A new bioinformatics insight into human cancer-associated proteins

FU-JUN LIU[1*], XIU-FENG HUA[2*] and WEN-JUAN WANG[3]

[1]Central Laboratory, [2]Department of Endocrinology, [3]Reproduction Medical Center, Yu-Huang-Ding Hospital/Qingdao University, Yantai 264000, Shandong, P.R. China

Received November 23, 2011; Accepted January 30, 2012

DOI: 10.3892/or.2012.1714

Abstract. Cancer is a complex disease caused by multiple factors including genetic mutations, and environmental factors. Cancer-associated proteins are potential biomarkers or targets for diagnostic and therapeutic interventions in cancer. The Universal Protein Resourse (UniPort) is a well-annotated comprehensive resourse for protein sequence records. In the present study, we performed data mining of UniProt proteins as a proteomic resource. We generated a catalog of 1653 cancer-associated proteins including 344 secretory proteins and 300 cell surface proteins. Integrated bioinformatic analysis including ontological classification, functional enrichment and pathway construction were performed. These proteins could serve as a reference for further studies to discover cancer targets, and the enriched bioinformatic analysis provides new insights into cancer proteomics research.

Introduction

Cancer is a complex disease that is caused by multiple factors including genetic mutations, and environmental factors (1,2). Cancer biomarker discovery is crucial for both cancer biology and clinical applications. These biomarkers can be DNA, RNA, miRNA or proteins (3), and the preferred one being the protein (4).

Biomarkers are indicators of the specific biological status and can be useful for diagnosis and early detection of cancers, assessment of prognosis, and treatment monitoring (5). The development and improvement of biotechnologies have allowed researchers to perform high throughput analysis of genomes, transcriptomes and proteomes in health and disease, and identify hundreds of potential biomarkers (6). However, less than two dozen cancer biomarkers are currently approved by the Food and Drug Administration (FDA) (7), including only nine protein biomarkers in the blood (8). Due to the lack of the sensitivity and specificity of these known biomarkers (9), researchers continue to look for more meaningful targets. Among these platforms, proteomics is particularly promising for the discovery of biomarkers (10). Proteomic technologies can provide high-throughput and in-depth analysis. Many of these studies have been performed on different cancer types including lung (11), breast (12), colorectal (13), bladder (14), prostate (15), head and neck (16) and ovarian cancers (17). The different cancer proteomics have generated a set of putative cancer biomarkers. However, these results should be verified and validated before they can be used in clinical detection, and the specificity and reproducibility also need to be addressed (18). The major challenge of these studies is how to decipher the results and extrapolate them into clinical applications. Proteins with altered expressions in cancer do not act as individual units, but are involved in certain pathways and play different biological functions. A wide variety of cancers may also be linked to the same pathways that affect tumorigenesis and progression through altering protein expressions. So, once these pathways are known, it should be easier to monitor different aspects of cancer progression and to target therapeutic strategies by focusing on pathways instead of individual proteins. The enriched pathways or functions may be the most probable cause of cancer (19). Therefore, bioinformatics should be applied in discovering cancer-associated biomarkers. Identification and characterization of cancer proteins at integrated levels are key steps to improving our understanding of cancer biology, as well as cancer diagnosis and therapeutics. The study will also facilitate the exploration of associations between individual proteins and cancer biology.

Despite the increasing pace of data generation, efforts on organizing these data for useful exploitation are still limited. The Universal Protein Resource (UniProt, http://www.uniprot.org) has comprehensively covered 20244 reviewed proteins with detailed updated annotations (20). This well-annotated database on human proteome provides possibilities for screening new biomarkers or therapeutic targets for cancer, and also a good background for integrated functional interpretation of cancer biology.

In the present study, we performed an alternative strategy to explore cancer-associated proteins following analysis by integrated bioinformatics. This bioinformatics insight into cancer-associated protein profiles can potentially provide clues

Correspondence to: Dr Fu-Jun Liu, Central Laboratory, Yu-Huang-Ding Hospital/Qingdao University, Yantai 264000, Shandong, P.R. China
E-mail: zxsys008@126.com

*Contributed equally

Key words: bioinformatics, biomarker, cancer, cancer/testis antigen, proteome

for identifying new functional modules in cancer and can be applied for the understanding of the underlying tumorigenesis process.

Materials and methods

Selection of cancer-associated proteins. UniProt (Release 2011_10, http://www.uniprot.org) was used to select cancer-associated proteins. All human proteins were downloaded from the UniProt database including all annotations. Cancer-associated proteins were further selected manually from the downloaded data using keywords 'cancer', 'tumor', or 'carcinoma', and corresponding regulation levels were recorded as 'overregulation', 'downregulation' and 'no-annotation'.

Bioinformatics analysis

Ontological analysis. All cancer-associated proteins were classified broadly into several catalogs according to the GO annotation (www.geneontology.org) and functions reported in the literature.

Enrichment bioinformatics analysis. Protein IDs were uploaded to DAVID (http://david.abcc.ncifcrf.gv/) and the enrichment analyses of GO terms including biological process, molecular function, and cellular component were performed by using the functional clustering annotation tools. The default options with high classification stringency were used, and finally cluster names were extracted from the most biologically relevant GO term assigned to that cluster. For comparative functional analysis of the selected top five cancers, proteins IDs extracted from each cancer type were submitted into PANTHER (www.pantherdb.org) to determine functional categories by using gene expression tools. Each list is compared to the reference list using the binomial test for molecular function and protein class terms in PANTHER.

Pathway analysis. Ingenuity pathway analysis v8.0-2803 (IPA), (Ingenuity Systems, www.ingenuity.com) was used to analyze pathways and networks involved in the cancer-associated proteins. The following settings were used: reference set, ingenuity knowledge base (genes only); network analysis, drect and indirect relationships; molecules per network, 35; networks per analysis, 25; all species, tissues and cell lines were used for the analysis. IPA uses Fisher's exact test to determine which pathways (canonical pathways, toxicity pathways or biological functions) are significantly linked to the input protein set compared with the whole ingenuity knowledge base.

Selection of secretory and cell surface proteins. Secretory proteins and cell surface proteins are promising biomarkers. All cancer-associated proteins were compared with a serum/plasma proteome to select secretory proteins, and compared with cell surfaceome to select cell surface proteins. Gene Ontology (GO) was used to further filter out the result.

Results

Overview of cancer-associated proteins. By screening the reviewed proteins deposited in the UniProt database, we obtained 8756 proteins with tissue specificity description. Those proteins that have the key word 'cancer', 'tumor' or 'carcinoma' were considered as cancer-associated proteins. Finally, 1653 cancer-associated proteins were obtained, and more than 45% proteins were overexpressed in cancers and 19% were underexpressed. Fig. 1 shows the distribution of protein numbers in different cancer types. Apart from tumor/cancer cells, the top one was breast cancer.

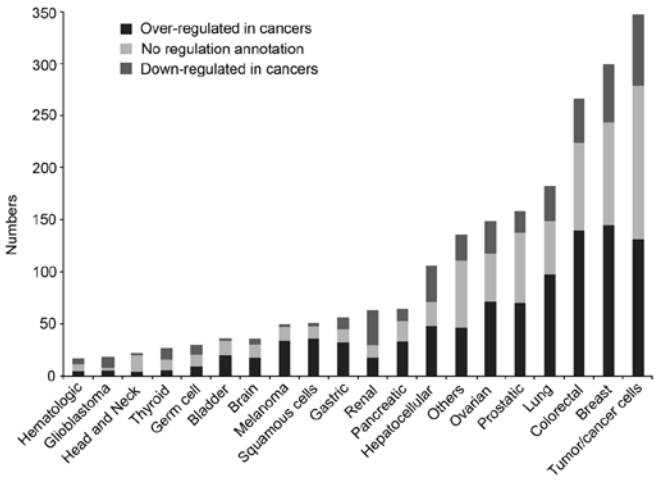

Figure 1. Distribution of proteins in different types of cancers.

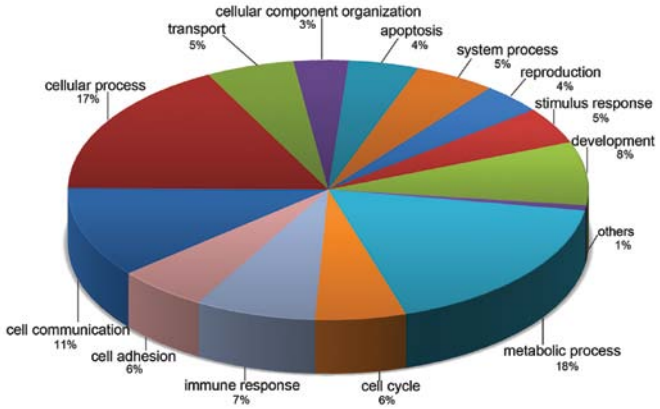

Figure 2. Broad functional classification of selected cancer-associated proteins.

Bioinformatics analysis

Ontological analysis. Functional classification analysis according to the gene ontological (GO) annotations revealed that the proteins were clustered into several categories. Of these molecular processes, the majority (18%) of the proteins were involved in metabolic processes and 17% in cellular processes (Fig. 2).

Functional clustering. To explore the enrichment functions of these proteins, their associated biological processes, molecular function and cellular component were determined by meas of the DAVID functional annotation clustering. Functional annotation clustering demonstrated that these proteins were associated significantly with five annotation

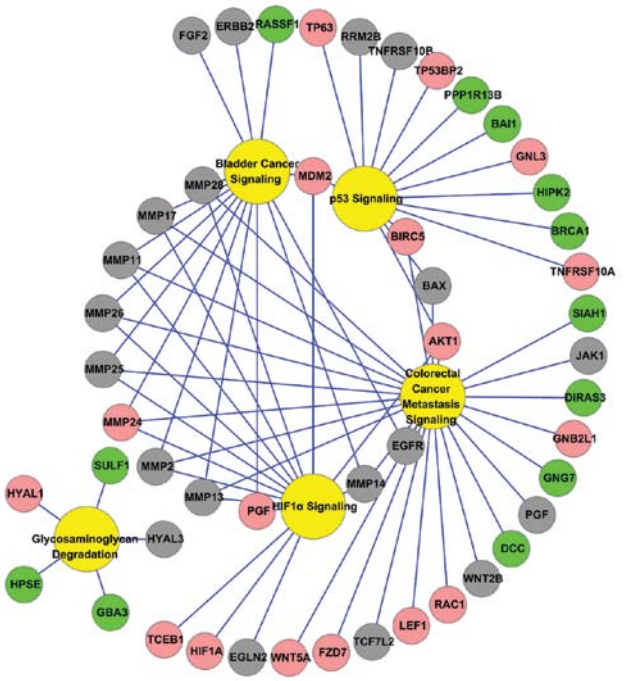

Figure 3. Diagram of the top five pathways enriched by IPA and related proteins. Green circles represent underexpressed proteins; red circles represent overexpressed proteins, and grey circles represent proteins with no regulation annotations.

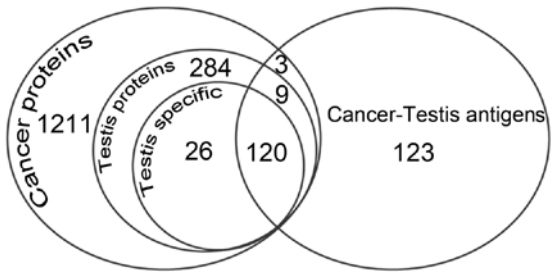

Figure 4. Comparison of cancer-associated proteins with testis proteins and cancer-testis proteins.

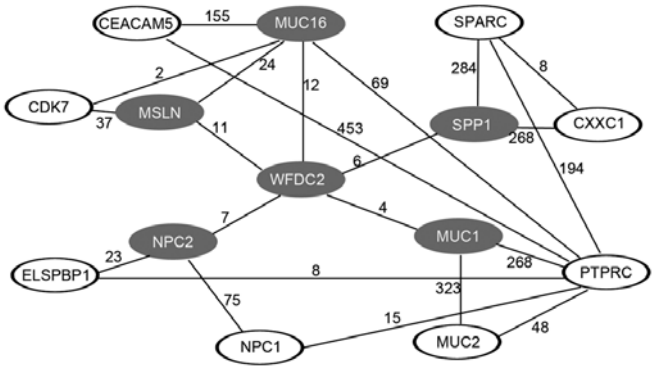

Figure 5. Literature networks of WFDC2 and related proteins with biological relationship. Five grey background proteins have direct relationship with WFDC2, and the blank background proteins have indirect relationships. The numbers represent the related number of references available in PubMed.

clusters. The first one included GO terms for apoptosis; these proteins were involved in the regulation of cell death. The second cluster included the GO terms for the cellular component extracellular region; these proteins consisted of secreted glycoproteins. The third cluster included GO terms for cell adhesion. The fourth cluster included GO terms for angiogenesis; these proteins were involved in vasculature development. The last cluster included terms for cell cycle which was related to cell division. Functional enrichment analysis of different cancer types showed that increased proportion of apoptosis and reproduction-related proteins were found in breast cancer. Cell adhesion, the immune system and reproduction-related proteins were overexpressed in ovarian cancer.

Pathway enrichment analysis. A more detailed analysis of pathways and networks were performed by using Ingenuity pathway analysis (IPA) tools. IPA generated 25 networks using all cancer-associated proteins and 18 networks using breast cancer proteins. Some canonical pathways were involved in these networks, and Fig. 3 shows the top five canonical pathways involved in the cancer proteins: bladder cancer signaling, colorectal cancer metastasis signaling, HIF1α signaling, p53 signaling, and glycosaminoglycan degradation pathways.

Cancer/testis proteins, secretory proteins and cell surface proteins. Comparison of cancer proteins with our previous study showed that 439 testis proteins including 146 testis-specific proteins were included in this study. One hundred and thirty-two common proteins are presented in the cancer proteins and cancer-testis antigens (Fig. 4).

Comparison with previous published data of human serum/plasma and cell surface proteins (surfaceome) showed that a total of 344 secretory proteins and 300 cell surface proteins were included in the cancer-associated protein database.

Literature relationship of HE4 protein with associated proteins. The HE4 (WFDC2) protein, a well-known epididymal protein, was selected as an example to evaluate its functions through PubMed retrieval. Literature retrieval showed that HE4 was significantly associated with five proteins (MUC16, MUC1, SPP1, MSLN and NPC2) (Fig. 5). These proteins were mainly involved in cell adhesion and sperm maturation processes.

Discussion

With the development and improvement of proteomic technology, proteomes of a wide range of cell types and disease processes have been identified and compared. Many studies were carried out to find cancer biomarkers in various cancer types through comparing the status of the disease (21). The obtained data offered the opportunities for the diagnosis and effective therapy of cancer. However, many studies lack experimental verification and validation, and functional analysis needs to be performed before they could be used as diagnostic markers or therapeutic targets.

In the present study, 1653 cancer-associated proteins were selected from 20244 reviewed human proteins by analyzing

the public database. These cancer-associated proteins were mainly classified into 20 specific cancer types, and most of them were cast to a broad group as tumor/cancer cells. To explore their biological relationship with cancer, we performed an integrated bioinformatics analysis. Ontological analysis indicated that most of these cancer-associated proteins were classified into different functional groups that may be involved in different aspects of cancer biology. Functional clustering analysis also indicated that these cancer-associated proteins were mainly significantly related to five GO terms: apoptosis, cell adhesion, cellular component extracellular region, angiogenesis and cell cycle. The proteins that functioned as cell adhesion (6%), cell cycle (6%) and apoptosis (4%) molecules are well-known to be involved in cancer processes (22). More than fifty cell adhesion proteins were selected in the present study. These cell adhesion molecules may play roles in intercellular and cell-extracellular matrix interactions of cancer, leading to cancer invasion or metastasis (23), or participating in signal transduction, cell growth and differentiation (24). E-cadherin is one prominent adhesion molecule that forms the E-cadherin-catenin complex which plays a role in epithelial cell-cell adhesion and differentiation (25), especially serving as a potent invasion/tumor suppressor of breast cancer (26). Some studies indicated that downregulation of E-cadherin was relevant to several pathways including the integrin-linked kinase (ILK) signaling pathway (27,28). The ILK signaling pathway is a prominent enrichment pathway in breast cancer, and it may play important roles in hormonal cancer progression (29).

Another important functional cluster was the extracellular region which consisted of glycoproteins and cell surface proteins that are excellent targets for diagnostic and therapeutic interventions. Of the 1653 proteins, 344 (21%) were reported to be present in the human serum/plasma and should be promising as biomarkers, while cell surface proteins have been deemed to serve as ideal therapeutic targets, and many monoclonal antibodies against them are approved for therapeutic applications, particularly in cancer therapy (30). Three hundred (18%) surface proteins were identified, such as ITGA9 and ITGA2B which were involved in the ILK signaling pathway.

More than two hundred cancer/testis antigens have been listed in the CT database (http://www.cta.lncc.br/). The striking feature of these proteins is their restricted expression to the testis and low or no expression in normal tissues (31). They can thus be used as potential cancer vaccine targets. One hundred and thirty-two, CT antigens (59%) and 439 testis proteins (146 testis-specific proteins) were included in the cancer-associated protein profiles; further study of these proteins is warranted. Among the 132 CT antigens, only 11 proteins were identified as cell surface proteins. The result was consistent with the characteristics of CT antigens which are mainly located in the intracellular compartment. Among these, two disintegrin metalloproteinases (ADAMs), ADAM2 and ADAM29, have been suggested to play a role in melanoma progression (32).

Apart from bioinformatics analysis, increasing references in PubMed will also provide us biological insight into cancer research (33). WFDC2 was selected as an example to explore its literature relationship with other proteins. The result was displayed as a network including 14 proteins, 5 of which were directly relevant to WFDC2. These proteins were involved in cell adhesion and sperm maturation processes.

In conclusion, this study provides new bioinformatics insight into the cancer proteome, which integrates cancer-associated proteins coming from a reviewed database to explore functions and pathways in cancer. Further studies are warranted to substantiate the enriched functions and pathways. The studies will advance our understanding of cancer biomarker discovery, and also facilitate biological interpretation of cancer biology in a network context.

References

1. Pfeifer GP and Hainaut P: Next-generation sequencing: emerging lessons on the origins of human cancer. Curr Opin Oncol 23: 62-68, 2011.
2. Ajani J and Allgood V: Molecular mechanisms in cancer: what should clinicians know? Semin Oncol 32: 2-4, 2005.
3. Aggarwal C, Somaiah N and Simon GR: Biomarkers with predictive and prognostic function in non-small cell lung cancer: ready for prime time? J Natl Compr Cancer Netw 8: 822-832, 2010.
4. Sung HJ and Cho JY: Biomarkers for the lung cancer diagnosis and their advances in proteomics. BMB Rep 41: 615-625, 2008.
5. Kim SY: A new paradigm for cancer therapeutics development. BMB Rep 43: 383-388, 2010.
6. Abu-Asab MS, Chaouchi M, Alesci S, Galli S, Laassri M, Cheema AK, Atouf F, VanMeter J and Amri H: Biomarkers in the age of omics: time for a systems biology approach. OMICS 15: 105-112, 2011.
7. Dunn BK, Wagner PD, Anderson D and Greenwald P: Molecular markers for early detection. Semin Oncol 37: 224-242, 2010.
8. Ludwig JA and Weinstein JN: Biomarkers in cancer staging, prognosis and treatment selection. Nat Rev Cancer 5: 845-856, 2005.
9. Wistuba II, Gelovani JG, Jacoby JJ, Davis SE and Herbst RS: Methodological and practical challenges for personalized cancer therapies. Nat Rev Clin Oncol 8: 135-141, 2011.
10. Schmitz-Spanke S and Rettenmeier AW: Protein expression profiling in chemical carcinogenesis: a proteomic-based approach. Proteomics 11: 644-656, 2011.
11. Hassanein M, Rahman JS, Chaurand P and Massion PP: Advances in proteomic strategies toward the early detection of lung cancer. Proc Am Thorac Soc 8: 183-188, 2011.
12. Goncalves A and Bertucci F: Clinical application of proteomics in breast cancer: state of the art and perspectives. Med Princ Pract 20: 4-18, 2011.
13. Tjalsma H: Identification of biomarkers for colorectal cancer through proteomics-based approaches. Expert Rev Proteomics 7: 879-895, 2010.
14. Schwamborn K, Gaisa NT and Henkel C: Tissue and serum proteomic profiling for diagnostic and prognostic bladder cancer biomarkers. Expert Rev Proteomics 7: 897-906, 2010.
15. Larkin SE, Zeidan B, Taylor MG, Bickers B, Al-Ruwaili J, Aukim-Hastie C and Townsend PA: Proteomics in prostate cancer biomarker discovery. Expert Rev Proteomics 7: 93-102, 2010.
16. Schaaij-Visser TB, Brakenhoff RH, Leemans CR, Heck AJ and Slijper M: Protein biomarker discovery for head and neck cancer. J Proteomics 73: 1790-1803, 2010.
17. Zhang B, Barekati Z, Kohler C, Radpour R, Asadollahi R, Holzgreve W and Zhong XY: Proteomics and biomarkers for ovarian cancer diagnosis. Ann Clin Lab Sci 40: 218-225, 2010.
18. Issaq HJ, Waybright TJ and Veenstra TD: Cancer biomarker discovery: opportunities and pitfalls in analytical methods. Electrophoresis 32: 967-975, 2011.
19. Sawyers CL: The cancer biomarker problem. Nature 452: 548-552, 2008.
20. Liu F, Wang H and Li J: An integrated bioinformatics analysis of mouse testis protein profiles with new understanding. BMB Rep 44: 347-351, 2011.
21. Dinicola S, D'Anselmi F, Pasqualato A, Proietti S, Lisi E, Cucina A and Bizzarri MA: Systems biology approach to cancer: fractals, attractors, and non-linear dynamics. OMICS 15: 93-104, 2011.

22. Coradini D, Casarsa C and Oriana S: Epithelial cell polarity and tumorigenesis: new perspectives for cancer detection and treatment. Acta Pharmacol Sin 32: 552-564, 2011.
23. Martin TA, Mason MD and Jiang WG: Tight junctions in cancer metastasis. Front Biosci 16: 898-936, 2011.
24. Kim SH, Turnbull J and Guimond S: Extracellular matrix and cell signalling: the dynamic cooperation of integrin, proteoglycan and growth factor receptor. J Endocrinol 209: 139-151, 2011.
25. Wijnhoven BP, Dinjens WN and Pignatelli M: E-cadherin-catenin cell-cell adhesion complex and human cancer. Br J Surg 87: 992-1005, 2000.
26. Berx G and Van Roy F: The E-cadherin/catenin complex: an important gatekeeper in breast cancer tumorigenesis and malignant progression. Breast Cancer Res 3: 289-293, 2001.
27. Tan C, Costello P, Sanghera J, Dominguez D, Baulida J, de Herreros AG and Dedhar S: Inhibition of integrin linked kinase (ILK) suppresses beta-catenin-Lef/Tcf-dependent transcription and expression of the E-cadherin repressor, snail, in APC-/- human colon carcinoma cells. Oncogene 20: 133-140, 2001.
28. Bravou V, Klironomos G, Papadaki E, Taraviras S and Varakis J: ILK over-expression in human colon cancer progression correlates with activation of beta-catenin, down-regulation of E-cadherin and activation of the Akt-FKHR pathway. J Pathol 208: 91-99, 2006.
29. Cortez V, Nair BC, Chakravarty D and Vadlamudi RK: Integrin-linked kinase 1: role in hormonal cancer progression. Front Biosci (Schol Ed) 3: 788-796, 2011.
30. da Cunha JP, Galante PA, de Souza JE, de Souza RF, Carvalho PM, Ohara DT, Moura RP, Oba-Shinja SM, Marie SK, Silva WA Jr, *et al*: Bioinformatics construction of the human cell surfaceome. Proc Natl Acad Sci USA 106: 16752-16757, 2009.
31. Grigoriadis A, Caballero OL, Hoek KS, da Silva L, Chen YT, Shin SJ, Jungbluth AA, Miller LD, Clouston D, Cebon J, *et al*: CT-X antigen expression in human breast cancer. Proc Natl Acad Sci USA 106: 13493-13498, 2009.
32. Wei X, Moncada-Pazos A, Cal S, Soria-Valles C, Gartner J, Rudloff U, Lin JC, NISC Comparative Sequencing Program, Rosenberg SA, López-Otín C and Samuels Y: Analysis of the disintegrin-metalloproteinases family reveals ADAM29 and ADAM7 are often mutated in melanoma. Hum Mutat 32: E2148-E2175, 2011.
33. Polanski M and Anderson NL: A list of candidate cancer biomarkers for targeted proteomics. Biomark Insights 1: 1-48, 2007.