G&I Genomics & Informatics

eISSN 2234-0742
Genomics Inform 2016;14(1):2-11
http://dx.doi.org/10.5808/GI.2016.14.1.2

REVIEW ARTICLE

Network-Based Protein Biomarker Discovery Platforms

Minhyung Kim, Daehee Hwang*

Department of New Biology and Center for Plant Aging Research, Institute for Basic Science,
Daegu Gyeongbuk Institute of Science and Technology, Daegu 42988, Korea

The advances in mass spectrometry-based proteomics technologies have enabled the generation of global proteome data from tissue or body fluid samples collected from a broad spectrum of human diseases. Comparative proteomic analysis of global proteome data identifies and prioritizes the proteins showing altered abundances, called differentially expressed proteins (DEPs), in disease samples, compared to control samples. Protein biomarker candidates that can serve as indicators of disease states are then selected as key molecules among these proteins. Recently, it has been addressed that cellular pathways can provide better indications of disease states than individual molecules and also network analysis of the DEPs enables effective identification of cellular pathways altered in disease conditions and key molecules representing the altered cellular pathways. Accordingly, a number of network-based approaches to identify disease-related pathways and representative molecules of such pathways have been developed. In this review, we summarize analytical platforms for network-based protein biomarker discovery and key components in the platforms.

Keywords: biomarkers, LC-MS/MS, network analysis, proteomics

Introduction

For a last decade, mass spectrometry (MS)-based proteomic technologies have emerged as a core technology to measure protein expression and the sites with post-translational modifications (PTMs) on large scales. However, MS-based proteome analysis has been possible for only a limited number of proteins and also limited in accurately detecting low abundant proteins and the peptides with low abundant PTMs [1-4]. Recent advances in high-resolution peptide separation, comprehensive fractionation, and high performance MS considerably improved the proteome size and depth (increased numbers of proteins and PTM sites measured) and also the accuracy in quantitative information of proteomic data (abundances of proteins and PTM sites) [5]. Moreover, MS-based proteome analysis required large amounts of samples to measure reliably proteins and PTM sites. To resolve this problem, the methods for serial enrichments of different PTMs from the same sample have been developed to significantly reduce the sample amount required [6]. These advanced MS-based proteomic technologies have facilitated the generation of proteome and PTM profiles in tissue and body fluid (plasma/serum, urine, ascites, cerebrospinal fluid, synovial fluid, saliva, and tear) samples collected from the patients with a broad spectrum of human diseases.

Comparative proteomic analysis of samples from the patients and healthy control subjects is commonly applied to identify protein biomarker candidates [7-10]. In this analysis, proteome profiles are first obtained for tissue or body fluid samples collected from the patients with a target disease and also healthy subjects, and the proteins showing altered abundances, called differentially expressed proteins (DEPs), in disease samples, compared to control samples are then identified (discovery phase) [10]. In the protein biomarker discovery platforms, about hundred DEPs are selected as an initial set of biomarker candidates, and their altered expression are then verified in a small cohort of the patients, which is independent to the patients used in the discovery phase (verification phase) [5]. However, the number of DEPs is often larger than 100, and how to select the initial set of biomarker candidates is not straightforward, although fold-changes and/or associations of the DEPs with pathophysiological processes in the target disease can be used as the criteria for selection of the initial set of biomarker candidates. Next, for the initial candidates whose disease-

Received November 25, 2015; Revised January 6, 2016; Accepted January 7, 2016
*Corresponding author: Tel: +82-54-279-2393, Fax: +82-54-279-8409, E-mail: dhhwang@postech.ac.kr

related alterations in their abundances were confirmed in the small cohort during the verification phase, their validity as biomarkers is tested in a large cohort of the patients (validation phase) [5]. Finally, the biomarker candidates whose altered expression was confirmed in the large cohort are selected as the final set of biomarkers.

Recently, a number of studies have reported that cellular pathways can serve as better indicators of disease states than individual molecules [11-17]. Pathway enrichment analysis provides a list of cellular pathways enriched by the DEPs, and a set of cellular pathways related to pathophysiological processes can be then selected as altered cellular pathways in the target disease. Next, we can focus on a subset of DEPs that are involved in the selected cellular pathways. Furthermore, network analysis for this subset of DEPs enables effective identification of key molecules that represent the selected cellular pathways as protein biomarker candidates. Using this approach, an initial set of biomarker candidates can be selected more effectively than the conventional criteria, such as fold-changes and associations of the DEPs with the target disease. Thus, several network-based approaches to identify disease-related cellular pathways and representative proteins of such pathways have been developed. In this review, we summarize bioinformatics methods for network-based protein biomarker discovery and also key components in these network-based methods.

Results

Peptide and protein identification

After obtaining proteomic data from tissue or body fluid samples using liquid chromatography–tandem mass spectrometry (LC-MS/MS) analysis, the tandem mass spectrometry (MS/MS) spectra are first searched against a protein sequence database (e.g., SWISS-Prot or UniProt) to identify the peptide sequences for individual MS/MS spectra (peptide/protein identification) (Fig. 1). The detected proteome size determines the depth to understand disease-related cellular networks based on the proteome data. Thus, a sufficient size of the detected proteome is a prerequisite in effective discovery of protein biomarkers based on disease-related networks. To ensure an adequate proteome size, accurate identification of the peptides for MS/MS spectra is important through database search using the engine, such as Mascot [18], SEQUEST [19], MS-GF+ [20], or Paragon (Table 1) [21]. During the database search, a score is assigned by the search engine to each peptide that can be generated from the sequence database, which reflects the degree of the agreement of the measured MS/MS spectrum to the theoretical spectrum of the peptide. For example, SEQUEST assigns the scores of X-corr, deltaCorr, SPrank, and SP value to the peptide, whereas MS-GF+ assigns $-\log_{10}$(E-value) to the peptide. For each measured MS/MS spectrum, the peptide with the largest score value is then selected, which defines a

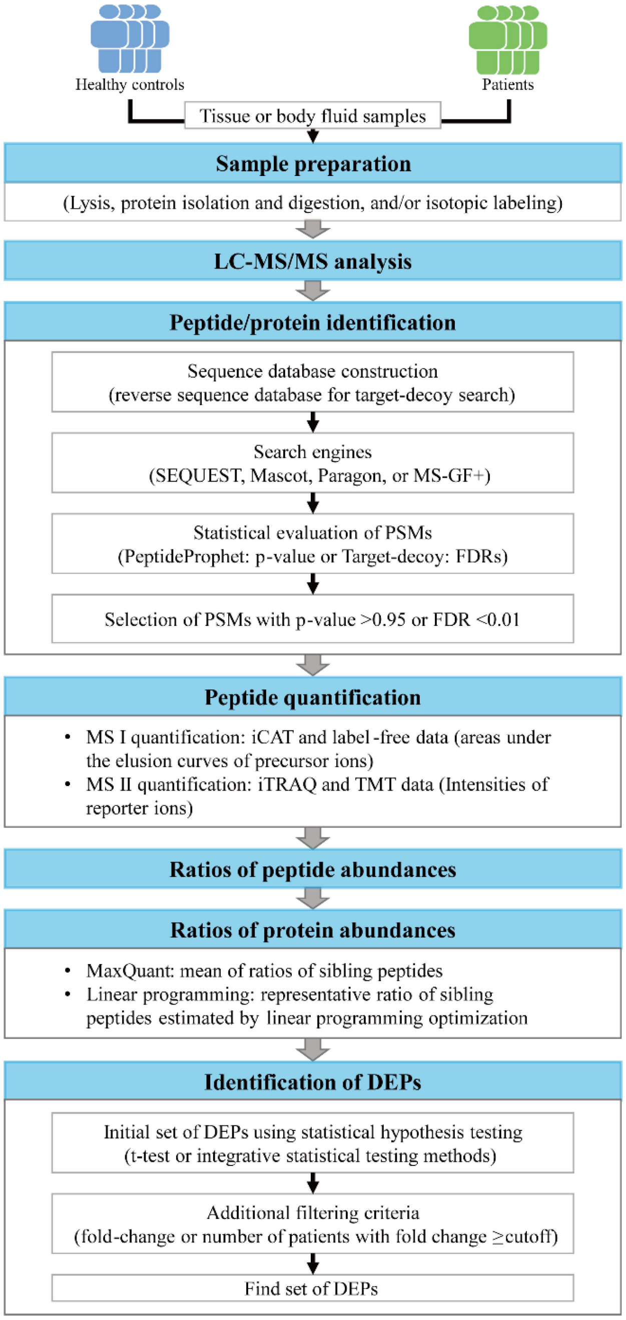
Healthy controls
Patients
Tissue or body fluid samples
Sample preparation
(Lysis, protein isolation and digestion, and/or isotopic labeling)
LC-MS/MS analysis
Peptide/protein identification
Sequence database construction
(reverse sequence database for target-decoy search)
Search engines
(SEQUEST, Mascot, Paragon, or MS-GF+)
Statistical evaluation of PSMs
(PeptideProphet: p-value or Target-decoy: FDRs)
Selection of PSMs with p-value >0.95 or FDR <0.01
Peptide quantification
• MS I quantification: iCAT and label-free data (areas under the elusion curves of precursor ions)
• MS II quantification: iTRAQ and TMT data (Intensities of reporter ions)
Ratios of peptide abundances
Ratios of protein abundances
• MaxQuant: mean of ratios of sibling peptides
• Linear programming: representative ratio of sibling peptides estimated by linear programming optimization
Identification of DEPs
Initial set of DEPs using statistical hypothesis testing
(t-test or integrative statistical testing methods)
Additional filtering criteria
(fold-change or number of patients with fold change ≥cutoff)
Find set of DEPs

Fig. 1. Basic data analysis pipeline for liquid chromatography–tandem mass spectrometry (LC-MS/MS) data. This pipeline includes peptide/protein identification, peptide/protein quantification, and identification of differentially expressed proteins (DEPs). Each of bioinformatics analyses in the pipeline is schematically shown together with the concepts and the tools. PSMs, peptide-spectrum matches; FDRs, false discovery rates; iCAT, isotope-coded affinity tag; iTRAQ, isobaric tag for relative and absolute quantitation; TMT, tandem mass tags.

Table 1. Resources for network-based protein biomarker discovery

	Tool	Website
Peptide and protein identification	SWISS-Prot or UniProt	http://www.ebi.ac.uk/uniprot
	Mascot	http://www.matrixscience.com/
	SEQUEST	http://fields.scripps.edu/sequest/
	MS-GF+	http://bix-lab.ucsd.edu/pages/viewpage.action?pageId=13533355
	Paragon	http://sciex.com/products/software/proteinpilot-software
	PeptideProphet	http://peptideprophet.sourceforge.net/
	Trans-Proteomic Pipeline (TPP)	http://www.proteomecenter.org/software.php
	Compid	http://compid.aavalla.net/
	MSblender	http://www.marcottelab.org/index.php/MSblender
Protein quantitation	MaxQuant	http://www.coxdocs.org/doku.php?id=maxquant:start
Functional enrichment analysis	Kyoto Encyclopedia of Genes and Genomes (KEGG)	http://www.genome.jp/kegg/
	DAVID	http://david.ncifcrf.gov/
	PANTHER	http://pantherdb.org/
	MetaCore	http://portal.genego.com/
	Ingenuity Pathway Analysis (IPA, QIAGEN Redwood City)	http://www.qiagen.com/ingenuity
	Gene set enrichment analysis (GSEA)	http://software.broadinstitute.org/gsea/index.jsp
	Signaling pathway impact analysis (SPIA)	http://vortex.cs.wayne.edu/projects.htm
Network modeling and analysis	Human protein reference database (HPRD)	http://www.hprd.org/
	Biological general repository for interaction datasets (BioGRID)	http://thebiogrid.org/
	Biomolecular interaction network database (BIND)	http://metadatabase.org/wiki/BIND_-_Biomolecular_Interaction_Network_Database
	Search tool for recurring instances of neighbouring genes (STRING)	http://string-db.org/
	Molecular INTeraction database (MINT)	http://mint.bio.uniroma2.it/mint/
	EdgeExpressDB (FANTOM4-EEDB)	http://fantom.gsc.riken.jp/4/edgeexpress/about/
	Transcriptional regulatory element database (TRED)	http://cb.utdallas.edu/cgi-bin/TRED/tred.cgi?process=home
	Molecular signatures database (MSigDB)	http://software.broadinstitute.org/gsea/msigdb/index.jsp
	jActiveModules	http://apps.cytoscape.org/apps/jactivemodules
	ResponseNet	http://netbio.bgu.ac.il/respnet/
	NetWalker	https://netwalkersuite.org/
	clusterMaker	http://apps.cytoscape.org/apps/clustermaker
Integrative analysis of proteomic data with other global data	HotNet	http://compbio.cs.brown.edu/projects/hotnet/
	SteinerNet	http://fraenkel.mit.edu/steinernet/

peptide-spectrum match (PSM).

Next, the statistical significance of the PSMs resulted from the database search is evaluated using a statistical method, such as PeptideProphet [22] in the Trans-Proteomic Pipeline [23] or the target-decoy method (Table 1) [24]. For example, for each PSM, PeptideProphet combines the scores (X-corr, deltaCorr, SPrank, and SP value) to an F score, and a mixture of Gaussian and Gamma distributions is fitted to the distribution of F scores for all PSMs. Using the mixture distribution, the probability of a PSM being true (PeptideProphet p-value) is estimated [25]. Finally, the PSMs with PeptideProphet p-value > 0.95 are selected as the correct PSMs. In the target-decoy method, a reverse sequence database was generated by inversing the reference protein sequence in the database and then included in the database prior to the database search. Using the score values for the PSMs in which the peptide sequence was obtained from the reverse sequence database, false discovery rates (FDRs) for all PSMs are estimated [24]. Finally, the PSMs with FDR < 0.01 are selected as the correct PSMs.

Recent studies have shown that the best coverage of the detected proteome can be obtained by combining the outputs from multiple search engines. For example, Compid enables the integration of Paragon and Mascot (Table 1) [24] by assigning the peptides with higher scores from the two search engines to the MS/MS spectra, which can leads to unreliable false positive rates. Also, MSblender integrates the search scores from the search engines into a probability score for every possible PSM and then estimates FDRs for the PSMs in a reliable manner [26]. This method identifies more PSMs than any single search engine at the same FDR. After identifying all the PSMs from the database search, a list of the detected proteins are further identified. Using the SEQUEST search outputs, the probability that a protein is correctly identified (ProteinProphet p-value) is calculated by statistically combining the PeptideProphet p-values of the peptides derived from the protein, and the proteins with PeptideProphet p-value $>$ 0.99 are selected as the proteins correctly identified [22]. In the target decoy method, the proteins with more than two non-redundant peptides with FDR $<$ 0.01 can be considered as the reliable proteins [27].

Protein quantitation

Several LC-MS/MS approaches provide quantitative information of the peptides identified from the database search (protein quantification) (Fig. 1). First, prior to LC-MS/MS analysis, the peptides can be labeled using the isotopic agents, such as isotope-coded affinity tag (iCAT) [28], isobaric tag for relative and absolute quantitation (iTRAQ) [29], or tandem mass tags (TMT) [30]. For the iCAT data, the abundance of an identified peptide is estimated as the area under the elusion curve of the peptide. On the other hand, for the iTRAQ or TMT data, the peptide abundance is estimated as the intensities of the reporter ions in the MS/MS data [31]. Second, when no isotope labeling was done, called label-free LC-MS/MS analysis, the peptide abundance is estimated as the area under the elusion curve of the peptide as done for the iCAT data. Next, the relative abundance of an identified peptide is calculated as the ratio of the peptide abundances between patient and control samples. From the iCAT data, the ratios of the heavy and light peptides are calculated as the relative peptide abundance. To estimate the relative abundances of a protein, a list of the peptides derived from the proteins are identified and the relative abundances of these peptides are then combined using a method for protein quantitation, such as the quantification tool in MaxQuant software [32] or a linear programming method [33].

Identification of differentially expressed proteins

The biomarkers should reflect the alteration of disease-related pathophysiological processes in the patient samples, compared to controls. Thus, the DEPs between control and patient samples are identified as the biomarker candidates (identification of DEPs) (Fig. 1). A number of statistical methods have been developed for identification of the DEPs, such as t-test [34] or integrative statistical methods [35, 36],

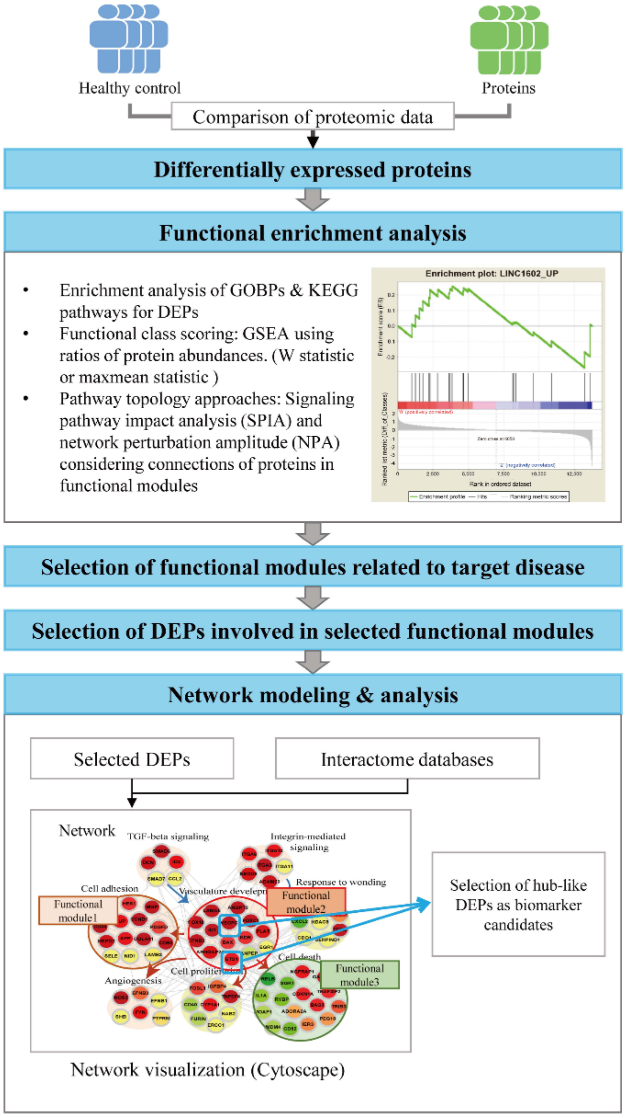
Healthy control
Proteins
Comparison of proteomic data
Differentially expressed proteins
Functional enrichment analysis
• Enrichment analysis of GOBPs & KEGG pathways for DEPs
• Functional class scoring: GSEA using ratios of protein abundances. (W statistic or maxmean statistic)
• Pathway topology approaches: Signaling pathway impact analysis (SPIA) and network perturbation amplitude (NPA) considering connections of proteins in functional modules
Enrichment plot: LINC1602_UP
Enrichment score (ES)
Ranked list metric (PreRanked)
Rank in ordered dataset
Zero cross at 6659
'1' (positively correlated)
'2' (negatively correlated)
Enrichment profile — Hits — Ranking metric scores
Selection of functional modules related to target disease
Selection of DEPs involved in selected functional modules
Network modeling & analysis
Selected DEPs
Interactome databases
Network
TGF-beta signaling
Integrin-mediated signaling
Response to wonding
Cell adhesion
Vasculature developm
Functional module1
Functional module2
Functional module3
Cell death
Cell prolifer
Angiogenesis
Selection of hub-like DEPs as biomarker candidates
Network visualization (Cytoscape)

Fig. 2. Schematics for functional enrichment and network analyses of differentially expressed proteins (DEPs). Three types of functional enrichment analyses are shown. The output from gene set enrichment analysis (GSEA) is displayed as an example. Also, an example of network model is shown, which includes functional modules 1–3 (cell adhesion, vasculature development, and cell death, respectively). Node colors in the network model represent up- (red) and down-regulation (green), the color gradient denotes $\log_2$-fold-changes of the proteins, and gray lines represent the connections between the nodes. GOBPs, gene ontology biological processes; KEGG, Kyoto Encyclopedia of Genes and Genomes; TGF, tumor growth factor.

with the multiple comparison correction. These methods compare the relative protein abundances between control and patient samples and then estimate the significance (p-value or FDR) of the difference in the relative protein abundances between control and patient samples methods. The proteins with p-value or FDR < 0.01 or 0.05 are then selected as the DEPs. Furthermore, additional constraints are used to select more reliable DEPs. For example, for each DEP, the number of the patients showing the fold-changes larger than a cutoff (1.5- or 2-fold) is counted, and a subset of DEPs with the number of patients larger than a certain percentage (50% of the patients) can be selected as the reliable DEPs [37]. Using this criterion, we can focus on the DEPs that are likely to show their alterations in the abundance in at least more than half of the patients when they were used as biomarker candidates for newly collected patient samples.

Functional enrichment analysis

The alterations in protein abundances under disease conditions reflect simple clinical symptoms that are commonly observed in many other diseases, such as inflammation or immune responses, or the alterations of the pathophysiological processes specific to the target disease, such as aggregation of toxic proteins in neurodegenerative diseases. Thus, the DEPs that are involved in the disease-related pathophysiological processes can represent more effectively the alterations specific to the target disease (Fig. 2). To effectively understand the pathophysiological processes altered under disease conditions, functional enrichment analysis is often performed for the DEPs using diverse enrichment tools. For example, the enrichment analysis of gene ontology biological processes (GOBPs) or Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways can be applied to the DEPs using DAVID [38] and PANTHER [39], and commercial tools, such as MetaCore [40] and Ingenuity Pathway Analysis (IPA, QIAGEN Redwood City, http://www.qiagen.com/ingenuity) (Table 1). A list of GOBPs or KEGG pathways with enrichment p-value < 0.05 are selected as the pathophysiological processes or cellular pathways altered under disease conditions. Of these selected GOBPs and KEGG pathways, a subset of the GOBPs related to the target disease is further selected based on the knowledge of the target disease. Finally, a subset of the DEPs that are involved in the selected GOBPs or KEGG pathways related to the target disease can be selected as an initial set of biomarker candidates, considering that these DEPs are likely to be specifically associated with the target disease.

No enrichment analysis mentioned above consider quantitatively the differences in fold-changes of the DEPs, but consider equivalently the DEPs that were selected using a fixed cutoff (p-value and/or fold-changes). To remedy this problem, several functional class scoring methods have been developed, such as gene set enrichment analysis (GSEA) (Table 1, Fig. 2) [41]. In the GSEA, all D-changes, and the enrichment of functional modules defined in the database, such as the molecular signatures database (MSigDB) (Table 1), by the top or bottom of the ranked protein list is statistically evaluated. Several alternative module-level statistics to the conventional GSEA have been also used, including Kolmogorov-Smirnov statistic and the maxmean statistic [42]. Moreover, as an alternative enrichment analysis, several pathway topology approaches have been developed, such as signaling pathway impact analysis (SPIA) [43] and network perturbation amplitude (NPA) (Table 1, Fig. 2) [26]. These approaches consider whether the proteins involved in functional modules defined by GOBP or GSEA database (MSigDB) interact with each other in cellular networks. In the SPIA, a conventional overrepresentation measure, as in GOBP enrichment analysis and a topology measure of the pathway are combined to identify functional modules in which the DEPs are significantly connected. In the NPA, the upper and lower tiers are defined as cellular pathways (e.g., mitogen-activated protein kinase pathway activation) and target genes/proteins regulated by the pathways in the upper tier, respectively, and positive and negative causal relationships are defined by the links between the upper and lower tiers. Then, NPA evaluates the causal relationships by quantitatively summarizing whether the expression changes of the downstream nodes are consistent with pathway activation, inactivation, or no change. These analyses provide a list of functional modules significantly altered under disease conditions. Similar to the case of the GOBP enrichment analysis, a subset of function modules related to the target disease can be identified, and a subset of DEPs involved in the identified functional modules can be selected as an initial set of biomarker candidates.

Network modeling and analysis

Functional molecules or cellular networks altered under disease conditions interact closely with each other to form a disease-perturbed cellular network. The network model enables identification of hub-like molecules that can serve as core indicators of the activities of cellular processes described in the network model. Thus, it is important to understand the disease-perturbed cellular network to identify the core indicators as the biomarker candidates. The core molecules and their associated cellular pathways have been further suggested as therapeutic targets or the biomarker candidates that can be used to evaluate the efficacy of the treatments (e.g., drug efficacy) for the target disease. Hence, the network-based approaches have been employed for protein biomarker discovery.

The network-based approaches first reconstruct a disease-perturbed cellular network model describing the interactions of the DEPs or a subset of the DEPs involved in the selected GOBPs related to the target disease using protein-protein (PPIs), protein-DNA (PDIs), and/or protein-metabolite interactions (PMIs) in the following interactome databases (Table 1, Fig. 2): (1) PPI databases: human protein reference database (HPRD) [44], biological general repository for interaction datasets (BioGRID) [45], biomolecular interaction network database (BIND) [46], search tool for recurring instances of neighbouring genes (STRING) [47], and Molecular INTeraction database (MINT) [48]; (2) PDI databases: EdgeExpressDB (FANTOM4-EEDB) [49], transcriptional regulatory element database (TRED) [50], MSigDB [41], MultiNet [51], and MetaCore [40]; and (3) PMI database: KEGG pathway database [52].

Next, the network model are analyzed to identify network modules or clusters each of which includes a set of the nodes densely connected in the network (Fig. 2). For network clustering or modularization, a number of methods, such as jActiveModules [53], ResponseNet [54], NetWalker [55], and clusterMaker [56], have been developed (Table 1). These methods can be categorized into two groups. One group, such as clusterMaker, searches for the node clusters based on the network topology information such that the nodes in the same cluster have dense connections, but sparse connections between the different clusters [56]. The other group, such as jActiveModules, uses the quantitative fold-changes of the nodes between control and patient samples measured by LC-MS/MS together with the network topology information. In this group, the node clusters are identified such that the nodes in the same clusters are densely connected and also show similar alteration patterns in protein abundance [53].

For functional interpretation of the node clusters resulted from network clustering, functional enrichment analyses mentioned above are then performed for the nodes in each network cluster. Of the node clusters, prior to the enrichment analysis, we can focus on the major network clusters including large numbers of the DEPs. The functional enrichment analysis then provide a subset of the major network clusters whose functions (e.g., GOBPs or KEGG pathways) are associated with the disease-related cellular processes. This subset can be considered as potential determinants of the disease-perturbed cellular network. Finally, the hub-like DEP in each major network cluster is selected as a biomarker candidate that can reflect the perturbation of the disease-related cellular process associated with the network cluster. Mitra *et al.* [57] demonstrated the power of the network-based biomarker discovery approach in prediction of metastatic cancers in human breast cancer. They identified network clusters with dysregulated expression using SigArSearch and showed that the network clusters were more accurate in distinguishing metastatic cancers from non-metastatic cancers, compared with individual cancer-gene markers.

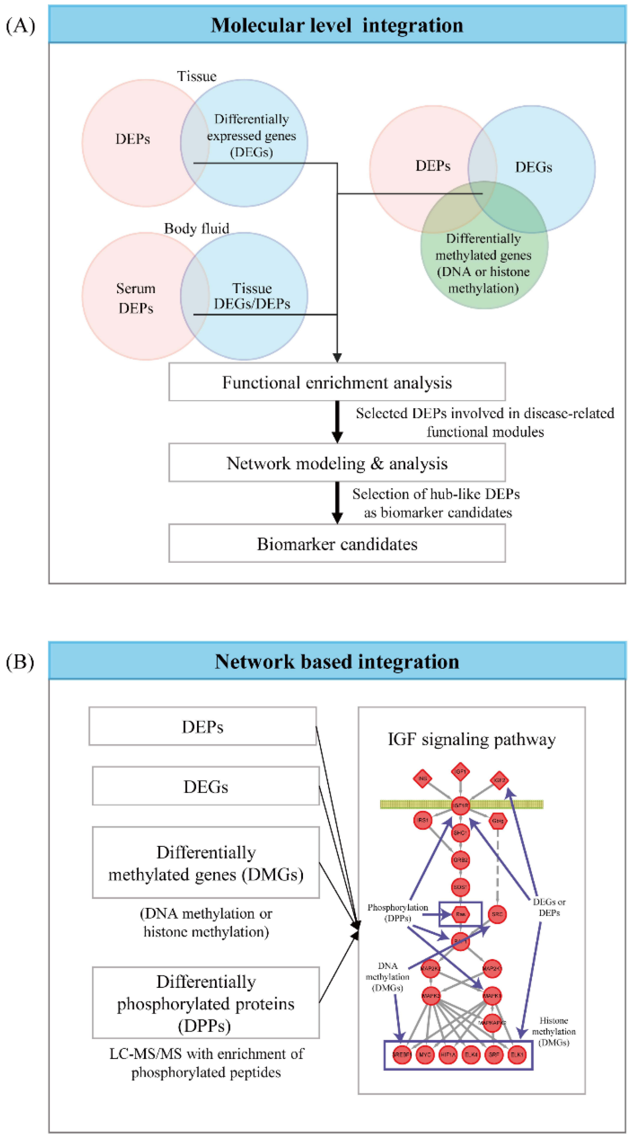
(A)
Molecular level integration
Tissue
DEPs
Differentially expressed genes (DEGs)
Body fluid
Serum DEPs
Tissue DEGs/DEPs
DEPs
DEGs
Differentially methylated genes (DNA or histone methylation)
Functional enrichment analysis
Selected DEPs involved in disease-related functional modules
Network modeling & analysis
Selection of hub-like DEPs as biomarker candidates
Biomarker candidates
(B)
Network based integration
DEPs
DEGs
Differentially methylated genes (DMGs)
(DNA methylation or histone methylation)
Differentially phosphorylated proteins (DPPs)
LC-MS/MS with enrichment of phosphorylated peptides
IGF signaling pathway
INS
IGF1
IGF2
IGF1R
IRS1
Gbg
SHC1
GRB2
SOS1
Ras
SRC
RAF1
MAP2K2
MAP2K1
MAPK3
MAPK1
MAPKAPK2
SREBF1
MYC
HIF1A
ELK4
SRF
ELK1
Phosphorylation (DPPs)
DNA methylation (DMGs)
DEGs or DEPs
Histone methylation (DMGs)

Fig. 3. Integrative analysis of proteomic data with other types of global data. The concepts for molecular level (A) and network based integration (B) are shown. In the top panel, two example cases (one for tissue data and the other for serum data) are illustrated, and also one example case for integrative analysis of three types of global datasets (proteomic, mRNA expression, and DNA/histone methylation data) is shown. In the bottom panel, network based integration of four types of global data using differentially expressed molecules from them are schematically displayed. DEPs, differentially expressed proteins; IGF, insulin-like growth factor; LC-MS/MS, liquid chromatography–tandem mass spectrometry.

Integrative analysis of proteomic data with other global data

Other global data than proteomic data, such as mRNA expression, mutation, DNA methylation, histone modification, can provide complementary information that enables to select reliable biomarker candidates when they are integrated with the proteomic data. The integrative analysis of proteomic data with other types of global datasets can be categorized into two groups (Fig. 3): (1) the molecular level and (2) the network-level integration methods [58]. The proteome data are contaminated with the noises coming from numerous sources of technical variability during LC-MS/MS experiments and biological variability in patient samples during sample collection and preparation [59, 60]. The integration of disparate global datasets reduces the artifacts of the noises in selection of the DEPs by enabling us to focus on the reliable DEPs showing consistent changes between control and patient samples in multiple types of global data [58].

In the molecular level integration method, the DEPs that show consistent alterations in the other global datasets (e.g., mRNA expression data) are searched, assuming that they are more reliable indicators of the perturbation of cellular processes under disease conditions (Fig. 3). Biomarker candidates can be then selected from these reliable DEPs. For example, the DEPs selected from the tissue data can be supported by the consistency in the altered expression of their corresponding mRNAs. Also, the DEPs can be further supported by the consistency in the changes of copy number variations or DNA and/or histone methylations of the corresponding genes. Moreover, the reliability of the DEPs selected from the serum proteome data can be supported by the altered expression of their corresponding mRNAs or proteins in the tissues from the target organ. Hyung *et al.* [61] demonstrated the value of the integrative analysis of serum proteome data with mRNA expression and proteome data obtained from the tissues in human breast cancers. They selected the DEPs between the serum samples collected from sensitive and resistant patients to a combinatorial chemotherapy using doxorubicin and docetaxol and then further selected the DEPs with altered expression of the corresponding mRNAs and proteins in the breast tissue samples. During the validation of biomarker candidates, they showed that the DEPs with the consistent alteration in the tissues showed higher accuracy in their validation using western blotting for a validation cohort of independent serum samples.

Individual patients can show the variation in altered molecules of a cellular pathway though the same pathway is consistently perturbed under disease conditions. Given the variation, these molecules cannot be selected as DEPs using the conventional statistical methods, thereby leading to the failure to identify biomarkers that reflect the alteration of such cellular pathways. Thus, it has been addressed that cellular pathways can serve as a more reliable indicator of the altered cellular processes under disease conditions, compared to individual molecules [58]. Different types of molecules (mRNA/proteins, microRNAs, DNA methylations, and metabolites) can represent distinct layers of cellular networks. Thus, a cellular network can be modeled in a multi-layered network where each layer can be delineated by a distinct type of the molecules. For example, transcriptional and microRNA regulatory networks are defined by interactions of transcription factors and microRNAs, respectively, with their target mRNAs. Also, cellular signaling networks are defined by protein-protein interactions (kinase- substrate interactions), while DNA methylation networks are defined by interactions of methyltransferases and demethylases with their target DNAs and mRNAs. The integration of multiple global datasets for different types of molecules enables us to decode the multi-layered cellular networks associated with the target disease. A number of the tools have been developed to understand the subnetworks (network clusters) of the multi-layered networks whose perturbations are collectively indicated by different types of global datasets, including IPA (QIAGEN Redwood City, http://www.qiagen.com/ingenuity), HotNet [62], or SteinerNet (Table 1) [63]. Although these tools can be applied for integration of proteomic data with other types of global data, they have been used mainly for integration of disparate genomic and transcriptomic data. Recently, Shi *et al.* [64] developed NetGestalt that can be used for integration of multi-dimensional global datasets including proteomic data. For example, using NegGestalt, Zhu *et al.* [65] identified KRAS and AKAP12 subnetworks that can play important roles in pathogenesis of colorectal cancers by integrating proteomic data with mutation, copy number variation, DNA methylation, and mRNA expression data generated from colon tissues and cells. Finally, the hub-like molecules in the subnetworks are selected as key indicators of the alterations of the disease-related processes associated with the subnetworks. For example, Iliopoulos *et al.* [66] identified a set of the biomarker candidates by performing the network- based integration of the proteome and microRNA expression data. Also, The Cancer Genome Atlas (TCGA) research network identified the notch signaling pathway as a key molecule representing the pathogenesis of ovarian cancers by integrating somatic mutation, copy number variation, and mRNA expression data using HotNet analysis [67].

Discussion

In this review, we summarized a battery of bioinformatics analyses for network-based biomarker discovery using LC-MS/MS data. These analyses include peptide/protein identification using database search engines, peptide/protein quantification from MS or MS/MS data, identification of DEPs using statistical methods, functional enrichment analysis of DEPs using diverse enrichment tools, and network modeling and analysis of DEPs, as well as integrative analysis of disparate global datasets with the LC-MS/MS data.

Peptide identification using database search has been known as a main challenge. Of the MS/MS spectra measured, about half of them are commonly mapped to peptide sequences. The inclusion of PTMs in database search can map the unidentified MS/MS spectra to peptide sequences. However, the PTM inclusion exponentially increases the search time and thus it would be practically impossible to include the most common five PTMs, phosphorylation, glycosylation, ubiquitination, methylation, and acetylation in database search. Recently, although several search tools, such as MODa [68], with improved search speeds have been developed, there is still significant needs for efficient search tools that can include the five common PTMs in the search.

Moreover, protein quantification has been also one of the challenges in MS-based proteomic analysis. Relative protein abundances are estimated by combining the ratios of abundances of the peptides derived from the protein. The same peptide sequences can be derived from multiple proteins due to the redundancy in their sequences, which can cause the discrepancy in the ratios of the peptides from the same protein between control and patient samples. Also, technical and biological variability can add the discrepancy to the peptide ratios. Because of these sources of the discrepancy in the peptide ratios, it has been difficult to estimate the relative ratios of protein abundances between control and patient samples. Recently, although the optimization-based quantification tools, such as linear programming [33], have been developed, there have been significant needs for the quantification methods with improved accuracy.

Due to the issue with protein quantification, identification of DEPs can be also erroneous when inaccurate ratios of protein abundances exist. Additional criteria, such as the number of patients showing fold-changes larger than a cutoff (e.g., 1.5-fold), are used to reduce the error (false-positives) in identification of DEPs. The error in identification of DEPs can further propagate into functional enrichment analysis and network modeling and analysis for the DEPs. Finally, the integrative analysis of disparate global datasets suffers from different scales of fold-changes and the size of detected molecules in the different types of global datasets. For example, mRNA fold-changes between control and patient samples can be relatively larger than the fold-changes of protein abundances, which may cause the bias toward mRNA fold-changes during the integration of fold-changes of mRNA and protein abundances. Also, differentially expressed microRNAs between control and patient samples are smaller than the DEPs or mRNAs, which can lead to the bias toward the DEPs or mRNAs during the integration. Various integration strategies have been proposed to reduce the intrinsic bias from the differences in the size of disparate global datasets. For example, in these methods, mRNA and protein abundance ratios are first integrated and the data with small sizes are separately integrated to the outputs from the initial integration of mRNA and protein data. However, regarding the scale difference in fold-changes, there have been the needs for the algorithms that can effectively normalize the different scales of fold-changes in different datasets.

Acknowledgments

This work was supported by the grant from Institute for Basic Science (IBS-R013-G1-2015-a00) funded by the Korean Ministry of Science, ICT & Future Planning.

References

1. Biomarkers Definitions Working Gruop. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin Pharmacol Ther* 2001;69:89-95.
2. Hawkridge AM, Muddiman DC. Mass spectrometry-based biomarker discovery: toward a global proteome index of individuality. *Annu Rev Anal Chem (Palo Alto Calif)* 2009;2:265-277.
3. Aebersold R, Anderson L, Caprioli R, Druker B, Hartwell L, Smith R. Perspective: a program to improve protein biomarker discovery for cancer. *J Proteome Res* 2005;4:1104-1109.
4. Li D, Chan DW. Proteomic cancer biomarkers from discovery to approval: it's worth the effort. *Expert Rev Proteomics* 2014;11:135-136.
5. Frantzi M, Bhat A, Latosinska A. Clinical proteomic biomarkers: relevant issues on study design & technical considerations in biomarker development. *Clin Transl Med* 2014;3:7.
6. Anderson NL, Anderson NG. The human plasma proteome: history, character, and diagnostic prospects. *Mol Cell Proteomics* 2002;1:845-867.
7. Zubarev RA. The challenge of the proteome dynamic range and its implications for in-depth proteomics. *Proteomics* 2013;13:723-726.
8. Wang P, Whiteaker JR, Paulovich AG. The evolving role of mass spectrometry in cancer biomarker discovery. *Cancer Biol*

Ther 2009;8:1083-1094.

9. Shen J, Wang W, Wu J, Feng B, Chen W, Wang M, *et al*. Comparative proteomic profiling of human bile reveals SSP411 as a novel biomarker of cholangiocarcinoma. *PLoS One* 2012;7:e47476.
10. Chen JH, Ni RZ, Xiao MB, Guo JG, Zhou JW. Comparative proteomic analysis of differentially expressed proteins in human pancreatic cancer tissue. *Hepatobiliary Pancreat Dis Int* 2009;8: 193-200.
11. Giles RH, van Es JH, Clevers H. Caught up in a Wnt storm: Wnt signaling in cancer. *Biochim Biophys Acta* 2003;1653:1-24.
12. Leiserson MD, Blokh D, Sharan R, Raphael BJ. Simultaneous identification of multiple driver pathways in cancer. *PLoS Comput Biol* 2013;9:e1003054.
13. Moore RA, Faris R, Priola SA. Proteomics applications in prion biology and structure. *Expert Rev Proteomics* 2015;12: 171-184.
14. Aitman TJ, Boone C, Churchill GA, Hengartner MO, Mackay TF, Stemple DL. The future of model organisms in human disease research. *Nat Rev Genet* 2011;12:575-582.
15. Baranzini SE. The genetics of autoimmune diseases: a networked perspective. *Curr Opin Immunol* 2009;21:596-605.
16. Ding Y, Chen M, Liu Z, Ding D, Ye Y, Zhang M, *et al*. atBioNet: an integrated network analysis tool for genomics and biomarker discovery. *BMC Genomics* 2012;13:325.
17. Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Network-based classification of breast cancer metastasis. *Mol Syst Biol* 2007; 3:140.
18. Baker ES, Liu T, Petyuk VA, Burnum-Johnson KE, Ibrahim YM, Anderson GA, *et al*. Mass spectrometry for translational proteomics: progress and clinical implications. *Genome Med* 2012; 4:63.
19. Eng JK, McCormack AL, Yates JR. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* 1994; 5:976-989.
20. Kim S, Gupta N, Pevzner PA. Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *J Proteome Res* 2008;7:3354-3363.
21. Polaskova V, Kapur A, Khan A, Molloy MP, Baker MS. High-abundance protein depletion: comparison of methods for human plasma biomarker discovery. *Electrophoresis* 2010; 31:471-482.
22. Nesvizhskii AI, Keller A, Kolker E, Aebersold R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem* 2003;75:4646-4658.
23. Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem* 2002;74:5383-5392.
24. Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* 2007;4:207-214.
25. Kaji H, Saito H, Yamauchi Y, Shinkawa T, Taoka M, Hirabayashi J, *et al*. Lectin affinity capture, isotope-coded tagging and mass spectrometry to identify N-linked glycoproteins. *Nat Biotechnol* 2003;21:667-672.
26. Kim W, Bennett EJ, Huttlin EL, Guo A, Li J, Possemato A, *et al*. Systematic and quantitative assessment of the ubiquitin-modified proteome. *Mol Cell* 2011;44:325-340.
27. Carapito C, Lane L, Benama M, Opsomer A, Mouton-Barbosa E, Garrigues L, *et al*. Computational and Mass-Spectrometry-Based Workflow for the Discovery and Validation of Missing Human Proteins: Application to Chromosomes 2 and 14. *J Proteome Res* 2015;14:3621-3634.
28. Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol* 1999;17:994-999.
29. Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, Hattan S, *et al*. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics* 2004;3:1154-1169.
30. Thompson A, Schäfer J, Kuhn K, Kienle S, Schwarz J, Schmidt G, *et al*. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal Chem* 2003;75:1895-1904.
31. Zhang Y, Fonslow BR, Shan B, Baek MC, Yates JR 3rd. Protein analysis by shotgun/bottom-up proteomics. *Chem Rev* 2013; 113:2343-2394.
32. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* 2008;26: 1367-1372.
33. Dost B, Bandeira N, Li X, Shen Z, Briggs SP, Bafna V. Accurate mass spectrometry based protein quantification via shared peptides. *J Comput Biol* 2012;19:337-348.
34. Barallobre-Barreiro J, Didangelos A, Schoendube FA, Drozdov I, Yin X, Fernández-Caggiano M, *et al*. Proteomics analysis of cardiac extracellular matrix remodeling in a porcine model of ischemia/reperfusion injury. *Circulation* 2012;125:789-802.
35. Hwang D, Smith JJ, Leslie DM, Weston AD, Rust AG, Ramsey S, *et al*. A data integration methodology for systems biology: experimental verification. *Proc Natl Acad Sci U S A* 2005;102: 17302-17307.
36. Hwang D, Rust AG, Ramsey S, Smith JJ, Leslie DM, Weston AD, *et al*. A data integration methodology for systems biology. *Proc Natl Acad Sci U S A* 2005;102:17296-17301.
37. Tsai CF, Hsu CC, Hung JN, Wang YT, Choong WK, Zeng MY, *et al*. Sequential phosphoproteomic enrichment through complementary metal-directed immobilized metal ion affinity chromatography. *Anal Chem* 2014;86:685-693.
38. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 2009;4:44-57.
39. Mi H, Lazareva-Ulitsky B, Loo R, Kejariwal A, Vandergriff J, Rabkin S, *et al*. The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res* 2005; 33:D284-D288.
40. Ekins S, Nikolsky Y, Bugrim A, Kirillov E, Nikolskaya T. Pathway mapping tools for analysis of high content data. *Methods Mol Biol* 2007;356:319-350.
41. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide ex-

pression profiles. *Proc Natl Acad Sci U S A* 2005;102:15545-15550.

42. Beck HC, Nielsen EC, Matthiesen R, Jensen LH, Sehested M, Finn P, *et al*. Quantitative proteomic analysis of post-translational modifications of human histones. *Mol Cell Proteomics* 2006;5:1314-1325.
43. Xu G, Paige JS, Jaffrey SR. Global analysis of lysine ubiquitination by ubiquitin remnant immunoaffinity profiling. *Nat Biotechnol* 2010;28:868-873.
44. Peri S, Navarro JD, Kristiansen TZ, Amanchy R, Surendranath V, Muthusamy B, *et al*. Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res* 2004;32:D497-D501.
45. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* 2006;34:D535-D539.
46. Bader GD, Betel D, Hogue CW. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res* 2003;31:248-250.
47. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, *et al*. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* 2015;43:D447-D452.
48. Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, *et al*. MINT: the Molecular INTeraction database. *Nucleic Acids Res* 2007;35:D572-D574.
49. Severin J, Waterhouse AM, Kawaji H, Lassmann T, van Nimwegen E, Balwierz PJ, *et al*. FANTOM4 EdgeExpressDB: an integrated database of promoters, genes, microRNAs, expression dynamics and regulatory interactions. *Genome Biol* 2009;10:R39.
50. Jiang C, Xuan Z, Zhao F, Zhang MQ. TRED: a transcriptional regulatory element database, new entries and other development. *Nucleic Acids Res* 2007;35:D137-D140.
51. Feist P, Hummon AB. Proteomic challenges: sample preparation techniques for microgram-quantity protein analysis from biological samples. *Int J Mol Sci* 2015;16:3537-3563.
52. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* 2012;40:D109-D114.
53. Ideker T, Ozier O, Schwikowski B, Siegel AF. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* 2002;18 Suppl 1:S233-S240.
54. Lan A, Smoly IY, Rapaport G, Lindquist S, Fraenkel E, Yeger-Lotem E. ResponseNet: revealing signaling and regulatory networks linking genetic and transcriptomic screening data. *Nucleic Acids Res* 2011;39:W424-W429.
55. Komurov K, Dursun S, Erdin S, Ram PT. NetWalker: a contextual network analysis tool for functional genomics. *BMC Genomics* 2012;13:282.
56. Morris JH, Apeltsin L, Newman AM, Baumbach J, Wittkop T, Su G, *et al*. clusterMaker: a multi-algorithm clustering plugin for Cytoscape. *BMC Bioinformatics* 2011;12:436.
57. Mitra K, Carvunis AR, Ramesh SK, Ideker T. Integrative approaches for finding modular structure in biological networks. *Nat Rev Genet* 2013;14:719-732.
58. Wang J, Zuo Y, Man YG, Avital I, Stojadinovic A, Liu M, *et al*. Pathway and network approaches for identification of cancer signature markers from omics data. *J Cancer* 2015;6:54-65.
59. Wang J, Zhang Y, Marian C, Ressom HW. Identification of aberrant pathways and network activities from high-throughput data. *Brief Bioinform* 2012;13:406-419.
60. Oishi N, Kumar MR, Roessler S, Ji J, Forgues M, Budhu A, *et al*. Transcriptomic profiling reveals hepatic stem-like gene signatures and interplay of miR-200c and epithelial-mesenchymal transition in intrahepatic cholangiocarcinoma. *Hepatology* 2012;56:1792-1803.
61. Hyung SW, Lee MY, Yu JH, Shin B, Jung HJ, Park JM, *et al*. A serum protein profile predictive of the resistance to neoadjuvant chemotherapy in advanced breast cancers. *Mol Cell Proteomics* 2011;10:M111.011023.
62. Vandin F, Upfal E, Raphael BJ. Algorithms for detecting significantly mutated pathways in cancer. *J Comput Biol* 2011;18:507-522.
63. Huang SS, Fraenkel E. Integrating proteomic, transcriptional, and interactome data reveals hidden components of signaling and regulatory networks. *Sci Signal* 2009;2:ra40.
64. Shi Z, Wang J, Zhang B. NetGestalt: integrating multidimensional omics data over biological networks. *Nat Methods* 2013;10:597-598.
65. Zhu J, Shi Z, Wang J, Zhang B. Empowering biologists with multi-omics data: colorectal cancer as a paradigm. *Bioinformatics* 2015;31:1436-1443.
66. Iliopoulos D, Malizos KN, Oikonomou P, Tsezou A. Integrative microRNA and proteomic approaches identify novel osteoarthritis genes and their collaborative metabolic and inflammatory networks. *PLoS One* 2008;3:e3740.
67. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature* 2011;474:609-615.
68. Na S, Bandeira N, Paek E. Fast multi-blind modification search through tandem mass spectrometry. *Mol Cell Proteomics* 2012;11:M111.010199.