Disease Markers 28 (2010) 233–239
DOI 10.3233/DMA-2010-0696
IOS Press



# A systems approach to the proteomic identification of novel cancer biomarkers

Sharon Pitteri and Sam Hanash*
*Fred Hutchinson Cancer Research Center, Seattle, WA, USA*

**Abstract**. The proteomics field has experienced rapid growth with technologies achieving ever increasing accuracy, sensitivity, and throughput, and with availability of computational tools to address particular applications. Given that the proteome represents the most functional component encoded for in the genome, a systems approach to disease investigations and biomarker identification benefits substantially from integration of proteome level studies. Here we present proteomic approaches that have allowed systematic searches for potential cancer markers by integrating cancer cell profiling with additional sources of data, as illustrated with recent studies of ovarian cancer.

## 1. Introduction

Previously, systematic interrogation of the proteome has represented a substantial challenge that limited integration of proteomics into systems biology. However substantial progress has been made in developing strategies for mining complex proteomes. Given the multiple facets of the proteome that encompass quantitative levels of proteins, their sub-cellular distribution, alternative splicing, cleavages and numerous post-translational modifications, their activity states, and occurrence of proteins in complexes, mining the proteome requires the use of suitable technologies for particular applications [1].

Proteomics is increasingly relied upon for understanding basic cellular and physiologic processes in an unbiased systems-wide fashion [2]. From a disease point of view proteomics has the potential to yield critical pathways, novel targets for therapeutics, biomarkers relevant to risk assessment and early disease detection, and molecular classification and monitoring of disease response to therapy and progression [3]. Here, we use ovarian cancer as a case study to address the contributions of proteomics to biomarker identification through integration of data from profiling cancer cells with other sources of data.

*Corresponding author: Sam Hanash, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N., M5-C800, PO Box 19024, Seattle, WA 98109, USA. E-mail: shanash@fhcrc.org.

## 2. Interrogating cancer sub-cellular proteomes

A multitude of approaches for cell protein analysis have yielded quantitative information with respect to cellular protein composition and responses to environmental changes and to gene manipulations. Label free methods have been successfully utilized to derive quantitative data from mass spectrometry [4]. Alternatively, in vitro stable isotope labeling of proteins or of amino acids in cell culture has become widely used to study the proteomes of various cell types and how they change in response to various conditions [5]. For isotopic labeling, cells are cultured in the presence of an amino acid, such as lysine, for which all the $^{12}C$ are substituted by $^{13}C$. Incorporation of the isotopically labeled amino acid occurs during cell growth, protein synthesis, and turnover. Isotopic labeling allows “light” and “heavy” proteomes to be distinguished by mass spectrometry. In one study [6] SILAC based quantitative mass spectrometry was relied upon to identify tyrosine phosphorylated proteins in isogenic human bronchial epithelial cells and human lung adenocarcinoma cell lines, expressing mutant EGFR or a mutant KRAS allele. Tyrosine phosphorylation of signaling molecules was found to

ISSN 0278-0240/10/$27.50 

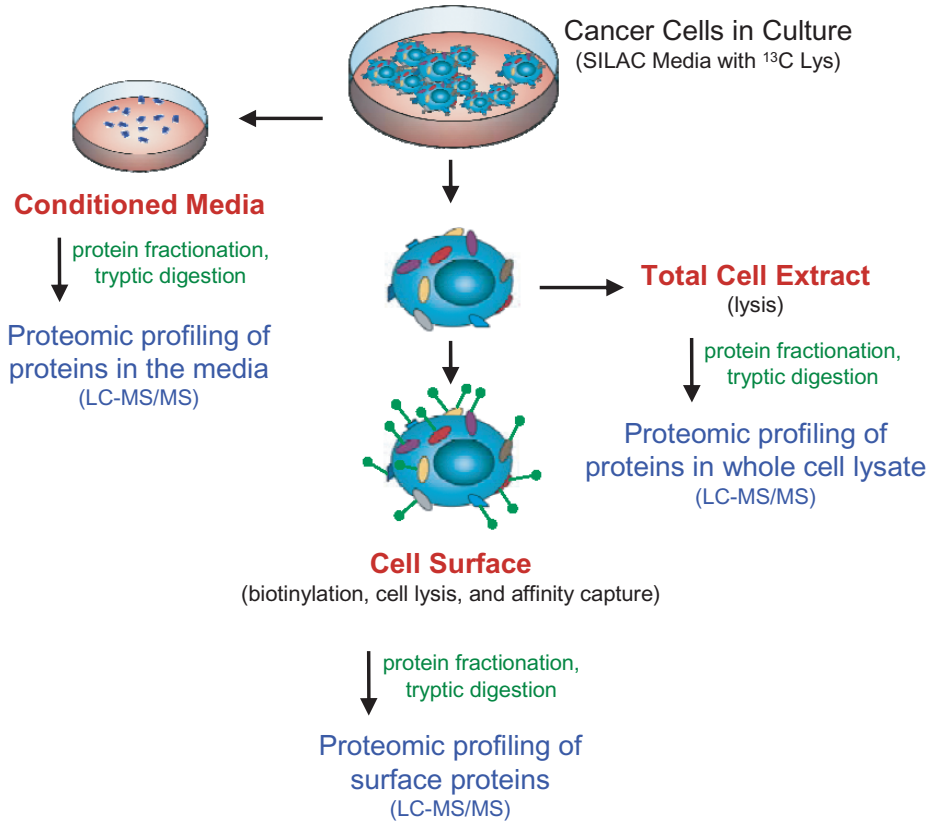


Fig. 1. Cell based proteomics discovery workflow.

be greater in cells expressing mutant EGFR compared to mutant wild type EGFR or mutant KRAS.

A broader interrogation of cellular processes using proteomics encompasses in-depth analysis of whole cell lysates, of proteins localized to various compartments notably the cell surface which mediates cell-cell interactions and responses to changes in the microenvironment, and proteins that are secreted or otherwise released into the extracellular compartment (Fig. 1).

This approach is currently being applied by our group to characterize the proteome repertoires of epithelial cancer cell types including lung, colon, breast, and ovarian cancer, to determine the extent of their similarities and differences as part of an effort to identify novel biomarkers and targets for therapeutics. Several findings have emerged from our studies of ovarian cancer cells that encompassed established cell lines as well as primary ovarian cancer cells isolated from ascites fluid [7]. Proteomic data from cancer cells are integrated with gene expression data, and with plasma proteome profiles from mice bearing ovarian tumors and plasmas from newly diagnosed women with ovarian cancer to identify potential ovarian cancer markers.

Separate sets of proteins released into culture media and of cell surface proteins from ovarian cancer cell lines and primary tumor cells using total cell lysates as a reference were subjected to fractionation and high resolution mass spectrometry. Some 6,400 proteins were identified in this analysis with high confidence [7]. Such a depth of analysis is remarkable and likely indicates that proteins have been identified for most RNAs that are translated into proteins. Abundance of identified proteins in sub-proteomes were estimated based on spectral counts [8]. Occurrence of proteins in particular sub-proteomes was largely consistent with predictions made from database searches and sequence analysis, with some exceptions particularly related to cleavages of surface anchored proteins and release of cleaved fragments into the medium. Interestingly, a large proportion of the shed proteins were found to be related to processes of cell adhesion and cell movement (Table 1).

It is interesting to note that a large number of proteins that have been previously characterized as candidate markers for ovarian cancer based on gene expression studies or based on discoveries of tumor antigens were identified as highly enriched in the secreted/shed

Table 1

Proteins enriched (> 10-fold) in both cell surface and conditioned media related to cell adhesion and/or cell movement

| | | | | | PROTEIN ENRICHMENT[a] | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | OVCAR3 | | CaOV3 | | ES2 | | Ascites Cells | |
| IPI | Gene Name | Description | Cell Adhesion[b] | Cell Movement[b] | Secreted | Surface | Secreted | Surface | Secreted | Surface | Secreted | Surface |
| IPI00013897 | ADAM10 | ADAM 10 | X | X | 2.9 | 4.0 | 2.0 | 3.0 | 30.7 | 54.0 | 0.3 | 5.5 |
| IPI00440932 | ADAM9 | ADAM 9 | X | | 30.1 | 18.0 | – | 3.7 | 82.2 | 11.9 | 1.7 | 3.4 |
| IPI00022229 | APOB | APOLIPOPROTEIN B-100 | | X | 4.2 | 6.7 | – | 94.5 | – | 245.9 | 16.6 | 20.3 |
| IPI00006608 | APP | AMYLOID BETA A4 | X | X | 24.9 | 2.7 | – | – | 47.4 | 81.1 | 6.0 | – |
| IPI00296992 | AXL | AXL RECEPTOR TYROSINE KINASE | X | X | – | – | 6.0 | – | 53.0 | 13.6 | – | – |
| IPI00215767 | B4GALT1 | BETA-1,4-GALACTOSYLTRANSFERASE 1 | X | X | 49.4 | – | 95.6 | 17.0 | 8.4 | – | 0.7 | – |
| IPI00290085 | CDH2 | NEURAL CADHERIN | X | X | 114.0 | – | – | – | 47.4 | 13.8 | 10.9 | 16.8 |
| IPI00028931 | DSG2 | DESMOGLEIN-2 | X | | 3.3 | 3.6 | 2.0 | 14.9 | – | 3.5 | 15.9 | 35.5 |
| IPI00013744 | ITGA2 | INTEGRIN ALPHA-2 | X | X | 11.1 | 28.6 | – | 12.9 | – | 18.8 | – | 13.9 |
| IPI00027087 | L1CAM | NEURAL CELL ADHESION MOLECULE L1 | X | X | 7.5 | 51.9 | 4.7 | 5.4 | – | – | 53.8 | 30.0 |
| IPI00010333 | MDK | MIDKINE | | X | 91.4 | 23.6 | 80.6 | – | 4.5 | 0.9 | 25.8 | – |
| IPI00023814 | NEO1 | NEOGENIN | X | X | – | – | 15.9 | 27.7 | – | 16.1 | – | 6.8 |
| IPI00022284 | PRNP | MAJOR PRION PROTEIN | | X | 4.2 | – | 6.0 | – | 48.8 | 67.4 | 13.4 | 6.8 |
| IPI00107831 | PTPRF | RECEPTOR-TYPE TYROSINE-PROTEIN PHOSPHATASE F | X | X | 2.5 | – | 2.4 | 6.9 | 45.6 | 64.1 | 38.2 | 98.8 |
| IPI00168358 | RGMB | RGM DOMAIN FAMILY, MEMBER B | X | | 4.2 | – | – | – | 30.7 | 11.1 | – | – |
| IPI00027434 | RHOC | RHO-RELATED GTP-BINDING PROTEIN RHOC | | X | – | – | 11.0 | 11.7 | – | – | – | – |
| IPI00011564 | SDC4 | SYNDECAN-4 | | X | 7.5 | – | 11.0 | – | – | – | 30.7 | 12.5 |
| IPI00749245 | SFRP1 | SECRETED FRIZZLED-RELATED PROTEIN 1 | | X | | – | 20.9 | – | 74.8 | 27.0 | – | – |
| IPI00219663 | TCIRG1 | VACUOLAR PROTON TRANSLOCATING ATPASE 116 KDA SUBUNIT A | | X | 4.2 | 12.3 | – | 6.3 | – | – | 10.9 | 28.3 |
| IPI00018219 | TGFB1 | TRANSFORMING GROWTH FACTOR-BETA-INDUCED PROTEIN IG-H3 | X | X | – | – | 160.2 | 14.4 | – | – | 15.9 | 6.8 |
| IPI00031008 | TNC | TENASCIN | X | X | 26.8 | – | – | 3.7 | 283.2 | 435.2 | 85.3 | 42.7 |

[a]The enrichment was calculated based on the ratio of normalized spectral count in (conditioned media or cell surface) / total cell extract.
[b]As annotated by Ingenuity Pathway Analysis Software.

fraction or on the surface of ovarian cancer cells. The list includes IGFBP2, IGFBP3, KLK6, KLK7, KLK9, MDK, MUC16 (CA125), PROS1, SLPI, TIMP1, and WFDC2 (HE4).

## 3. Profiling of biological fluids for candidate cancer markers

Proteomics is particularly promising for the analysis of biological fluids and biomarker discovery. Biomarkers are indicators of specific physiologic and pathologic states. Cancer biomarkers can aid in diagnosis and/or patient management by defining subtypes and predicting or monitoring response to treatment and disease progression or regression [3]. The search for biomarkers that can be assayed in biological fluids notably serum and plasma has relied on various strategies each with its own advantages and disadvantages related in part to depth of analysis and throughput. Serum and plasma from which serum is derived are among the most accessible biological materials and have been relied upon for cancer screening, diagnosis or disease monitoring as in the case of CA125 for ovarian cancer, CA19.9 for pancreatic cancer, and PSA for prostate cancer. The plasma proteome is particularly challenging for in-depth analysis because of its complexity and because of the vast dynamic range of its protein constituents. While mass spectrometry has evolved sufficiently to detect and identify femtomoles of peptides, the dynamic range of detection is still a limiting factor and currently does not exceed 3–4 orders of magnitude for complex mixtures such as plasma [9].

Three basic strategies have facilitated achievement of in-depth analysis of the serum and plasma proteomes: i) removal of high abundance proteins, such as albumin and immunoglobulins that interfere with the detection of less abundant proteins, by means such as immunodepletion [10]; ii) fractionation of samples by chromatographic or other means of separation resulting in individual fractions with reduced complexity, a process that is better suited for the capabilities of mass spectrometry [11] and iii) targeted analysis of particular protein or peptide subsets such as glycoproteins [12]. To study changes in the plasma proteome with disease state, we have employed a strategy involving depletion of abundant proteins, isotopic labeling, extensive fractionation and high resolution mass spectrometry (Fig. 2). The integration of cancer cell findings with plasma proteome findings, has enabled our group to identify several novel cancer biomarkers for epithelial tumors [13,14] as presented here for ovarian cancer.

We applied the in-depth quantitative plasma proteomics strategy shown in Fig. 2, to a mouse model of ovarian cancer [13]. 106 proteins showed increased levels in tumor bearing mice compared to controls. 58 of the 106 proteins were also found to be secreted or shed from ovarian cancer cells using the strategy shown in Fig. 1, while the remainder consisted primarily of host-response proteins. Data integration across studies yielded eight proteins, ascertained to represent mostly secreted proteins that were common to mouse plasma and human cancer cells. These proteins were found to be significantly upregulated in a set of plasmas from ovarian cancer patients. Five of the eight proteins (GRN, IGFBP2 RARRES2, TIMP1, and CD14) were confirmed to be upregulated in a second independent set of ovarian cancer plasmas, including early stage disease. This integrated proteomic approach is an effective approach to identify potential circulating biomarkers.

## 4. Beyond quantitative analysis of protein levels

Although substantial depth of analysis is currently feasible to identify and quantify proteins in complex mixtures, there is a need to expand the reach of proteomic based biomarker applications to mine post-translational modifications that are disease related. Glycoproteins represent a subset of proteins that have important functions such as cell-cell interaction and exhibit alterations in disease states. Several technologies are available to capture and analyze cell and tissue glycoproteins [15–17]. An illustrative study focused on the effects of cell-cell interactions on N-linked glycans in epithelial cells [18]. N-glycans were purified from whole cell lysates and then detected by high performance liquid chromatography and mass spectrometry. GlcNAc-containing N-glycans, which are dependent on N-acetylglucosaminyltransferase III (GnT-III), were found to be substantially increased in cells cultured under dense conditions compared with those cultured under low-density conditions. Concordant increases in expression levels and activities of GnT-III but not other glycosyltransferases were also found. Disruption of E-cadherin-mediated adhesion by treatment with EDTA or a neutralizing anti-E-cadherin antibody abolished the up-regulation of expression of GnT-III. The data suggested that E-cadherin-dependent pathway regulated GnT-III expression. This study emphasizes the in-

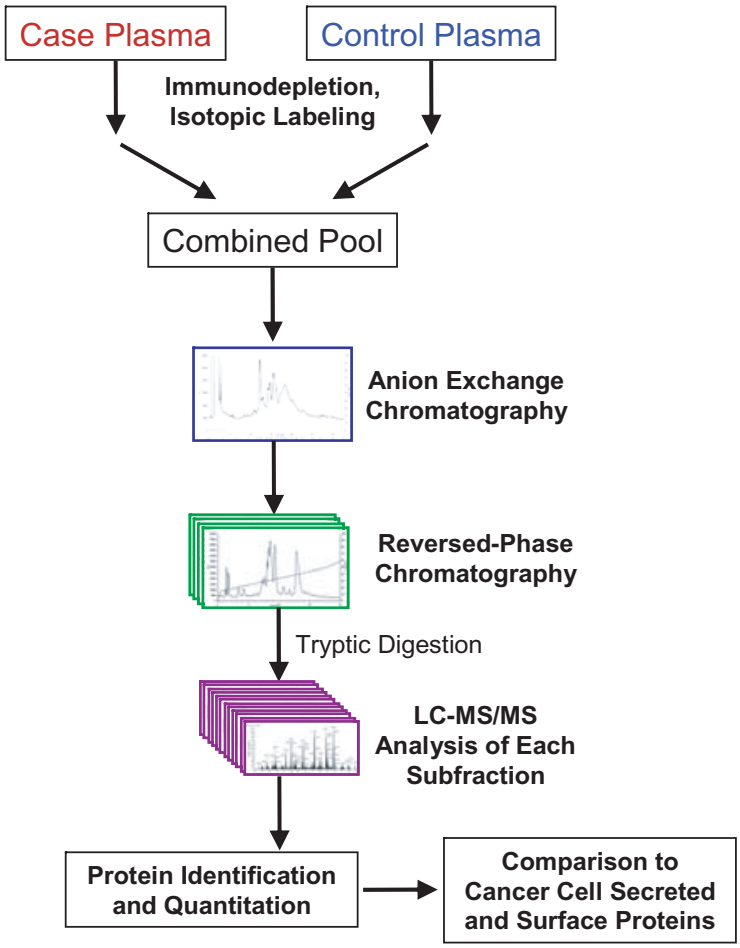


Fig. 2. A multi-step analysis of plasma proteins for potential cancer markers.

terplay of glycosylation enzymes with other regulatory systems and their impact on glycoproteins.

In our studies of ovarian cancer, a substantial fraction of candidate markers identified represented glycoproteins. An analysis of glycan structures associated with these proteins associated with the cancer state likely would improve the specificity of the candidate markers identified.

Aside from studies of protein levels and modifications, other approaches that are focused on function likely would be informative with respect to disease specific alterations. Such approaches include activity-based protein profiling which relies on active site-directed probes to interrogate, for example, the functional state of enzyme families [19]. The delineation of enzyme activities selectively associated with disease processes has the potential to yield a rich source of targets for diagnosis and therapy. In one study [20], an active site-directed chemical probe for profiling histone deacetylases in native proteomes and live cells was used to profile both the activity state of HDACs and the binding proteins that regulate their function. The probe was applied to assess differences in acetylase content and complex assembly in human disease models.

## 5. Databasing of proteomics data to facilitate systems interrogations

A large number of databases have been developed for depositing and retrieving proteomic datasets including PRIDE [21], PeptideAtlas [22], UniPep [23], the Global Proteome Machine [24], Proteopedia [25], and Proteome Commons and its Tranche file-sharing system [www.tranche.proteomecommons.org].

While initial analysis of proteomics data focuses on identifying individual proteins of interest, computational tools have led to the ability to discern rel-

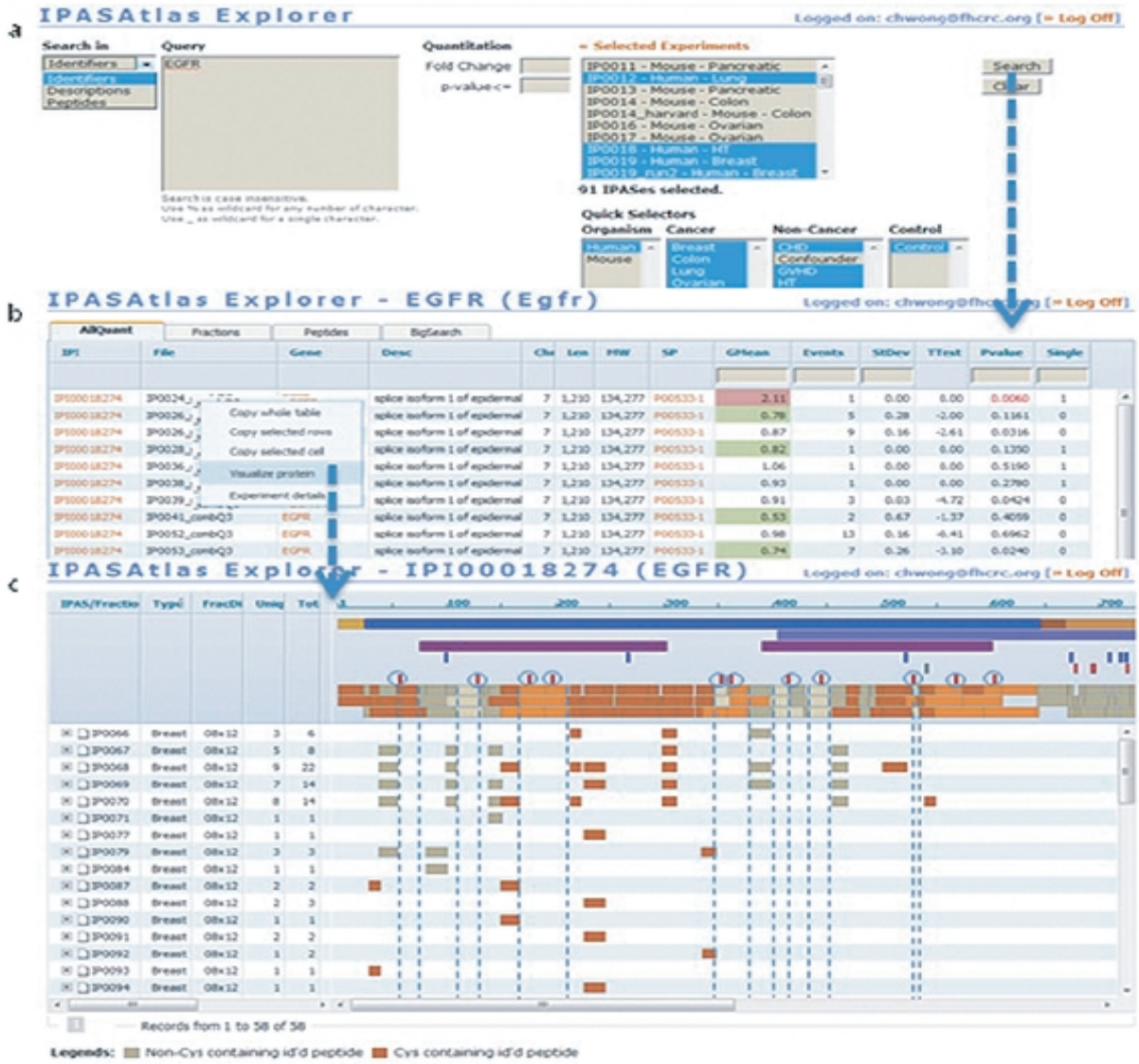


Fig. 3. Construction of a plasma proteome database to provide easy and consistent access to identification and quantification results to facilitate user query and analysis. A menu (a) allows users tolook up the normalized expression ratio of their protein(s) of interest in plasma across multiple experiments (b). The data is cross-linked to other public databases such as IPI, SwissProt and GeneCard. (c) SwissProt annotation is integrated into the protein visualization to assist user with interpretation. The circled red bar in (c) are glycosites of EGFR. Lack of identified peptides containing glycosites is likely a result of glycosylation.

evant networks and pathways from the data. There are currently no standard approaches for this purpose. However approaches being applied to transcriptomic data also have utility for proteomic data including Gene Ontology (http://www.geneontology.org), Kyoto Encyclopedia of Genes and Genomes (KEGG http://www.genome.jp/kegg), Ingenuity Pathway Analysis (http://www.ingenuity.com), GeneGo by Metacore (http://www.genego.com), gene set enrichment, etc. From such analyses, overrepresentation of particular pathways may be discerned which has the added advantage of increasing confidence in particular protein assignments. Therefore the approach of integrative genomics, assimilating data and information from multiple sources also has relevance to proteomics.

Integration of multiple datasets is valuable for deriving hypotheses, and developing a list of candidate markers to guide subsequent studies. A recent effort to facilitate mining of proteomics data for biomarker discovery is BiomarkerDigger (http://biomarkerdigger.org),

with its automated data analysis and search functions. Its metadata-gathering function searches available proteome databases for protein-protein interaction, Gene Ontology annotations, protein domain information, and tissue expression profiles, and integrates it into protein dataset profiles accessible by search functions in BiomarkerDigger.

Under construction by our group is an integrated database to manage protein identification and quantification information from various studies (Fig. 3). The system is intended to provide easy and consistent access to identification and quantitationresults which facilitates user query and analysis. Users may look up normalized expression ratios of their proteins of interest in plasma across multiple experiments. The data is cross-linked to other public databases such as IPI, SwissProt, and GeneCard. To assist users in their interpretation, SwissProt annotation is integrated into the protein visualization.

## 6. Conclusion

A near exhaustive analysis of the protein content of cells and biological fluids spanning six to seven logs is currently possible. These advances should facilitate integration of proteomic data with other sources of data, thus empowering a systems approach to disease investigations.

## References


[1] M. Mann and N.L. Kelleher, Precision proteomics: the case for high resolution and high mass accuracy, *Proc Natl Acad Sci U S A* **105**(47) (2008), 18132–18138.

[2] B.F. Cravatt, G.M. Simon and J.R. Yates, 3rd, The biological impact of mass-spectrometry-based proteomics, *Nature* **450**(7172) (2007), 991–1000.

[3] S.M. Hanash, S.J. Pitteri and V.M. Faca, Mining the plasma proteome for cancer biomarkers, *Nature* **452**(7187) (2008), 571–579.

[4] A.S. Haqqani, J.F. Kelly and D.B. Stanimirovic, Quantitative protein profiling by mass spectrometry using label-free proteomics, *Methods Mol Biol* **439** (2008), 241–256.

[5] S.E. Ong and M. Mann, A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC), *Nat Protoc* **1**(6) (2006), 2650–2660.

[6] U. Guha et al., Comparisons of tyrosine phosphorylated proteins in cells expressing lung cancer-specific alleles of EGFR and KRAS, *Proc Natl Acad Sci U S A* **105**(37) (2008), 14112–14117.

[7] V.M. Faca et al., Proteomic analysis of ovarian cancer cells reveals dynamic processes of protein secretion and shedding of extra-cellular domains, *PLoS ONE* **3**(6) (2008), e2425.

[8] H. Liu, R.G. Sadygov and J.R.R. Yates, A model for random sampling and estimation of relative protein abundance in shotgun proteomics, *Analytical Chemistry* **76**(14) (2004), 4193–4201.

[9] M. Mann and N.L. Kelleher, Special Feature: Precision proteomics: The case for high resolution and high mass accuracy, *Proc Natl Acad Sci U S A*, 25 Sep 2008, p. Epub ahead of print.

[10] T. Liu et al., Evaluation of multiprotein immunoaffinity subtraction for plasma proteomics and candidate biomarker discovery using mass spectrometry, *Mol Cell Proteomics* **5**(11) (2006), 2167–2174.

[11] H. Wang and S. Hanash, Intact-protein based sample preparation strategies for proteome analysis in combination with mass spectrometry, *Mass Spectrom Rev* **24**(3) (2005), 413–426.

[12] Y. Zhou, R. Aebersold and H. Zhang, Isolation of N-linked glycopeptides from plasma, *Anal Chem 79*(15) (2007), 5826–5837.

[13] S.J. Pitteri et al., Integrated proteomic analysis of human cancer cells and plasma from tumor bearing mice for ovarian cancer biomarker discovery, *PLoS One* **4**(11) (2009), e7916.

[14] V.M. Faca et al., A mouse to human search for plasma proteome changes associated with pancreatic tumor development, *PLoS Med* **5**(6) (2008), e123.

[15] A.D. Taylor et al., Towards an integrated proteomic and glycomic approach to finding cancer biomarkers, *Genome Med* **1**(6) (2009), 57.

[16] Y.Y. Zhao et al., Functional roles of N-glycans in cell signaling and cell adhesion in cancer, *Cancer Sci* **99**(7) (2008), 1304–1310.

[17] K.T. Pilobello and L.K. Mahal, Deciphering the glycocode: the complexity and analytical challenge of glycomics, *Curr Opin Chem Biol* **11**(3) (2007), 300–305.

[18] J. Iijima et al., Cell-cell interaction-dependent regulation of N-acetylglucosaminyltransferase III and the bisected N-glycans in GE11 epithelial cells. Involvement of E-cadherin-mediated cell adhesion, *J Biol Chem* **281**(19) (2006), 13038–13046.

[19] B.F. Cravatt, A.T. Wright and J.W. Kozarich, Activity-based protein profiling: from enzyme chemistry to proteomic chemistry, *Annu Rev Biochem* **77** (2008), 383–414.

[20] C.M. Salisbury and B.F. Cravatt, Activity-based probes for proteomic profiling of histone deacetylase complexes, *Proc Natl Acad Sci U S A* **104**(4) (2007), 1171–1176.

[21] P. Jones et al., PRIDE: a public repository of protein and peptide identifications for the proteomics community, *Nucleic Acids Res* **34**(Database issue) (2006), D659–D663.

[22] E.W. Deutsch et al., Human Plasma PeptideAtlas, *Proteomics* **5**(13) (2005), 3497–3500.

[23] H. Zhang et al., UniPep – a database for human N-linked glycosites: a resource for biomarker discovery, *Genome Biol* **7**(8) (2006), R73.

[24] R. Craig, J.P. Cortens and R.C. Beavis, Open source system for analyzing, validating, and storing protein identification data, *J Proteome Res* **3**(6) (2004), 1234–1242.

[25] S. Mathivanan et al., Human Proteinpedia enables sharing of human protein data, *Nat Biotechnol* **26**(2) (2008), 164–167.