteomics studies.[6,7] One major drawback however is data inconsistency (in protein identification) across repeated mass spectrometric runs of the same sample (by the same operator or different laboratories). Although this could be partially resolved by increasing the number of repeated injections,[8,9] it is not always practical.

With the human HCC iTRAQ data set as a reference, we proposed a set of complementary methods that can help to overcome incomplete data coverage and inconsistency, and present functional information by combining iTRAQ data with network and pathway information. The mapping of genomic data onto biological networks is not a novel concept unto itself.[10] In Ramakrishnan et al., it was demonstrated that the use of gene interaction network information can meaningfully expand the repertoire of proteins returned via MS analysis.[10] In microarray screens, genomic expression combined with network analysis can yield important information on how expression variation relates to differences between observed states.[11] As closely connected genes tend to be involved in similar functions, network annotation can complement clusters obtained via fold change analysis.[12] Microarray and deep sequencing methods tended to provide a much larger information pool relative to proteomic platforms. In addition, the general reproducibility is more amendable to statistical analyses. However, they are less able to provide information at the functional level. Alternatively, the smaller information yield and general inconsistency in iTRAQ screens impedes analysis. The former increases false negatives that are not taken into consideration during analysis. As a result many important pathways and components are missed. We show here that recovering shared components or closely associated neighbors can attenuate this problem. The latter case is harder to solve. Because the various protein (data) points are not all accounted for, it is difficult to establish if a protein is indeed differential. For example, if only one out of six samples reports a protein as being differential and there is no reading in the other five, it is harder to establish if the protein is truly important. It is possible to increase the confidence by increasing sample size, but not always feasible. Alternatively, it is not unreasonable to hypothesize that lowly supported differential proteins could become important if they are linked to high confidence differential proteins in the network, indicating that they share similar properties or a dysregulation in any of them could be phenotypically equivalent. A scoring function that takes into account low and high confidence proteins in a cluster can help recover important information that may be lost due to overstringent filtering. Here, we looked to two elements of biological networks that can be applied toward improving MS analysis. These are the clusters/cliques, and the chains (or biological pathways).

Clusters in networks are functionally related components in a protein−protein interaction network (PPIN) as they correspond to protein complexes or tightly interconnected subcomponents in biological pathways.[13,14] A cluster may be a clique, or composed of several overlapping cliques that is, a completely connected sub graph. Many clusters are probably not cliques since the latter is strictly mathematically defined. To find interesting clusters in the mapped network, clique percolation is a useful method for detecting overlapping cliques given a defined core size.[15] Other suitable cluster finding methods include the Girvan-Newman algorithm,[16] CMC,[14] MCODE,[17] and MCL.[18] Because clusters are strongly connected internally, those that contain a high proportion of detected differential proteins can yield novel testable targets. They also allow discovery of closely associated differential proteins. A second advantage of using cluster information is recovery of interesting associations between proteins that are lost due to using a threshold.

A biological pathway, also referred to as a chain, is composed of several biological molecules known to be involved in a specific biological system. This can be metabolic, signaling, etc. Chains are better defined biologically than clusters due to curation and established literature sources. For the same reason, chain information is also relatively scarce. Despite the fact that chains are better annotated, surprisingly, the components within a chain are not always agreed on across various data sources. Chain analysis is also generally not straightforward due to difficulties in extracting information from data repositories such as the Kyoto Encyclopedia of Genes and Genomes (KEGG), or costs incurred from subscribing to commercial databases such as Ingenuity Pathway Analysis (IPA). Furthermore, it is known that despite being better annotated than protein complexes, pathway agreement across various databases such as KEGG and IPA is also generally low.[19] In addition, their overlaps are also low. To improve the coverage of pathway information, we built and utilized an integrated Pathway database, PathwayAPI[19] so that robust data evaluation could be performed.

Combining network cleaning with community finding methods, we uncovered several biologically interesting clusters, one of which comprises of a heavily targeted protein kinase (which is found in most poor patients and is also strongly differentially expressed) and surrounded by an array of DNA repair enzymes, each of which is found in less than half the patients.

The reference HCC data was also comprised of two biological stages (moderate and poor). To better understand how these two stages could be understood in terms of pathway progression, we developed a method of tracing links based on average path lengths from pathways found in both (mod and poor) to poor only.

Our method differs from previous approaches as first, we based our analysis on an integrated gene interaction network that has been evaluated for functional coherence. Second, we performed a similar analysis on an integrated and curated biological pathway database consisting of the widely used KEGG, WikiPathways as well as IPA. Third, by combining both pathway and functional interaction analyses, it is possible to derive a more comprehensive understanding of the experimental data. Finally, it is observed that expansion of the data to incorporate highly connected neighbors is crucial in overcoming data sparseness in MS. Detecting HCC−associated molecular changes could help manage HCC−staging, improve surveillance and also in development of stage-specific therapeutic intervention or biomarker discovery.

MATERIALS AND METHODS

Tissues

Liver tissues were obtained from 12 male patients diagnosed with HCC and suffered from cirrhosis with chronic HBV infection. There was no metastasis at the point of surgery. Tissues collected were grouped according to histology report; 5 had moderately differentiated HCC (mod) and 7 had poorly differentiated HCC (poor). Supplementary Table 1 (Supporting Information) shows the designated identifier and their HCC grade. Paired tissues were obtained from each patient, one from the adjacent nontumor region (normal) and the other from the tumor region of the resected liver. The tissues after resection were immediately snap-frozen in liquid nitrogen and stored at −151 °C until use. Usage of these samples was approved by the National University Hospital Ethics committee.[20]

Tissue Sample Preparation

Human liver tissues were ground into a fine powder in liquid nitrogen and subsequently solubilized in a cocktail of 7 M urea, 2 M thiourea, 4% (w/v) CHAPS, 10 mM Tris supplemented with 1× HALT protease inhibitor cocktail, 50 μg/mL DNase I and 50 μg/mL RNase A. The lysates were then centrifuged at 50 000× *g* for 2 h at 15 °C to remove any insoluble cell debris. The supernatant was stored at −80 °C. All protein estimations were carried out using the Coomassie Plus Protein Assay Reagent kit with minor modifications. Bovine serum albumin provided in the kit was used as the standard.[21]

Quantitative Proteomics using iTRAQ

Protein lysates from either the nontumor or tumor were first precipitated using the 2-D Clean-Up kit. The protein pellets were subsequently resuspended in either dissolution buffer (500 mM triethylammonium bicarbonate and 0.1% (w/v) SDS) for iTRAQ labeling. iTRAQ labeling and processing of the samples were carried out as described by the protocol with minor modifications and using the reagents provided from Applied Biosystems. One-hundred micrograms of protein from each sample was reduced with 50 mM of TCEP at 6 °C for 1 h and subsequently alkylated with 200 mM of methyl methanethiosulfonate (MMTS) for 10 min at room temperature. Each sample was diluted to achieve a final concentration of 0.05% (w/v) SDS prior to trypsinization at 37 °C for 16 h. Following this, each tryptic digest was labeled for 1 h with one of the four isobaric amine-reactive tags. The labeling was carried out at random ensuring that 2 pairs of patient tissues were labeled as follows: Channel 114 (nontumor); Channel 115 (tumor); Channel 116 (nontumor); and Channel 117 (tumor samples). These four iTRAQ-labeled samples were then pooled and passed through a strong cation exchange cartridge as recommended by the manufacturer (Applied Biosystems). This eluate was further desalted using a Sep-Pak cartridge (Millipore), lypholised and reconstituted in appropriate buffers for 2-D LC.[22]

Two-Dimensional Liquid Chromatography Separation of Labeled Peptides

iTRAQ-labeled peptide mixtures was further separated using an Ultimate dual-gradient LC system (Dionex-LC Packings) with a Probot MALDI spotting device. A two-dimensional LC separation was performed as follows: the labeled peptide mixture was first dissolved in 2% (v/v) acetonitrile (ACN) containing 0.05% (v/v) TFA and injected into a 0.3 × 150 mm strong cation-exchange (SCX) column (FUS-15-CP, Poros 10S; Dionex-LC Packings) for the first dimensional separation. The mobile phase A was 5 mM KH2PO4 buffer, pH 3, 5% ACN and mobile phase B 5 mM KH2PO4 buffer, pH 3, 5% ACN + 500 mM KCl respectively. The flow rate was 6 μL/min. A total of 9 fractions were obtained using step gradients of mobile phase B: unbound, 0−5, 5−10, 10−15, 15−20, 20−30, 30−40, 40−50, 50−100% of B. The eluting fractions were captured alternatively onto two 0.3 × 1-mm trap column, washed with 0.05% TFA and followed by gradient elution in a 0.2 × 50-mm reverse-phase column (Monolithic PS-DVB; Dionex-LC Packings). The mobile phase used for this second-dimensional separation was 2% ACN with 0.05% TFA (A) and 80% ACN with 0.04% TFA (B). The gradient elution step was 0−60% B in 15 min at a flow rate of 2.7 μL/min. The LC fractions were mixed directly with MALDI matrix solution (7 mg/mL CHCA and 130 μg/mL ammonium citrate in 75% ACN) at a flow rate of 5.4 μL/min via a 25-nl mixing tee (Upchurch Scientific) before they were spotted onto a 192-well stainless steel MALDI target plate (Applied Biosystems) using a Probot Micro Precision Fraction collector (Dionex-LC Packings), at a speed of 5 s per well. 50 fmol of ACTH (18−39) peptide (m/z = 2465.199) was spiked into each well as internal standard.[23]

Mass Spectrometry Analysis and Database Search

The samples on the MALDI target plates were analyzed using a 4700 Proteomics Analyzer mass spectrometer (AB SCIEX) with MALDI source and TOF/TOF optics. MS/MS analyses were performed using nitrogen at collision energy of 1 kV and a collision gas pressure of 1 × 10^{-6} Torr. The GPS Explorer software Ver. 3.6 (AB SCIEX) was used to create and search files with the MASCOT search engine (version 2.1; Matrix Science) for peptide and protein identifications. The International Protein Index (IPI) human database (Version 3.31) was used for the search and this was restricted to tryptic peptides.One thousand shots were accumulated for each MS spectrum. For MS/MS, 6000 shots were combined for each precursor ion with signal-to-noise (S/N) ratio greater or equal to 100. For precursors with S/N ratio between 50 and 100, 10 000 shots were acquired. The resolution used to select the parent ion was 200. No smoothing was applied before peak detection for both MS and MS/MS, and the peaks were deisotoped. For MS/MS, only the peaks from 60 to 20 Da below each precursor mass, and with S/N greater than or equal to 10 were selected. Peak density was limited to 30 peaks per 200 Da, and the maximum number of peaks was set to 125. Cysteine methanethiolation, N-terminal iTRAQ labeling, and iTRAQ labeled-lysine were selected as fixed modifications while methionine oxidation was considered as a variable modification. One missed cleavage was allowed. Precursor error tolerance was set to 100 ppm while MS/MS fragment error tolerance was set to 0.4 Da. Maximum peptide rank was set to 2. iTRAQ quantification was performed using the GPS Explorer software and normalized among samples.[23] For MS/MS, only the peaks from 50 to 20 Da below each precursor mass, and the minimum S/N filter was designated at 10. The mass exclusion tolerance was 3 Da around 115.5 m/z. Peak density was limited to 50 peaks per 200 Da, and the maximum number of peaks was set to 80. iTRAQ ratios were calculated based on the areas of the iTRAQ reporter fragment peaks (114, 115, 116 and 117), and the ratios calculation included only peptides identified with C.I. % above cutoff thresholds as described below. The average iTRAQ ratio and standard deviation (S.D.) were determined using the GPS Explorer software (Ver. 3.6). In order to verify the identified proteins and degree of quantification, the same set of spectra were run on a different database se\arch algorithm, the Paragon algorithm in Protein Pilot 4 software (AB SCIEX). Autobias correction was applied and the Unused ProtScore was >1.3 (C.I.% > 95%).

Establishment of Differential Candidates

Proteins identified and quantified by iTRAQ (Supplementary Tables 3 and 4, Supporting Information) formed the basis of our seed selection. For each patient, a ratio was obtained for each protein by self-comparison to nontumorigenic liver tissue. The definition of a differential protein required passing two levels. First, the protein should meet the expression threshold of 1.25 and 0.8 (reciprocal of 1.25) for overexpressed and underexpressed proteins respectively. We chose a slightly lower cutoff

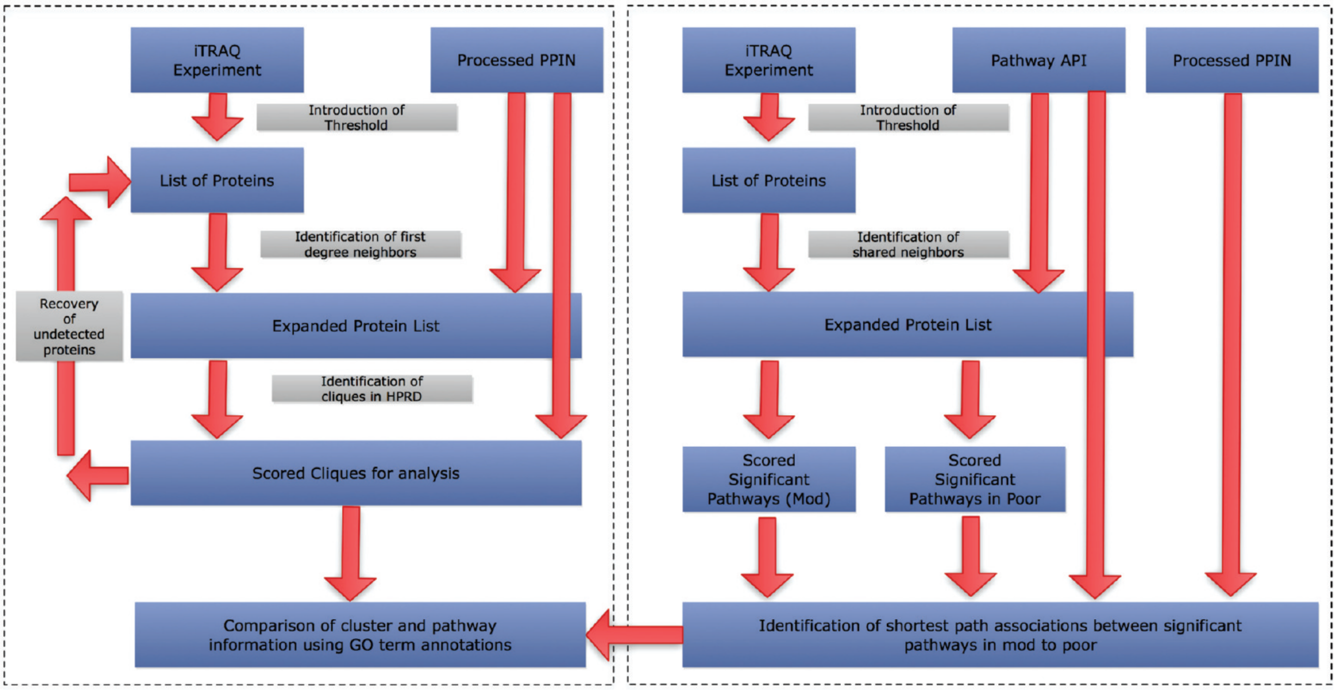

iTRAQ Experiment
Processed PPIN
Introduction of Threshold
List of Proteins
Identification of first degree neighbors
Recovery of undetected proteins
Expanded Protein List
Identification of cliques in HPRD
Scored Cliques for analysis
Comparison of cluster and pathway information using GO term annotations
iTRAQ Experiment
Pathway API
Processed PPIN
Introduction of Threshold
List of Proteins
Identification of shared neighbors
Expanded Protein List
Scored Significant Pathways (Mod)
Scored Significant Pathways in Poor
Identification of shortest path associations between significant pathways in mod to poor

Figure 1. Schematic of integrated analysis pipeline. The analysis pipeline can be broadly divided into two major components. On the left block, MS protein list is first filtered for seed proteins. The expansion step is done in relation to the cleaned protein−protein interaction network or PPIN (for information on how this is done, please refer to Materials and Methods). Clique analysis is then performed to obtain tightly connected clusters. The clusters are then scored and ranked. For pathway analysis, we used an integrated pathway database developed in-house (Pathway API). Similarly, the MS protein list is expanded by locating shared neighbors in the list of pathways. Pathways significant in mod and poor respectively were identified. The shortest path distances between the mod to poor pathways were then identified on an integrated gene network combining both pathway and PPIN information. Finally, the set of significant pathways can be compared to the significant clusters to identify overlaps in function and annotation.

as this would help bolster sensitivity. Traditionally, the significant cutoff threshold proposed was based on the standard deviation (S.D.) of all the ratios of the respective labeled peptides and this would theoretically be 1.3/0.77 (based on 1 ± 2 S.D.).[23] Second, the other requirement is that the protein has to be consistently detected in at least 3/5 of patients for the moderately differentiated samples and 5/7 patients for the poorly differentiated tumors. Of note, there were cases where proteins met the expression threshold but did not meet the second requirement and thus were removed (Supplementary Figure 1, Supporting Information). Taken together, proteins that meet the requirements in both filters (expression change and patient support) were maintained as seeds for neighbors and cluster analysis.

Protein−Protein Interaction Network (PPIN) Cleaning

An integrated PPIN was built comprising of data from HPRD,[24] BioGRID,[25] INTACT[26,27] and DIP[28−30] as well as data from literature.[31,32] This network was then filtered using CMC/FilterNadd, and the top 90% of highest scoring edges kept. The resultant combined network (pathway and PPIN information) displayed the properties of a typical PPIN (data not shown).

To evaluate the effect of network cleaning on the integrated network, we measured Gene Ontology (GO) Biological Process, Cellular Component and Molecular Function term coherence for every edge pair in the cleaned and uncleaned integrated network. Edge coherence is calculated by counting the number of shared GO terms in each category for every GO-annotated edge pair divided by the total number of considered edges.

The cleaned integrated PPIN was then used as the reference protein interaction network for cluster identification.

Identification of Functional Clusters as Overlapping Cliques

Figure 1 shows the overview of the integrated bioinformatics pipeline. To identify relevant subnets, we used differential proteins obtained above as seeds and mapped them onto the cleaned PPIN. They were then expanded to include their first-degree neighbors. Identification of overlapping clusters was performed using Palla's Clique Percolation Method.[15] In this paper, we will discuss the largest clusters obtained with $k = 5-8$. Smaller clusters ($k = 3-4$) are available upon request. The clusters were then scored and ranked by the following method.

$$S = \frac{\sum_{i=1}^{n} Ei}{n} \tag{1}$$

Where S is the calculated score and E is the expression value for a detected protein (if protein is underexpressed, then the reciprocal score is used).

Identification of Enriched Pathways and Tracking Pathway Associations

To find enriched pathways, we mapped the seeds onto our in-house developed database PathwayAPI,[19] which comprises of information extracted from KEGG,[33] WikiPathways[34,35] and IPA.[36] The online repository is available at http://pathwayapi.com/. There are 4268 nodes and 35307 edges, corresponding to

Table 1. Overlaps between Mascot and Paragon Protein Hits for All Samples

patient identifier	intersection (common to Mascot and ProteinPilot)	Mascot only	Protein Pilot only
199	197	28	63
131	246	16	120
215	499	33	250
196	247	16	119
200	198	28	62
207	498	33	251
126	531	30	287
155	531	30	287
203	586	41	180
120	586	41	180
157	696	34	376
187	697	36	375

544 pathways. Significance ($P \leq 0.05$) was calculated using the hypergeometric method in eq 2. For the hypergeometric test, shared neighbors that were recovered were also considered in the set of proteins in the poor or mod list. This helped to increase the significance of those pathways that have a higher preponderance of indirectly linked genes. The pathways were then scored and ranked by using eq 3.

$$h(x; N, n, k) = \frac{\binom{N-k}{N-x}\binom{k}{x}}{\binom{N}{n}} \tag{2}$$

Where N is the number of proteins in the pathway network, x is the number of proteins in the poor or mod list, m is the number of proteins in current pathway, and k is the number of same proteins between the poor or mod gene list and the pathway.

$$S = \frac{\sum_{i=1}^{n} Ei}{n} \tag{3}$$

where S is the calculated score and E is the expression value for a detected protein (if protein is underexpressed, then the reciprocal score is used).

We tracked the progression of moderate into poor stage by first taking into account all significant pathways that are common to both mod and poor stage. We then considered pathways that are found only in mod, and in poor stage and calculated the minimum and average distance needed to go from one pathway to another using the Floyd-Warshall algorithm.[37] Distance d, is defined as the number of steps (in terms of biological molecules) needed to reach one pathway from another. A distance of 0 implies shared common node(s). Average distance is the average number of steps from one pathway to another. That is, for each node N_x in first pathway, N_y in second pathway, find the minimum distance from N_x to N_y, then averaged over all combinations of N_x, N_y.

RESULTS AND DISCUSSION

Result Correlation between Mascot and Paragon

Combining multiple search engines has the dual advantages of checking reproducibility and providing complementary data from the same raw results.[38] To leverage on the different search algorithms we performed a search using Paragon on ProteinPilot and Mascot.

Using a cutoff of 5% protein "local" FDR on Paragon, we found that the agreement, in terms of common protein identifications, between Mascot and Paragon was good (Table 1). Interestingly, it also appeared that Paragon consistently returned more protein hits (Table 1). That many, if not most of the Mascot hits were also found in the Paragon list showed that despite differences in database search algorithm methods, proteins, especially those of high confidence identifications (see below) were repeatedly identified. This is plausible in that multiple search engines adds to the confidence of original protein identifications by working through different algorithms and assigning previously unassigned high quality MS/MS spectrum to peptides.[39] To further determine the degree of similarity in both Mascot and Paragon, we performed correlation analysis on the intersection for proteins from all 12 patients of their list Rank and list Ratio scores (Supplementary Figure 2, Supporting Information). Our results indicated that for proteins agreed on by both Mascot and Paragon, there is a clear positive correlation for both ranks and ratio.

As Paragon was returning many more protein hits, it is of interest to establish if these additional Paragon unique proteins were of lower confidence. We performed a one-sided Wilcoxon Ranked Sum test on Paragon-only proteins and compared it to the distribution of ranks in Paragon. It was found that the distribution of ranks in the Paragon-only list was always significantly greater than that in the full data set. In 11 of 12 patient, the values were close to zero ($<2.2 \times 10^{-16}$) with only one patient giving a value of 0.00064. This confirms that the Paragon-only list is of lower confidence than the intersection set. Hence, only proteins that are supported by both Mascot and Paragon were retained for establishment of differential candidates.

Expansion by First-Degree Neighbors Improves Coverage Significantly

From the original data set of approximately 500 proteins detected from iTRAQ 2D LC−MS/MS we were able to expand to approximately 1500 proteins, an additional 1000 proteins, through the use of our bioinformatics pipeline (Figure 2). Proteins classified as "differential" given the expression ratio threshold and patient count criteria are used as "seeds" for first-degree neighbor expansion. A first-degree neighbor is a protein that has a direct connection or interaction with a seed. Second-degree neighbors, that is, proteins that are linked to seeds via a single intermediary, were also considered but due to the wide dispersal of the proteins in the network, resulted in covering most of the reference network. Hence, only first degree clique analysis results were retained.

A clique is a fully connected graph, and its size is denoted by the constant k. For instance, all cliques consisting of 3 proteins is referred as $k = 3$. To show that expansion of seeds is necessary, we mapped nonextended seeds onto the reference network, and retrieved only 31 edges. In the case of cliques obtained via nonextension, only four $k = 3$ cliques were returned. These are (FGA, FGB, FGG), (HSPA8, YWHAG, YQHAQ), (TGM2, TUBB, ALDOA) and (YQHAB, TUBB, YWHAG). Mapping first-degree neighbors and the seeds returned 3378 edges and returned a much larger number of clusters, approximately 120

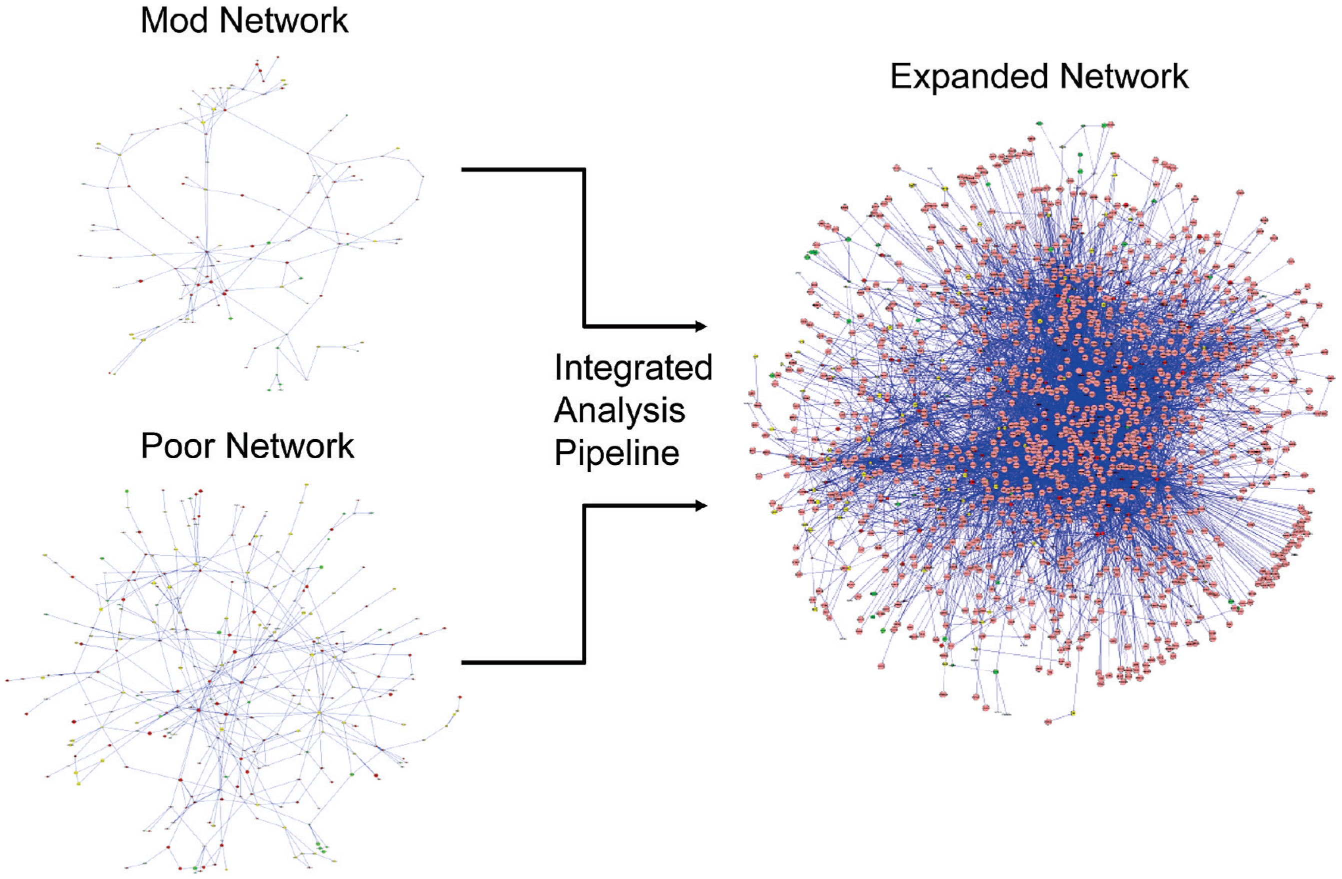


Figure 2. Expansion of candidate proteins from mod and poor. MS-detected proteins from mod and poor stages show smaller, sparser networks and there are too few targets for clustering analysis. Note that mod is a subset of poor, hence only proteins from the poor stage were expanded. Expansion of first order candidates helps to create a much larger set of candidates for further functional analysis.

unscored clusters from $k = 3$ to 8 (Figure 2). Score-able clusters (with at least one differential protein) provide some indication of function within HCC. We also recovered many other clusters that did not contain differential proteins but which appeared to be interesting and relevant to cancer (data not shown).

Cluster-Based Analysis Reveals Functional Relationships between Identified-Differential Proteins

Uncovering communities requires a well-annotated network with high coverage. Currently, PPINs are extremely noisy and are also incomplete. Agreement across various PPIN databases is also extremely low.[40] Combining PPINs can improve coverage but may give rise to compounded errors. Functional evaluation of edges in a PPIN is therefore an important first step. We used the algorithm CMC[14] to reduce the PPIN to only high confidence edges based on CD distance (Czekanowski-Dice Distance). We also introduced a GO term coherence cutoff as a second filter.

Although in the cleaning proess, about half of edges were lost, it corresponded to a 3−4 fold enrichment for GO term coherence. While integration of several PPINs improved coverage, data quality is also important. We demonstrated here that by coupling integration to our data cleaning algorithm, there is an appreciable improvement (3−5 times) in data quality (Supplementary Table 2, Supporting Information).

For brevity, we limited our discussion to the larger clusters. That is, overlapping cliques with a core comprised of at least 5−8 proteins. Some smaller clusters might be meaningful and could be isolated through scoring function, but these were few and usually were subsets of a larger cluster. Moreover, larger clusters provided an initial list of higher confidence proteins that could aid in understanding the biological significance of the data. To build the clusters, all poor stage proteins found to be differential, and their first-degree neighbors, were used (mod differential proteins are a subset of the poor differential proteins; Figure 3). The top-ranked clusters were found associated with expected functions such as stress, DNA damage, apoptosis and differentiation (Figure 4A).

One interesting cluster is the PRKDC cluster which comprised of six members, PRKDC, XRCC6, PCNA, XRCC5, WRN and PARP1 (Figure 4B, top left). XRCC5/6, and PCNA and PARP1 are repair factors, while WRN is a nuclear protein that could be involved in maintaining genomic stability. PRKDC is a protein kinase that is capable of targeting p53, and found to be differential in a majority of poor patients (5 out of 7). It was interesting to note that the repair factors were all low count, between 1 and 2 patients each (Supplementary Tables 3 and 4, Supporting Information). It might be that mutations of these repair factors are crucial in affecting the functionality of this group of functionally close knitted proteins. In particular, these repair factors all appear to interact with PRKDC. Checking the cluster against a reference microarray database (Cancer Gene Expression Database[41]) indicated that the low count repair factors have been previously reported to be differentially expressed in earlier screens. This lends further support that the

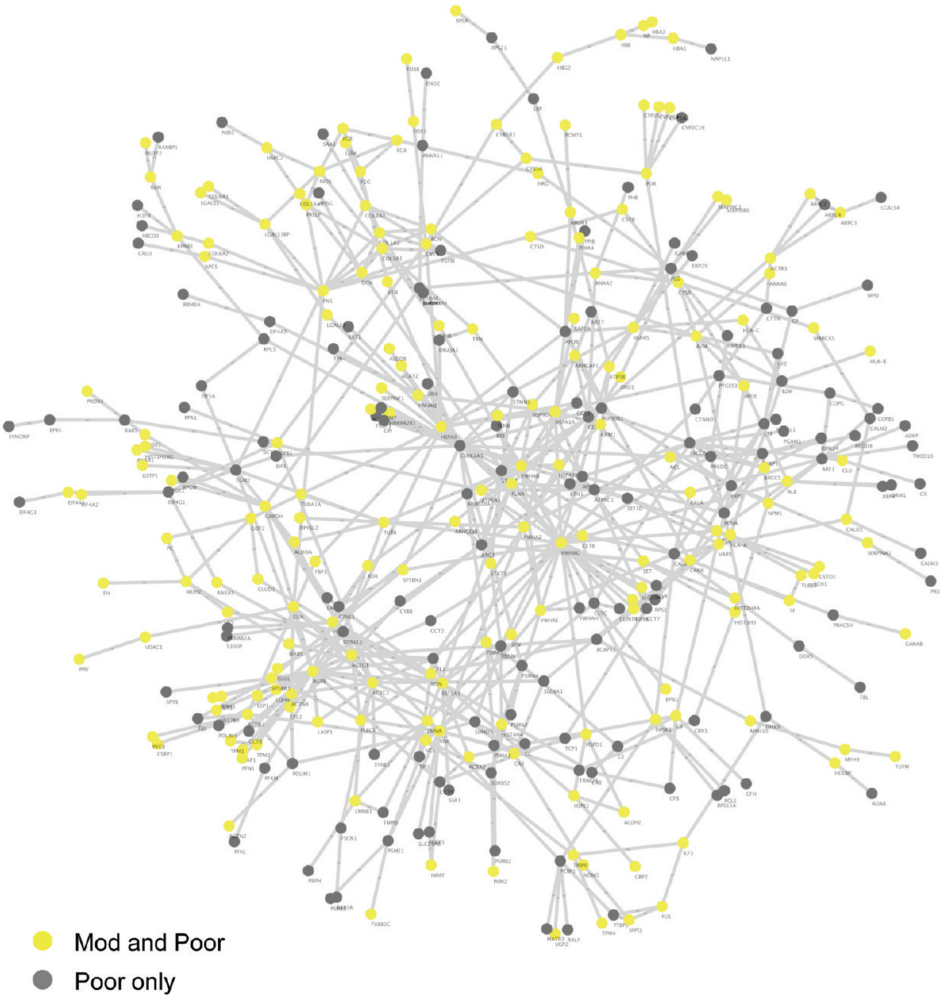


Figure 3. Network of differential candidate proteins in mod and poor. Network of poor proteins shows the emergence of a giant connected-component (gray and yellow nodes) by overlaying differential proteins onto HPRD. Since mods are a near perfect subset of poor we overlaid the former onto this network (yellow nodes) and found that while there appeared to be some pockets where mod proteins tended to aggregate, we found that overall, poor and mod proteins are interspersed.

cluster identification can identify meaningful biological relationships and establish a functional context for our data.

Only XRCC5 was found to be differentially expressed by comparing the PRKDC cluster to the patient data in the mod stage. Taking the ratio between the scores in this cluster for both mod and poor revealed this cluster to have the greatest score jump (that is, in poor, the score in poor is approximately 10 times greater than the mod stage). This cluster therefore appears to be important in the transition between mod and poor stage (Table 4). The exact functional significance and its possible role in triggering liver cancer progression to the poor stage require further wet-lab validation.

We also found TP53 in several high ranking clusters. TP53 is well established in cancer but was not detected in the iTRAQ screen. p53 is a small protein of low abundance and detection of low abundant proteins by MS-based proteomics have limited success. To this end, innovative protocols have emerged to improve the detection of low abundant proteins such as extensive fraction by MuDPIT[42] and targeted proteomics by MRM.[43] However, extensive time and resources are required before detecting such low abundant proteins is possible. As an alternative, our pipeline offers the prospect of detecting such proteins.

Recovery of Clique Proteins from MS Spectra

Based on the cliques isolated for analysis, we found a total of 160 unique proteins of interest. These were formed from 14 participating seed proteins. We combined Mascot- and Paragon-only proteins into an excess list and identified 23 of these in the cliques. However, this subset of 23 proteins did not possess any enrichment for ranks or ratio scores (Wilcoxon ranked sum test, $P = 0.512$).

Some MS spectra may match to a particular protein, but because their scores were below the defined cutoff threshold they may not be reported initially in the first round of data analysis. Our bioinformatics pipeline by improving the coverage of protein communities could have highlighted these proteins. In an iterative approach, we are empowered now to return to the original MS/MS spectra to look for evidence to support the existence of these recovered clique proteins. There are a few reasons for not reporting these proteins in the initial round of data analysis: (1) they did not satisfy the two unique peptides requirement, that is protein identification by a single peptide, (2) they were identified by short peptides, and (3) they were not consistently found in the patients. To this end, we selected several proteins (ACTR2, CDC42, GNB2L1, KIF5B, PPP2R1A,

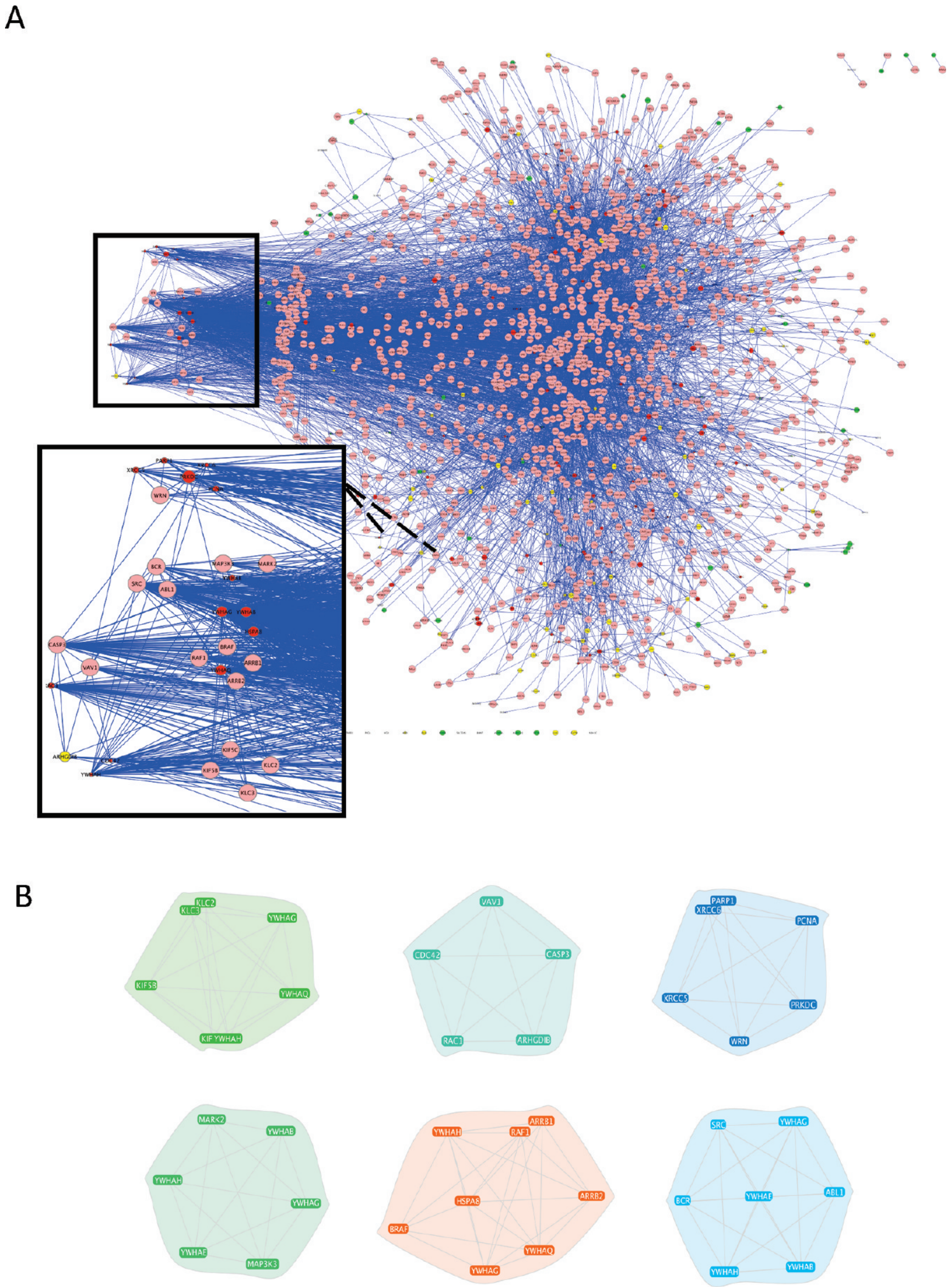


Figure 4. Top 6 Clusters. (A) In the context of the network, from the top down view, the top six clusters (yellow nodes) are shown to be related to each other. Highlighted in red are the links of clusters to each other and to the network. (B) Clusters are the top six ranked clusters obtained in the HCC poor stage. Each cluster consists of both MS-detected and undetected proteins. The GO term analysis reveals that many of the clusters are involved in cancer causing events such as apoptosis and cell growth. Calculation of scores in both the mod and poor stage, and comparing their ratio allows identification of the clusters most involved in the transition event.

PKACA and TOP1) from the top 34 cliques/chains and not detected by Paragon, and manually examined their GPS and Mascot search results and also their MS/MS-to-peptide assignments to verify the legitimacy of the predictions. Assessment of MS/MS spectra of their top ranked peptides revealed accurate *y* and *b* ion assignments and were of good quality (Expectation value <0.05) supporting and verifying the *in silico* expansion (Supplementary Figure 3); with the exception of KIF5B whose top scoring peptide had an Expectation value of >0.05. PKACA could not be found in any of the 12 patients, possibly due to the Mascot peptide significance *P*-value >0.05. Proteins of low copy numbers and high cellular turnover such as transcription factors

Table 2. Pathways Unique to Mod Stage

Mod only pathway
Antigen processing and presentation - *Homo sapiens* (human)
Cell adhesion molecules (CAMs) - *Homo sapiens* (human)
FGF Signaling
Id Signaling Pathway
IL-2 Signaling
IL-4 Signaling
Insulin Recpetor Signaling
Interferon type I
mTOR signaling pathway - *Homo sapiens* (human)
Natural killer cell mediated cytotoxicity - *Homo sapiens* (human)
Neurotrophin/TRK Signaling
Peptidoglycan biosynthesis - *Homo sapiens* (human)
T-cell receptor Signaling
Triacylglyceride Synthesis
Vitamin B6 metabolism - *Homo sapiens* (human)

and protein kinases however still cannot be found through retrospective assessment of original MS/MS data and highlighted the utility of our bioinformatics pipeline in circumventing such challenging MS tasks.

Chained-based analysis supports cluster-based analysis but reveals many more functional relationships

Cluster based analysis showed that the DNA damage cluster was most pertinent in the poor stage relative to the mod stage, exhibiting the largest jump in score ratio. Ranking the top scoring pathways in the poor stage showed that the top ranking pathways included mostly detoxification and metabolic pathways. It would appear therefore chain and cluster−based analysis yielded different results. A closer check against the top clusters GO terms against significant pathways in poor revealed that it was not so. The top PRKDC cluster was mapped to stress responses ($P = 3.09 \times 10^{-11}$). Significant pathways in poor include Oxidative stress, FAS pathway and Stress induction of HSP regulation, and ER stress responses. The PEBP1 cluster, also highly ranked in our list, comprising members PEBP1, AKT1, ESR1, HSP90AA1, CHUK,HSP90AB1, HSPA8,MAP3K7, CASP8, MAP3K8, SRC, TBK1, MAP3K14, IKBKB, IKBKG and NFKB2 corresponded to the term “nuclear receptor”. Interestingly, this set of proteins is also implicated in stress responses and immune response to infection. This is of particular relevance since the liver cancer subtype being considered is hepatitis induced. We also uncovered apoptosis-associated clusters, which matched to apoptotic pathways in the poor stage (RAF1,MAP3K5,YWHAZ,CDC25A,YWHAE and ABL1, BCR, YWHAB, YWHAH, SRC, YWHAE and YWHAG). Proteins involved with focal adhesions, cell−cell junctions and adherens which were strongly represented in the PARD cluster (CDC42, RAC1, YWHAH, PRKCI, YWHAZ, PARD3, PRKCG, PARD6A and PARD6B) were also found in the poor list. The good correspondence between top cluster GO annotation and significant pathways suggested that the results were congruent. However, the chains revealed that cluster analysis has several limitations in understanding the functional and biological differences between the two stages. Because cluster analysis does not consider the functional biological units a priori, it is limited by the size cutoff for the clusters. Moreover, if many of the relationships do not exist in at least a clique of size 3, it would not be detectable. Chain analysis, however, was not as useful in yielding novel target proteins, or unveiling new interactions, as they are already well studied to begin with.

Chain-based Analysis Reveals Cancer Progression Mainly Occurs in Mod Stage while Poor Stage Exhibits Most Damage-Specific Effects

HCC progression can be categorized into three stages: (i) well-differentiated, where HCC cells resemble hepatocytes, are hypovascularised and considered the early form of HCC, (ii) poorly differentiated, which is characterized by diffused growth suggesting advanced HCC progression, and (iii) moderate HCC, which is an intermediate stage between the two. We compared pathways that are shared in both mod and poor, and ranked them by their scores to obtain a list of pathways most affected in the mod stage, and poor stage respectively. For pathways that are significantly common ($P \leq 0.05$) in both mod and poor, dysregulation of the cytoskeleton, cell−cell interactions and immune responses are strongly represented. In the mod stage only pathway list (Table 2), there appears to be strong enrichment for immune-specific responses. These include pathways involved in antigen presentation and processing, T-cell receptor signaling, and proteins involved with InterLeukin-2/4 and interferon pathways. Differential proteins found in these pathways may be able to provide suitable marker candidates for early detection for patients that have yet to develop to mod stage. HCC is a complex, multistep process that commonly develops against a cirrhotic background (largest risk factor) that arises from chronic liver inflammation. HBV and HCV infection are known etiological factors to HCC, accounting for approximately 80% of all HCC cases.[21] Patients who are seropositive with chronic HBV infection are 5−15-fold more likely to develop HCC.[44] HBV and HCV infections induce liver inflammation by continuous cycle of hepatocyte death and their regeneration. This leads to the development of chronic hepatitis, liver fibrosis and cirrhosis, eventually leading to HCC. In particular HCV RNA and core proteins have been implicated in T-cell activation[45] and evade immune-mediated cell death by interactions with Interferon-α.[46] HCV core proteins have also been shown to interact with MAPK signaling to modulate cell proliferation. Our results that inflammation-associated pathways were overrepresented were in line with the conjecture that chronic inflammation drives hepatocytes to a malignant phenotype into the mod stage. Major signaling pathways activated during HCC progression include Insulin/IGF-1/IRS-1/MAPK and Wnt/Frizzled/beta-catenin signaling.[47]

Our analysis found the enrichment of MAPK and Wnt signaling pathways in both mod and poor, confirming evidence that aberrant signaling of these two important signaling cascades help shape cancer transformation. Inflammation induces oxidative stress and the latter is another important factor contributing toward HCC. We did not find oxidative stress in the top 20; however, it ranked 67 ($P = 1.27 \times 10^{-1}0$) in mod and 87 in poor ($P = 2.20 \times 10^{-13}$; data not shown). Signaling pathways such as mTOR, insulin and MAPK are known to contribute to cancer metabolic transformation. In poor only pathways, there was enrichment in metabolic pathways (Table 3). Histologically, HCC is classified as poor when specimens are noted to be highly vascularised and suggest advanced HCC progression. To the extent that several metabolic processes were over-represented, we reasoned that the observed perturbation in metabolic pathways represented a gross effect on the transformed liver cells beyond a “threshold of no return”. In addition, they may confer an advantage that promotes the survivability of tumor cells.[48] This metabolic transformation creates a phenotype for sustained tumor growth and survival, cell-death signals resistance as characterized by the histological features of poorly differentiated HCC. Increased amino acid synthesis, lipid metabolism and synthesis were found enriched in poor. In conjunction with perturbations in G1 to S cell cycle control (and G2M DNA damage checkpoint regulation; Table 3), these observations suggest that amino acids and lipids synthesis are modulated for the actively dividing tumor cells. Interestingly, eicosanoids are known to promote secretion of angiogenic factors. Angiogenesis and vasculogenesis implement the formation of the vascular network characteristic of poorly differentiated HCC. Eicosanoids such as prostaglandins and leukotrienes modulate angiogenesis at different levels[49] and our data are congruent with this observation.

Chain Distances in Mod and Poor Reveals Key Roles in IL-2 Signaling and Monoterpenoid Biosynthesis, Respectively

In mod, the integrated network revealed that IL-2 signaling pathway appeared closely associated with many growth-signaling pathways such as TGF-β, PDGF, EGF and hepatocyte growth factor signaling. It is also closely associated to cancer associated signaling pathways such as Jak-Stat, as well as the stress-signaling pathway, SAPK/JNK. This suggests that the mod specific IL-2 pathway is important, and possibly quite involved in cancer progression. In poor, it appears that many of the metabolic process such as arachidonic acid, linoleic acid, as well as antitoxicity processes such as drug metabolism (CYP-P450), tetrachloroethane degradation, and styrene degradation are closely associated with the monoterpenoid synthesis pathway. The latter does not appear to be significantly involved in cancer. However, this pathway, and its constituent proteins may be closely related to many of the damage-associated events observed in poor. We also observed a close connection between p53 signaling (significantly present in both mod $P = 4.32 \times 10^{-5}$ and poor $P = 9.24 \times 10^{-8}$) and the Cell Cycle-G2M DNA Damage Checkpoint Regulation, which was significantly enriched in the poor stage ($P = 0.017$).

Table 3. Pathways Unique to Poor Stage

poor only pathway
Alkaloid biosynthesis II - *Homo sapiens* (human)
C21-Steroid hormone metabolism - *Homo sapiens* (human)
Cell Cycle-G2M DNA Damage Checkpoint Regulation
D-Arginine and D-ornithine metabolism - *Homo sapiens* (human)
Eicosanoid Synthesis
Ether lipid metabolism - *Homo sapiens* (human)
G1 to S cell cycle control
Glucocorticoid and Mineralcorticoid Metabolism
Glycerophospholipid metabolism - *Homo sapiens* (human)
Glycogen Metabolism
Jak-STAT signaling pathway - *Homo sapiens* (human)
Monoterpenoid biosynthesis - *Homo sapiens* (human)
Nuclear Receptors
Pantothenate and CoA biosynthesis - *Homo sapiens* (human)
Riboflavin metabolism - *Homo sapiens* (human)
Selenium metabolism/Selenoproteins
Steroid Biosynthesis

CONCLUSIONS

The integration of networks and pathways with proteomic data generated from iTRAQ 2DLC-MS/MS enhanced our understanding of the functional relationships of proteome changes during HCC progression. Using our developed pipeline on HCC samples, we were able to expand the proteomic data to recover common neighbors. This in turn made it possible to expand the set of dysregulated biological pathways. By applying both cliques and chains analyses our results suggest that HCC late stage is characterized by heavy metabolic defects which may be related to the large scale tissue damage characteristic of HCC. It also implies that intervention at the moderate stage is important in preventing further irreversible damage. This offers the opportunity to better characterize tumor proteomes of a small sample set and can be used for informed clinical decision-making for individual cases.

We also examined the possibility of integrating the outputs of several database search algorithms in enhancing the protein list for analysis. Although we found that the protein set in Mascot was largely covered by the Paragon, the large excess list in Paragon turned out to be of lower confidence. Given that the overlaps between Mascot and Paragon were strongly linearly correlated, it indicated that despite different approaches, they generally give rise to similar results. In future work, it may be worth exploring combining the results of even more database search algorithms to improve the confidence level of the protein set despite the time-consuming searches.

ASSOCIATED CONTENT

Supporting Information

Supplementary Figure S1. Histogram of Frequency Distribution of Patient Counts. Support for proteins given patient size is

Table 4. Clusters with Highest Score Jump Ratio[a]

clique	p_C	m_C	p_S	m_S	ratio	members
$K = 6\_0$	5	1	3.74	0.38	9.7	PRKDC XRCC6 XRCC5 WRN PARP1 PCNA
$K = 5\_1$	5	1	2.8	0.28	9.7	PRKDC TP53 XRCC6 XRCC5 WRN NCOA6 PARP1 PCNA
$K = 5\_3$	1	1	2.34	1.35	1.72	CHUK IKBKB MAP3K14 MAP3K7 PEBP1
$K = 5\_5$	2	2	2.17	2.03	1.06	YWHAE YWHAZ CDC25A RAF1MAP3K5
$K = 6\_5$	4	3	2.16	1.09	1.97	ABL1 BCR CDKN1B YWHAB YWHAE YWHAG YWHAH SRC
$K = 5\_1$	1	1	1.95	1.13	1.72	RAF1MAP2K1 PRKCZ PRKCD PEBP1MAPK1
$K = 5\_3$	1	1	1.59	1.66	0.96	YWHAZ EGFR KRT18 CBL PRKCA
$K = 5\_4$	3	1	1.4	1.03	1.34	YWHAH YWHAZ CDC42 PARD6A PRKCI PARD3 PRKCG PARD6B
$K = 5\_2$	1	1	1.1	0.28	3.82	TP53 PIAS4 SMAD3 SMAD2 KPNB1

[a] p_C, poor count; m_C, mod count; p_S, poor score; m_S, mod score.

small with most proteins only having support from one to two patients. Therefore, even if a protein is defined as "differential" given a large change, the lack of patient support may cause it to be filtered. Dotted lines correspond to the minimum number of patients we kept for each of HCC phases. Supplementary Figure S2. Ranks and ratio correlation analysis of proteins detected by Mascot and Paragon. The intersection for proteins agreed on by both Mascot and Paragon from all 12 patients were analyzed for their (A) list Rank and (B) list Ratio scores. Supplementary Figure S3. Mascot and MS/MS fragmentation information of clique recovered proteins. High quality MS/MS spectra and good Mascot information of Paragon-unique and clique recovered proteins retrospectively verified our *in silico* protein community expansion. While the Expectation value of KIF5B (D) is insignificant, its MS/MS spectrum demonstrated reasonably good ion assignment. Supplementary Table 1. Patient HCC grade and identifier. Supplementary Table 2. Edge GO term Coherence for cleaned and uncleaned integrated PPIN. Supplementary Table 3. List of Proteins detected from patients with moderate HCC. Supplementary Table 4. List of Proteins detected from patients with poor HCC. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Wong Lim Soon, PhD, School of Computing and Department of Pathology, Yong Loo Lin School of Medicine, National University of Singapore, Building COM1, 13 Computing Drive, Singapore 117417. E-mail: WongLS@Comp.nus.edu.sg. Tel: +65-6516-2902. Fax: +65-6779-7465.

Author Contributions

[†]These authors contributed equally to this work

ACKNOWLEDGMENT

The iTRAQ work was supported by Singapore Cancer Syndicate. L.S.W. and D.D. are supported by Singapore National Research Foundation grant NRF-G-CRP-2007-04-082(d). W. Goh is supported by a Singapore Agency for Science, Technology & Research (A*STAR) grant SERC 072 101 0016 and a Wellcome Trust Scholarship (83701/Z/07/Z). J.J. is supported by Singapore Ministry of Education Tier-2 grant MOE2009-T2-2-004. R.M.Z. and Y. H.L. were receipients of the NUS postgraduate scholarship. We are grateful to Lim Teck Kwang for this technical expertise on LC and data management.

REFERENCES

(1) Cox, J.; Mann, M. Is proteomics the new genomics? *Cell* **2007**, *130* (3), 395–8

(2) Mann, M.; Kelleher, N. L. Precision proteomics: the case for high resolution and high mass accuracy. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105* (47), 18132–8.

(3) Hanash, S.; Taguchi, A. The grand challenge to decipher the cancer proteome. *Nat. Rev. Cancer* **2010**, *10* (9), 652–60.

(4) El-Serag, H. B. Hepatocellular carcinoma: recent trends in the United States. *Gastroenterology* **2004**, *127* (5 Suppl 1), S27–34.

(5) Villanueva, A.; Minguez, B.; Forner, A.; Reig, M.; Llovet, J. M. Hepatocellular carcinoma: novel molecular approaches for diagnosis, prognosis, and therapy. *Annu. Rev. Med.* **2010**, *61*, 317–28.

(6) Ralhan, R.; Desouza, L. V.; Matta, A.; Chandra Tripathi, S.; Ghanny, S.; Datta Gupta, S.; Bahadur, S.; Siu, K. W. Discovery and verification of head-and-neck cancer biomarkers by differential protein expression analysis using iTRAQ labeling, multidimensional liquid chromatography, and tandem mass spectrometry. *Mol. Cell. Proteomics* **2008**, *7* (6), 1162–73.

(7) Sutton, C. W.; Rustogi, N.; Gurkan, C.; Scally, A.; Loizidou, M. A.; Hadjisavvas, A.; Kyriacou, K. Quantitative proteomic profiling of matched normal and tumor breast tissues. *J. Proteome Res.* **2010**, *9* (8), 3891–902.

(8) Gan, C. S.; Chong, P. K.; Pham, T. K.; Wright, P. C. Technical, experimental, and biological variations in isobaric tags for relative and absolute quantitation (iTRAQ). *J. Proteome Res.* **2007**, *6* (2), 821–7.

(9) Gramolini, A. O.; Kislinger, T.; Alikhani-Koopaei, R.; Fong, V.; Thompson, N. J.; Isserlin, R.; Sharma, P.; Oudit, G. Y.; Trivieri, M. G.; Fagan, A.; Kannan, A.; Higgins, D. G.; Huedig, H.; Hess, G.; Arab, S.; Seidman, J. G.; Seidman, C. E.; Frey, B.; Perry, M.; Backx, P. H.; Liu, P. P.; MacLennan, D. H.; Emili, A. Comparative proteomics profiling of a phospholamban mutant mouse model of dilated cardiomyopathy reveals progressive intracellular stress responses. *Mol. Cell. Proteomics* **2008**, *7* (3), 519–33.

(10) Ramakrishnan, S. R.; Vogel, C.; Kwon, T.; Penalva, L. O.; Marcotte, E. M.; Miranker, D. P. Mining gene functional networks to improve mass-spectrometry-based protein identification. *Bioinformatics* **2009**, *25* (22), 2955–61.

(11) Sivachenko, A. Y.; Yuryev, A.; Daraselia, N.; Mazo, I. Molecular networks in microarray analysis. *J. Bioinform. Comput. Biol.* **2007**, *5* (2B), 429–56.

(12) Chua, H. N.; Sung, W. K.; Wong, L. Using indirect protein interactions for the prediction of Gene Ontology functions. *BMC Bioinform.* **2007**, *8* (Suppl 4), S8.

(13) Chua, H. N.; Ning, K.; Sung, W. K.; Leong, H. W.; Wong, L. Using indirect protein-protein interactions for protein complex predication. *Comput. Syst. Bioinform. Conf.* **2007**, *6*, 97–109.

(14) Liu, G.; Wong, L.; Chua, H. N. Complex discovery from weighted PPI networks. *Bioinformatics* **2009**, *25* (15), 1891–7.

(15) Palla, G.; Derenyi, I.; Farkas, I.; Vicsek, T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **2005**, *435* (7043), 814–8.

(16) Newman, M. E.; Girvan, M. Finding and evaluating community structure in networks. *Phys. Rev. E: Stat. Nonlin. Soft Matter Phys.* **2004**, *69* (2 Pt 2), 026113.

(17) Bader, G. D.; Hogue, C. W. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinform.* **2003**, *4*, 2.

(18) Enright, A. J.; Van Dongen, S.; Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **2002**, *30* (7), 1575–84.

(19) Soh, D.; Dong, D.; Guo, Y.; Wong, L. Consistency, comprehensiveness, and compatibility of pathway databases. *BMC Bioinform.* **2010**, *11*, 449.

(20) Liang, C. R.; Leow, C. K.; Neo, J. C.; Tan, G. S.; Lo, S. L.; Lim, J. W.; Seow, T. K.; Lai, P. B.; Chung, M. C. Proteome analysis of human hepatocellular carcinoma tissues by two-dimensional difference gel electrophoresis and mass spectrometry. *Proteomics* **2005**, *5* (8), 2258–71.

(21) Zubaidah, R. M.; Tan, G. S.; Tan, S. B.; Lim, S. G.; Lin, Q.; Chung, M. C. 2-D DIGE profiling of hepatocellular carcinoma tissues identified isoforms of far upstream binding protein (FUBP) as novel candidates in liver carcinogenesis. *Proteomics* **2008**, *8* (23−24), 5086–96.

(22) Zubaidah, R. M. A comprehensive proteome analysis of hepatitis B virus- associated hepatocellular carcinoma; National University of Singapore, 2009.

(23) Tan, H. T.; Tan, S.; Lin, Q.; Lim, T. K.; Hew, C. L.; Chung, M. C. Quantitative and temporal proteome analysis of butyrate-treated colorectal cancer cells. *Mol. Cell. Proteomics* **2008**, *7* (6), 1174–85.

(24) Prasad, T. S.; Kandasamy, K.; Pandey, A. Human Protein Reference Database and Human Proteinpedia as discovery tools for systems biology. *Methods Mol. Biol.* **2009**, *577*, 67–79.

(25) Stark, C.; Breitkreutz, B. J.; Reguly, T.; Boucher, L.; Breitkreutz, A.; Tyers, M. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* **2006**, *34* (Database issue), D535–9.

(26) Aranda, B.; Achuthan, P.; Alam-Faruque, Y.; Armean, I.; Bridge, A.; Derow, C.; Feuermann, M.; Ghanbarian, A. T.; Kerrien, S.; Khadake, J.; Kerssemakers, J.; Leroy, C.; Menden, M.; Michaut, M.; Montecchi-Palazzi, L.; Neuhauser, S. N.; Orchard, S.; Perreau, V.; Roechert, B.; van Eijk, K.; Hermjakob, H. The IntAct molecular interaction database in 2010. *Nucleic Acids Res.* **2010**, *38* (Database issue), D525–31.

(27) Hermjakob, H.; Montecchi-Palazzi, L.; Lewington, C.; Mudali, S.; Kerrien, S.; Orchard, S.; Vingron, M.; Roechert, B.; Roepstorff, P.; Valencia, A.; Margalit, H.; Armstrong, J.; Bairoch, A.; Cesareni, G.; Sherman, D.; Apweiler, R. IntAct: an open source molecular interaction database. *Nucleic Acids Res.* **2004**, *32* (Database issue), D452–5.

(28) Xenarios, I.; Salwinski, L.; Duan, X. J.; Higney, P.; Kim, S. M.; Eisenberg, D. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* **2002**, *30* (1), 303–5.

(29) Xenarios, I.; Fernandez, E.; Salwinski, L.; Duan, X. J.; Thompson, M. J.; Marcotte, E. M.; Eisenberg, D. DIP: The Database of Interacting Proteins: 2001 update. *Nucleic Acids Res.* **2001**, *29* (1), 239–41.

(30) Xenarios, I.; Rice, D. W.; Salwinski, L.; Baron, M. K.; Marcotte, E. M.; Eisenberg, D. DIP: the database of interacting proteins. *Nucleic Acids Res.* **2000**, *28* (1), 289–91.

(31) Rual, J. F.; Venkatesan, K.; Hao, T.; Hirozane-Kishikawa, T.; Dricot, A.; Li, N.; Berriz, G. F.; Gibbons, F. D.; Dreze, M.; Ayivi-Guedehoussou, N.; Klitgord, N.; Simon, C.; Boxem, M.; Milstein, S.; Rosenberg, J.; Goldberg, D. S.; Zhang, L. V.; Wong, S. L.; Franklin, G.; Li, S.; Albala, J. S.; Lim, J.; Fraughton, C.; Llamosas, E.; Cevik, S.; Bex, C.; Lamesch, P.; Sikorski, R. S.; Vandenhaute, J.; Zoghbi, H. Y.; Smolyar, A.; Bosak, S.; Sequerra, R.; Doucette-Stamm, L.; Cusick, M. E.; Hill, D. E.; Roth, F. P.; Vidal, M. Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **2005**, *437* (7062), 1173–8.

(32) Stelzl, U.; Worm, U.; Lalowski, M.; Haenig, C.; Brembeck, F. H.; Goehler, H.; Stroedicke, M.; Zenkner, M.; Schoenherr, A.; Koeppen, S.; Timm, J.; Mintzlaff, S.; Abraham, C.; Bock, N.; Kietzmann, S.; Goedde, A.; Toksoz, E.; Droege, A.; Krobitsch, S.; Korn, B.; Birchmeier, W.; Lehrach, H.; Wanker, E. E. A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **2005**, *122* (6), 957–68.

(33) Kanehisa, M. Representation and analysis of molecular networks involving diseases and drugs. *Genome Inform.* **2009**, *23* (1), 212–3.

(34) Kelder, T.; Pico, A. R.; Hanspers, K.; van Iersel, M. P.; Evelo, C.; Conklin, B. R. Mining biological pathways using WikiPathways web services. *PLoS One* **2009**, *4* (7), e6447.

(35) Pico, A. R.; Kelder, T.; van Iersel, M. P.; Hanspers, K.; Conklin, B. R.; Evelo, C. WikiPathways: pathway editing for the people. *PLoS Biol.* **2008**, *6* (7), e184.

(36) Jimenez-Marin, A.; Collado-Romero, M.; Ramirez-Boo, M.; Arce, C.; Garrido, J. J. Biological pathway analysis by ArrayUnlock and Ingenuity Pathway Analysis. *BMC Proc.* **2009**, *3* (Suppl 4), S6.

(37) Floyd, R. W. Algorithm 97: Shortest Path. *Commun. ACM* **1962**, *5* (6), 345.

(38) Searle, B. C.; Turner, M.; Nesvizhskii, A. I. Improving sensitivity by probabilistically combining results from multiple MS/MS search methodologies. *J. Proteome Res.* **2008**, *7* (1), 245–53.

(39) Nesvizhskii, A. I.; Roos, F. F.; Grossmann, J.; Vogelzang, M.; Eddes, J. S.; Gruissem, W.; Baginsky, S.; Aebersold, R. Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides. *Mol. Cell. Proteomics* **2006**, *5* (4), 652–70.

(40) Mathivanan, S.; Periaswamy, B.; Gandhi, T. K.; Kandasamy, K.; Suresh, S.; Mohmood, R.; Ramachandra, Y. L.; Pandey, A. An evaluation of human protein-protein interaction data in the public domain. *BMC Bioinform.* **2006**, *7* (Suppl 5), S19.

(41) Kato, K.; Yamashita, R.; Matoba, R.; Monden, M.; Noguchi, S.; Takagi, T.; Nakai, K. Cancer gene expression database (CGED): a database for gene expression profiling with accompanying clinical information of human cancer tissues. *Nucleic Acids Res.* **2005**, *33* (Database issue), D533–6.

(42) Washburn, M. P.; Wolters, D.; Yates, J. R., 3rd Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **2001**, *19* (3), 242–7.

(43) Yang, X.; Lazar, I. M. MRM screening/biomarker discovery with linear ion trap MS: a library of human cancer-specific peptides. *BMC Cancer* **2009**, *9*, 96.

(44) Donato, F.; Tagger, A.; Gelatti, U.; Parrinello, G.; Boffetta, P.; Albertini, A.; Decarli, A.; Trevisi, P.; Ribero, M. L.; Martelli, C.; Porru, S.; Nardi, G. Alcohol and hepatocellular carcinoma: the effect of lifetime intake and hepatitis virus infections in men and women. *Am. J. Epidemiol.* **2002**, *155* (4), 323–31.

(45) Pachiadakis, I.; Pollara, G.; Chain, B. M.; Naoumov, N. V. Is hepatitis C virus infection of dendritic cells a mechanism facilitating viral persistence? *Lancet Infect Dis.* **2005**, *5* (5), 296–304.

(46) Melen, K.; Fagerlund, R.; Nyqvist, M.; Keskinen, P.; Julkunen, I. Expression of hepatitis C virus core protein inhibits interferon-induced nuclear import of STATs. *J. Med. Virol.* **2004**, *73* (4), 536–47.

(47) Branda, M.; Wands, J. R. Signal transduction cascades and hepatitis B and C related hepatocellular carcinoma. *Hepatology* **2006**, *43* (5), 891–902.

(48) Dang, C. V.; Semenza, G. L. Oncogenic alterations of metabolism. *Trends Biochem. Sci.* **1999**, *24* (2), 68–72.

(49) Wang, D.; Dubois, R. N. Eicosanoids and cancer. *Nat. Rev. Cancer* **2010**, *10* (3), 181–93.