Review

For reprint orders, please contact: reprints@futuremedicine.com

Biomarkers in Medicine

Identifying network-based biomarkers of complex diseases from high-throughput data

Zhi-Ping Liu
Department of Biomedical Engineering, School of Control Science & Engineering, Shandong University, Jinan, Shandong 250061, China
Tel.: +86 531 8839 2280
Fax: +86 531 8839 2205
zpliu@sdu.edu.cn

In this work, we review the main available computational methods of identifying biomarkers of complex diseases from high-throughput data. The emerging omics techniques provide powerful alternatives to measure thousands of molecules in cells in parallel manners. The generated genomic, transcriptomic, proteomic, metabolomic and phenomic data provide comprehensive molecular and cellular information for detecting critical signals served as biomarkers by classifying disease phenotypic states. Networks are often employed to organize these profiles in the identification of biomarkers to deal with complex diseases in diagnosis, prognosis and therapy as well as mechanism deciphering from systematic perspectives. Here, we summarize some representative network-based bioinformatics methods in order to highlight the importance of computational strategies in biomarker discovery.

First draft submitted: 1 December 2015; Accepted for publication: 5 January 2016;
Published online: 20 January 2016

Keywords: bioinformatics • complex disease • high-throughput data • network biomarker

Disease biomarkers provide tremendously valuable information for disease diagnosis and prognosis, as well as the prediction of therapy and treatment effectiveness [1]. Biomarkers distinguishing disease states highlight the potentiality of clinical applications in complex diseases [2], such as cancers [3,4], diabetes [5] and neurodegenerative disease [6]. The early diagnosis of disease and risk assessment by biomarkers provides the precious opportunity of designing rational therapy strategy for patients [7]. Prognosis markers track the effectiveness and efficiency of treatment. They indicate the body's responses to the surgical procedures and drug efficacy [8]. Biomarker identification is then a crucial topic in disease prevention and control [4], which has attracted attention from laboratories and clinical medicine as well as the general public [9].

Traditionally, biomarker identification is often based on body secretions and fluids, such as blood serum, urine, stool and saliva [10]. It is often found that genes and their products are dysregulated during the development of disease [11]. The state transitions from normal to diseased or aggravated circumstances will be reflected from the inflammatory, circulatory, digestive and other systems in various aspects of causality, passenger and outcome [4]. The internal cells from the body systems will trigger present (high) or absent (low) expression of certain genes [12], RNAs [13,14], proteins [10], metabolites [15] and other small molecules in response to disease occurrence and progression. Moreover, the internal cells and differentially generated molecules will be propagated into the bloodstream. The external and internal cells and molecules imply the disease phenotypes of various development stages and treatment conditions [1,2]. In traditional biomarker discovery, the molecules, for example, proteins, metabolites and cells, are often used as biomarker candidates. In the screening procedure, the candidates are often first selected according to some measurements

Future Medicine part of fsg

compared between control and disease samples [10]. Recently, gene mutation [16], miRNA [17], cirRNA [18] and DNA methylation [19] are also recognized as molecular signatures of complex diseases.

In general, accurate identification of biomarkers for complex diseases is very difficult [3]. The reasons are numerous, but they mainly come down to three. First is our limited understanding about disease pathogenesis. The pathogenesis of many diseases is still not very clear. Recently, these processes are often regarded with correlation to multiple factors and their interactions, such as gene mutations [20] and epigenetic modifications [21]. Moreover, multiple organs and tissues will be involved in disease occurrence and development, for example, pancreas, liver, adipose and some other tissues in Type II diabetes [22,23]. The biomarker discovery process of recognizing molecules and cells from their differences between control and disease states then becomes difficult. Second, biomarkers should be specific, stable and consistent for clinical applications. Subtypes of complex diseases and personalized genetic features should be considered in biomarker identification [24]. The diversity of human gene expression in tissues and populations makes identification much more difficult [25]. Third, difficulties underlie the techniques, machines and apparatus used for measuring disease samples. The noise and complexity in the body leads to bias in measured signals [26]. The data preprocessing and mining techniques also affect discovery accuracy [27]. These issues aggregate the difficulties of biomarker identification and limit their clinical applications. There is an urgent need to develop novel techniques for detection of more accurate biomarkers.

In the last two decades, more and more high-throughput techniques have been developed to measure genome-wide gene mutations [28], epigenomic profiling [29], transcriptome-wide gene expression [30], proteome-wide proteins [31] with their interactions [32] and the metabolome [33] simultaneously. These techniques are revolutionizing biomedicine research paradigms including biomarker discovery [34,35]. They provide data resources and alternative ways to identify better biomarkers, which are more stable, specific and consistent in the detection of disease signals.

Genes, RNAs, proteins, metabolites and their internal and external interactions orchestrate the intricate systems of facilitating the functionalities of cells [34]. For the numerous involved molecules, complex diseases have been regarded as a disorder of these systems [36]. Networks provide a distinct and rational framework of describing these interactions and organizing the available data simultaneously. Always, the molecules cooperate together in a form of network to perform their functions [36]. The nodes represent these molecules and the edges represent their physical and functional relationships [37]. The network provides a topological representation of a complex system and the data characterizes its specific condition via quantitatively measured values of a large number of molecules. It is inadequate for the traditional methods to identify individual molecules, in other words, isolated nodes of genes, RNAs or proteins in the network, as the biomarkers for indicating normal, predisease, disease and postdisease states [36].

To leverage the generated big data for biomarker discovery, computational methods of identifying the network components beyond individual nodes become important options to meet the challenging request. Here, we focus on these bioinformatics methods of identifying network-based biomarkers. The molecular interactions in various genetic information transmissions provide evidences for the complexity of molecular cooperation in cells during disease occurrence and development [35,36]. The edges linking these molecules represent their functional cooperation. And they should be identified rationally for detecting the critical signals of state changes or indicating the maximal possibility of disease recurrence in the near future. The subnetworks containing both nodes and edges, such as modules and pathways in the network, are highly expected to serve as better biomarkers of complex diseases [36,38].

In this paper, we provide a review of computational methods of identifying biomarkers of complex diseases from omics data. We formulate the identifications into network components, in other words, nodes, edges, subnetworks and global networks, respectively. We first summarize the available high-throughput databases for building networks in various levels from the genotype of gene interactions to the phenotype of disease associations. Then, we categorize the available methods into the identification of different network components for classifying phenotypic states. The bioinformatics techniques of discriminating these network components in two states, for example, control and disease, and multiple phenotypic states, for example, normal, early, medium and severe stages of disease progression, are summarized accordingly. Due to the complexity of data types, we simply focus our review on few data types such as genome-wide gene expression microarrays and proteome-wide protein interactions. The methods proposed by our own group will be introduced in more detail along with overviews of similar methods. Last but not least, we share the vision of further improving biomarker identification from data mining. We conclude that combination of computational and

experimental methods provides a broader method for biomarker discovery.

High-throughput data resources

In the posthuman genomics project era, more and more high-throughput data resources are available for characterizing complex diseases from multiple levels [39]. Table 1 lists some databases with their brief introductions. These disease-related omics or network datasets provide the possibility of discovering biomarkers applicable in clinical trials. At the genetic level, databases such as Online Mendelian Inheritance in Man [40] and GWAS Catalog [41] provide mutations in genome sequences by GWAS, next-generation sequencing or microarray techniques. The RegNetwork [42] collects the TF-miRNA cooperative transcriptional and post-transcriptional regulations, which facilitate the identification of dysfunctional regulations in complex diseases [43]. Gene Expression Omnibus (GEO) [44] and ArrayExpress [45] deposit the comprehensive gene expression data, which highly benefit the disorder detection from transcriptional level. Human Protein Atlas Database [46] and PRoteomics IDEntification (PRIDE) [47] record the high-throughput expression information of proteins. Search Tool for Recurring Instances of Neighbouring Genes (STRING) [48] and BioGrid [49] database provide the resources of protein–protein interactions (PPIs). From metabolic level, Kyoto Encyclopedia of Genes and Genomes (KEGG) [50] and Reactome [51] record the curated pathways from literature. From phenotypic level, DiseaseConnect [52] provides the associations of diseases. There are also some databases available for diagnosis biomarkers, such as BiomarkerDigger [53] and Urinary Protein Biomarker (UPB) [54], which identify potential biomarkers identified from proteomic data. For drug research, DrugBank [55] records the drug with its target information. These data sources improve the speed and quality of biomarker identification. Moreover, some big science initiatives have accelerated the availability of high-throughput data, such as the ENCyclopedia Of DNA Elements (ENCODE) [56] and modENCODE [57], ImmGene [58], Roadmap Epigenomics [29]. The Cancer Genome Atlas (TCGA) [39] and International Cancer Genome Consortium (ICGC) [59] research networks aim to characterize the molecular profiles of many cancers. These open resources provide the unprecedented opportunity and challenging tasks of identifying disease biomarkers.

Framework of identification

The identification of disease biomarker can be easily formulated into a classification problem. In machine learning, a classification problem is to identify the features which can distinguish different classes and accordingly categorize samples into them [73]. Equivalently, the discovery of disease biomarkers is to identify molecules, cells and other indicators which can classify different phenotypic states. The features are those biomarkers which we aim to identify for labeling the disease states of occurrence, development, deterioration, metastasis as well as treatment effect and survival time. Instinctively, the biomarkers are the features for delineating control and disease samples, or multiple development stages. In theory, the phenotype is denoted as Y, for example, control and disease, or different disease stages and the biomarker(s) as X. The classifier is trained to learn the function f between Y and X from data, which is to determine $Y = f(X)$. f is a generalized function which usually cannot be reconstructed in its explicit form. When the function f is learned by a classifier, the outcome Y can be determined easily given the biomarker value X is of a new client or patient.

Figure 1 illustrates the framework of identifying network-based biomarkers of complex diseases from high-throughput data (Figure 1A & F). The available methods of identifying biomarkers are essentially to discover the relationship between network components and disease phenotypes. The network components, such as nodes, a set of nodes, edges and subnetworks as modules and pathways and others (Figure 1C), can be used as the candidates to be screened as biomarkers by classification algorithms (Figure 1D).

As shown in Figure 1B, the network is a graphical representation of the relationships among objects. According to the theory of network biology [34] and network medicine [36], different contents are contained in the network components. The node is the individual molecule, in other words, gene, RNA, protein, metabolite, etc., which is an isolated factor of the network. The edge between two nodes indicates their relationship. The set of nodes implies the group of molecules that performs certain functions. Compared with a set of nodes, modules and pathways in the form of community structures refer to the local interconnected parts of a subnetwork in the global network, where nodes can be reached directly from each other through its contained edges on which the information can be transmitted smoothly. Often, the module is based on network topology [74] and the pathway is based on prior biological knowledge [50].

In collecting disease samples (often with controls), the training step is to train the classifier which characterizes the distinctive features underlying different phenotypes. The widely used classification algorithms

Table 1. Some high-throughput data resources for multilevel biomarker discovery.

Database	Description	Level	Website	Ref.
ArrayExpress	It stores high-throughput functional genomics data	Transcriptome	www.ebi.ac.uk/arrayexpress/	[45]
BioGrid	A PPI data repository through comprehensive curation efforts	Proteome	http://thebiogrid.org/	[49]
BiomarkerDigger	A versatile disease proteome database and analysis platform for the identification of plasma cancer biomarkers	Biomarker	www.biomarkerdigger.org/	[53]
CGAP	CGAP is to determine the gene expression profiles of normal, precancer and cancer cells	Multiple levels	http://cgap.nci.nih.gov/	[60]
dbGaP	dbGaP is a public respository for phenotype, genotype, sequence data and their associations	Multiple levels	www.ncbi.nlm.nih.gov/gap/	[61]
DiseaseConnect	A comprehensive web server for mechanism-based disease–disease connections	Phenome	http://disease-connect.org/	[52]
DrugBank	A unique resource of detailed drug data with comprehensive drug target information	Drug	www.drugbank.ca/	[55]
ENCODE	ENCODE is a project with the aim to identify all functional elements in the human genome sequence	Multiple levels	www.encodeproject.org/	[56]
GenBank	An annotated database for all publicly available DNA sequences	Genome	www.ncbi.nlm.nih.gov/genbank/	[62]
GEO	GEO is a data repository of microarray, next-generation sequencing and other forms of high-throughput data	Transcriptome	www.ncbi.nlm.nih.gov/geo/	[44]
GO	GO is a functional annotation system for consistent descriptions of gene products	Functome	http://geneontology.org/	[63]
GWAS catalog	A curated catalog of published	Genome and epigenome	www.ebi.ac.uk/gwas/	[41]
GWASdb	A database that collects genomic variants from GWAS with their functional annotations and disease classifications	Multiple levels	http://jjwanglab.org/gwasdb	[16]
HMDB	HMDB is a database which collects detailed information about small molecule metabolites in human	Metabolome	www.hmdb.ca/	[64]
HPRD	HPRD is a curated human PPI database	Proteome	www.hprd.org/	[65]
ICGC	The ICGC aims to generate comprehensive data of genomic abnormalities in tumors from 50 cancer types	Multiple levels	https://icgc.org/	[59]
IDBD	An infectious disease biomarker database	Biomarker	http://biomarker.cdc.go.kr/biomarker/About/AboutIDBD_en.jsp?m=IDBD	[66]
ImmGene	A collaborative group generating complete microarray dissections of gene expression and regulations	Multiple levels	www.immgen.org/	[58]

The databases are ordered alphabetically.
CGAP: Cancer genome anatomy project; dbGaP: Database of genotypes and phenotypes; ENCODE: Encyclopedia of DNA elements; GEO: Gene-expression omnibus; GO: Gene ontology; GWAS: genome-wide association study; HMDB: Human metabolome database; HPRD: Human protein reference database; ICC: International cancer genome consortium; KEGG: Kyoto encyclopedia of genes and genome; MOPED: Model organism protein expression database; OMIM: Online mendelian inheritance in man; PPI: Protein–protein interaction; PRIDE: Proteomics identification; STRING: Search tool for recurring instances of neighbouring gene; TCGA: The cancer genome atlas; TRED: Transcriptional regulatory element database; UPB: Urinary protein biomarker.

Table 1. Some high-throughput data resources for multilevel biomarker discovery (cont.).

Database	Description	Level	Website	Ref.
IntAct	IntAct provides a database system and analysis tools for molecular interactions	Proteome	www.ebi.ac.uk/intact/	[67]
KEGG	A database for understanding high-level functions and utilities of biological systems	Metabolome	www.genome.jp/kegg/	[50]
MOPED	MOPED is mass spectrometry proteomics studies on humans and model organisms	Proteome	www.proteinspire.org	[68]
OMIM	OMIM is a comprehensive database which categorizes the known disease genes and genetic phenotypes	Genome and epigenome	http://omim.org/	[40]
Pathway Commons	An integrated database for collecting available pathways in multiple organisms	Multiple levels	www.pathwaycommons.org/about/	[69]
PaxDB	A comprehensive absolute protein abundance database	Proteome	http://pax-db.org/	[70]
PRIDE	The PRIDE database is public data repository for proteomics data	Proteome	www.ebi.ac.uk/pride/archive/	[47]
Reactome	A curated and peer-reviewed pathway database	Metabolome	www.reactome.org/	[51]
RegNetwork	An integrated database of transcriptional and post-transcriptional regulatory networks	Transcriptome	www.regnetworkweb.org	[42]
Roadmap Epigenomics	Roadmap epigenomics mapping consortium was launched with the goal of producing a public resource of human epigenomic data to catalyze basic biology and disease-oriented research	Epigenome	www.roadmapepigenomics.org/	[29]
STRING	A database of known and predicted PPIs	Proteome	http://string-db.org/	[48]
TCGA	TCGA is a comprehensive and coordinated effort to generate multi-dimensional maps of the key genomic changes in major types and subtypes of cancer	Multiple levels	http://cancergenome.nih.gov/	[39]
The Human Protein Atlas	A proteomics database of protein expression in normal and cancer tissues	Proteome	www.proteinatlas.org/	[46]
TRED	TRED is a curated database of regulations between TF and target gene	Transcriptome	http://rulai.cshl.edu/TRED/	[71]
UniProt	A comprehensive resource of protein sequence and functional information	Proteome	www.uniprot.org/	[72]
UPB	A curated human and animal urine protein biomarker database	Biomarker	http://122.70.220.102/biomarker/index.asp	[54]

The databases are ordered alphabetically.
CGAP: Cancer genome anatomy project; dbGaP: Database of genotypes and phenotypes; ENCODE: Encyclopedia of DNA elements; GEO: Gene-expression omnibus; GO: Gene ontology; GWAS: genome-wide association study; HMDB: Human metabolome database; HPRD: Human protein reference database; ICC: International cancer genome consortium; KEGG: Kyoto encyclopedia of genes and genome; MOPED: Model organism protein expression database; OMIM: Online mendelian inheritance in man; PPI: Protein–protein interaction; PRIDE: Proteomics identification; STRING: Search tool for recurring instances of neighbouring gene; TCGA: The cancer genome atlas; TRED: Transcriptional regulatory element database; UPB: Urinary protein biomarker.

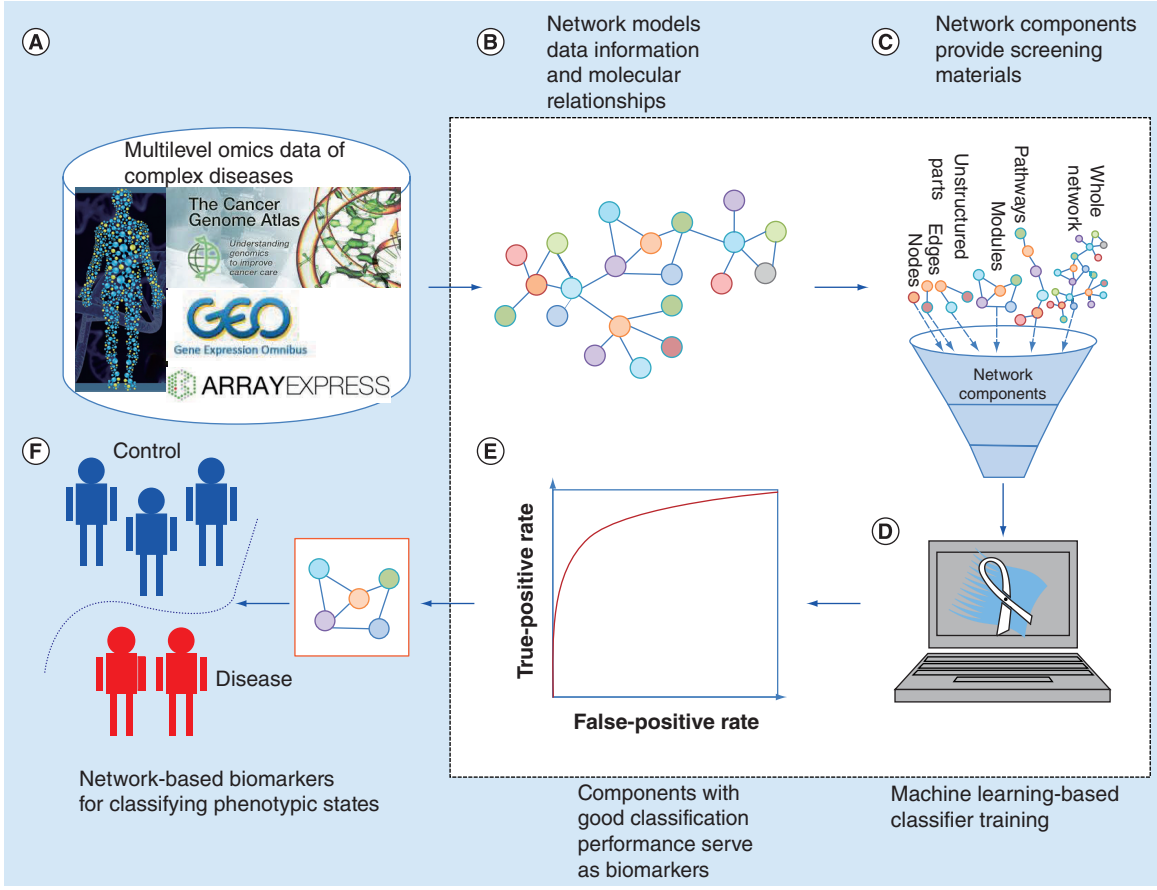

Figure 1. The framework of identifying network-based biomarkers from high-throughput data. **(A)** The available profiling data of complex diseases in multiple levels, which provide the resources for biomarker discovery. **(B)** The information and relationship among molecules, such as gene, RNA, protein and metabolites, are organized by a network model. **(C)** The network components such as nodes, edges, unstructured parts of network, modules and pathways provide the materials for identifying disease biomarkers. **(D)** Machine-learning-based classifier is built to select the specific features and abilities in these network components via the classification of different phenotypic states. **(E)** The validation and evaluation of these identified distinct network components served as biomarkers for distinguishing different conditions. **(F)** Network-based biomarkers are applied to distinguish phenotypic states, for example, control and disease.

are neural network [75], naive Bayesian [76], support vector machine [77,78] and random forest [79]. After training the relationships between network components and disease phenotypes, the biomarkers are those components which classify disease phenotypes with high sensitivity and specificity (Figure 1E). To evaluate the biomarker classification performance, *k*-fold (e.g., *k*=10) cross-validations or leave-one-out validations are usually implemented upon the training datasets. The application of these identified biomarkers is to extend the obtained biomarkers to unknown property samples. The biomarkers can be then used to diagnose diseases, predict drug effects and treatments and measure recurrence possibility and survival time.

As to nodes and node sets, edges, unstructured subnetworks, modules and pathways, the whole processes of identifying disease biomarkers will be slightly different according to their specific characteristics. We will review some computational methods with focus on their characteristics in the following sections.

Identifying node-based biomarkers

Table 2 lists some available bioinformatics methods for biomarker discovery from high-throughput data. The network model provides us a foothold to identify biomarkers. The first type of biomarker is the node in the network. Generally, node-based biomarker identification is to evaluate whether the node contains the distinct properties in different phenotypes, such as control and disease. Detecting the differential information of a molecule (i.e., a node of a network) across different states is the first-step effort in identification. Gene expression profiles are available for techniques such as microarray [30] and RNA-Seq [80]. The strategy of identifying node

biomarkers is to first screen all nodes in the network by statistically testing their differences over samples. Differentially expressed genes are then evaluated by their distinguishable power of classifying samples and then selected as biomarkers with further *in vivo* experimental validations. Obviously, the node-based methods used in gene expression profiles can also be used in proteomics, metabolomics and other datasets for identifying distinctive information across conditions.

Table 2. Bioinformatics methods for identifying biomarkers of complex diseases from high-throughput data.

Study (year)	Disease	Category	Data	Ref.
Golub *et al.* (1999)	Leukemia	Node	Gene expression	[81]
Guyon *et al.* (2002)	Multiple cancers	Node	Gene expression	[78]
Lu *et al.* (2005)	Multiple cancers	Node	miRNA expression	[13]
van de Vijver *et al.* (2002)	Breast cancer	Node	Gene expression	[82]
van 't Veer *et al.* (2002)	Breast cancer	Node	Gene expression	[12]
Efron *et al.* (2007)	Multiple complex diseases	Node set	Gene expression, gene groups of metabolic pathways	[83]
Subramanian *et al.* (2005)	Multiple complex diseases	Node set	Gene expression, gene groups of metabolic pathways	[84]
Tian *et al.* (2005)	Multiple complex diseases	Node set	Gene expression, gene groups of metabolic pathways	[85]
Bandyopadhyay *et al.* (2010)	DNA damage related to cancers	Edge	Multiple omics data	[86]
Liu *et al.* (2012)	Gastric cancer	Edge	Gene expression, PPI	[87]
Sahni *et al.* (2015)	Multiple diseases	Edge	Multiple omics data	[88]
Zhang *et al.* (2014)	Cholangiocarcinoma and diabetes	Edge	Gene expression	[89]
Chuang *et al.* (2007)	Breast cancer	Unstructured subnetwork	Gene expression, PPI	[90]
He *et al.* (2011)	CHD	Module	Gene expression, PPI	[91]
He *et al.* (2012)	HCC	Module	Gene expression	[92]
Hofree *et al.* (2013)	Multiple cancers	Module	Genomic mutation, gene expression, regulatory network, PPI, metabolic pathways, among others	[93]
Liu *et al.* (2011)	Alzheimer's disease	Module	Gene expression, PPI	[94]
Segal *et al.* (2004)	Multiple tumors	Module	Gene expression	[95]
Taylor *et al.* (2009)	Breast cancer	Module	Gene expression, PPI	[96]
Wen *et al.* (2012)	Colorectal cancer	Module	Gene expression, DNA methylation microarrays, PPI	[97]
Zhang & Horvath (2005)	Multiple diseases	Module	Gene expression	[98]
Lee *et al.* (2008)	Multiple cancers	Pathway	Gene expression, genes of metabolic pathways	[99]
Liu *et al.* (2012)	Multiple cancers	Pathway	Gene expression, PPI	[100]
Liu *et al.* (2013)	Time courses as cell cycles of complex diseases	Pathway	Gene expression, regulatory network, PPI	[101]
Chen *et al.* (2012)	Multiple complex diseases	Dynamical network biomarker	Gene expression, PPI	[102]

Note that these methods can be flexibly used to discover biomarkers of different diseases from different levels of data given suitable inputs and modifications. The methods are ordered by 'Category'.
CHD: Coronary heart disease; HCC: Hepatocellular carcinoma; PPI: Protein–protein interaction; Ref.: Reference.

For two phenotypic states, Student's *t*-test [103] is a widely used method to detect differential information. The method assumes the null hypothesis of gene expression, that there is no change in the control and disease samples, in other words,

$$H_0 : \mu_{control} = \mu_{disease}$$

Statistical testing is used to check whether the mean of gene expression contains significant change. The differentially expressed genes across control and disease provide the candidates for distinguishing the two states. Wilcoxon rank test [104] is a rank-based significance test when comparing two type of samples. The nonparametric statistical hypothesis test contains no population assumptions of normal distribution and can be used as an alternative to the former Student's *t*-test. Due to the population assumption of the test models and the limited availability of samples, significance analysis of microarrays (SAM) [105] is another widely used method to identify differential genes. Each gene is assigned to a score on the difference relative to the standard deviation of expression measurements, in other words,

$$d(\mathrm{i}) = \frac{\overline{x}_{\mathrm{c}}(\mathrm{i}) - \overline{x}_{\mathrm{d}}(\mathrm{i})}{s(\mathrm{i}) + s_0}$$

where

$$\overline{x}_{\mathrm{c}}(\mathrm{i})$$

and

$$\overline{x}_{\mathrm{d}}(\mathrm{i})$$

are defined as the mean expression values for gene *i* in control and disease, respectively. *s(i)* is the standard deviation of expression experiments and s_0 is a predefined constant. The significance is then evaluated by a permutation strategy [105].

More advanced techniques are used to discover the biomarkers in complicated conditions. When comparing more than two states, analysis of variance (ANOVA) techniques are often implemented [106]. ANOVA model is a general and powerful tool to identify differential information through multiple conditions. ANOVA F-test formulates an estimate of variation across conditions to an estimate of error variance. The test then identifies the significance of rejecting the null hypothesis if there is no variance change across the states. The gene expression data are often time series [107]. For this case, many statistical testing methods have been developed for time-course data. Widely used techniques include the edgeR (extraction of differential gene expression in R) package in Bioconductor [108]. It originally identifies statistically significant changes in expression over time by representing gene expression trajectories as cubic splines. EdgeR can be applied to differential expression in various levels such as gene, exon, transcript or tag. It also includes DESeq [109] for differential analysis of RNA-Seq [80] and ChIP-Seq [110] data. Short time-series expression miner (STEM) can identify significant genes from short time series microarrays based on a clustering method [107]. Functional principal component analysis (FPCA) method employs a functional principle component analysis method to remove the noise and ambiguity of the expression data for differential gene identification during time courses [111].

It is easy to know that the ranks of differentially significant scores of these genes are not changed after operating a statistical test. To select the differential genes by their p-values, there is often a need to define a threshold for statistical significance. False discovery rate [112], Akaike information criterion [113] and Bayesian information criterion [114] techniques of variable selection are often employed to control the false positive ratios in the discovery of differential genes. To investigate the differential information underlying a set of isolated nodes as a whole, for example, a gene set, some statistical testing methods have been proposed for identifying the enriched gene group as integrative biomarkers [83,84]. The differential genes or gene sets contain the distinctive characteristics to distinguish phenotypic conditions. They are the biomarker candidates, but they are not the determined biomarkers due to their complicated relationships with these phenotypes [94]. Computational classification powered evaluation and further experimental validations are the following steps for identifying biomarkers from them with more possibilities and confidences. Node-based methods are the fundamental strategies of identifying disease biomarkers [36]. Generally speaking, the other available network-based methods are built on these methods and philosophies.

There are three major points for the nonequivalence between differential genes and biomarkers. First is the generalization ability. Although we identify a gene or a gene set with differential information over different conditions, it might be a false positive and should be validated in large-scale samples, especially in independent samples. Considering the maturing process of these omics techniques and the disturbing noise in the data, some important issues also affect the generalization ability of these candidates. Second, the biomarkers should be determined by their pathological, dysfunctional and clinical implications. Often, hundreds or thousands

genes have been identified as differential genes [83,84]. There are too many candidates and possibilities, of which our focus is on a smaller range of genes. These genes should be checked for their dysfunctional implications in the physiological processes. The categories of causal, driver, passenger, response and housekeeping genes for the disease have not been determined yet. Some correlation metrics, such as Pearson's correlation coefficient (PCC) and mutual information, are beneficial to determine their relationship with phenotypic outcomes. However, they cannot be determined as causal disease factors or affected molecular entities only by optimizing the classification power. The specificity of gene signatures served as biomarkers is very important in clinical applications [82]. Third is the complexity of disease. The molecules in the cell, such as genes, gene products, metabolites and minerals, are linked together in the form of intricate networks for performing functions [34]. It is difficult to mark the disease by the independently isolated nodes in the network. Currently, it is known that the complex diseases are system disorders, which imply they are not caused only by a single gene and/or protein and/or metabolite [35,36]. It has also been found that the permutation widely exists in molecular interactions in disorders [88,115]. Moreover, from a systematic perspective, complex diseases are caused by various interactions and subnetwork communities, such as modules and pathways. The network medicine methods provide powerful alternatives to reveal biomarker mutations [116], genes [117], miRNAs [17], lncRNAs [118] and proteins [10] and metabolites [15] from high-throughput data.

Identifying edge-based biomarkers

Unlike node-based methods, edge-based biomarker discovery methods take molecule interactions in the network into consideration. Instead of observing the molecules in the network in isolation, edge-based methods regard them from a cooperation perspective. Compared with the disorders of isolated nodes, the disorders of molecular interactions [88,115,119,120] and then the subnetworks they are involved in [36] seem to be a more reasonable hypothesis for the pathogenesis of complex diseases. Edgotype refers to the edgetic perturbation in biomolecular interaction networks, which leads genetic variants to distinct phenotypic outcomes. The edgotype bridges the gap between genotype and phenotype through the rewiring interactions among molecules [121]. Edge-based methods are expected to uncover better biomarkers relating genotypes to phenotypes.

As mentioned, many edge-based methods are based upon node-based methods. In the same philosophy of detecting discriminant information through control and disease states, we combined node significance and edge significance in two phenotypic states together by Fisher's method (shown in Figure 2A) to identify the activated linkages during the disease progression of Alzheimer's disease (AD) [122], as well as those in different AD brain regions [94,123]. The combined significance as weights are overlaid on a highly-qualified PPI network. Then, we identified the context-specific protein interactions with rewiring characteristics from these edge-weighted networks. The edgotype provides a powerful alternative to bridge genotype and phenotype [121,124], then serves as a biomarker of indicating the phenotypic signals of AD [87].

We also provided an edge-based method of identifying the biomarkers of gastric cancer from a weighted PPI network. By distinguishing the two states of control and disease, we identified a protein subnetwork of these differential edges which accurately classified the disease and control samples. As shown in Figure 2B, in contrast to the former node-based methods of differential genes, the PCC between a pair of interacting proteins in the human PPI network was calculated based on the samples in each group. Subsequently, only those PPIs with high correlation coefficients were reserved with the assumption that these PPIs with low correlation coefficients do not occur in the corresponding samples. The differential PPIs between control and disease samples were then identified, which illustrate the dynamic changes of rewiring interactions across control and disease status. The differential PPIs are obtained by combining a specific PPI network in control and a specific PPI network in disease while removing their common edges. They indicate the differential information of cooperation between these proteins in gastric cancer. The differential interactions form subnetworks and can significantly distinguish control from disease samples. The computational method provides a novel approach to detecting diagnostic biomarkers of stomach cancer [87].

The former methods provide direct evidence that edge-based methods are effective for identifying biomarkers. These identified biomarkers are very meaningful in the mechanisms of deciphering complex diseases. Moreover, there are topological neighbors and functional groups in the network [36,87]. From this perspective, the edge-based biomarkers outperform node-based identifications in the classification performance and pathogenesis meaning [87]. Recently, another edge-based method has been proposed for its importance and prevalence [89].

Identifying subnetwork-based biomarkers

For subnetwork-based biomarker discovery, there are several categories in these subnetworks, in other

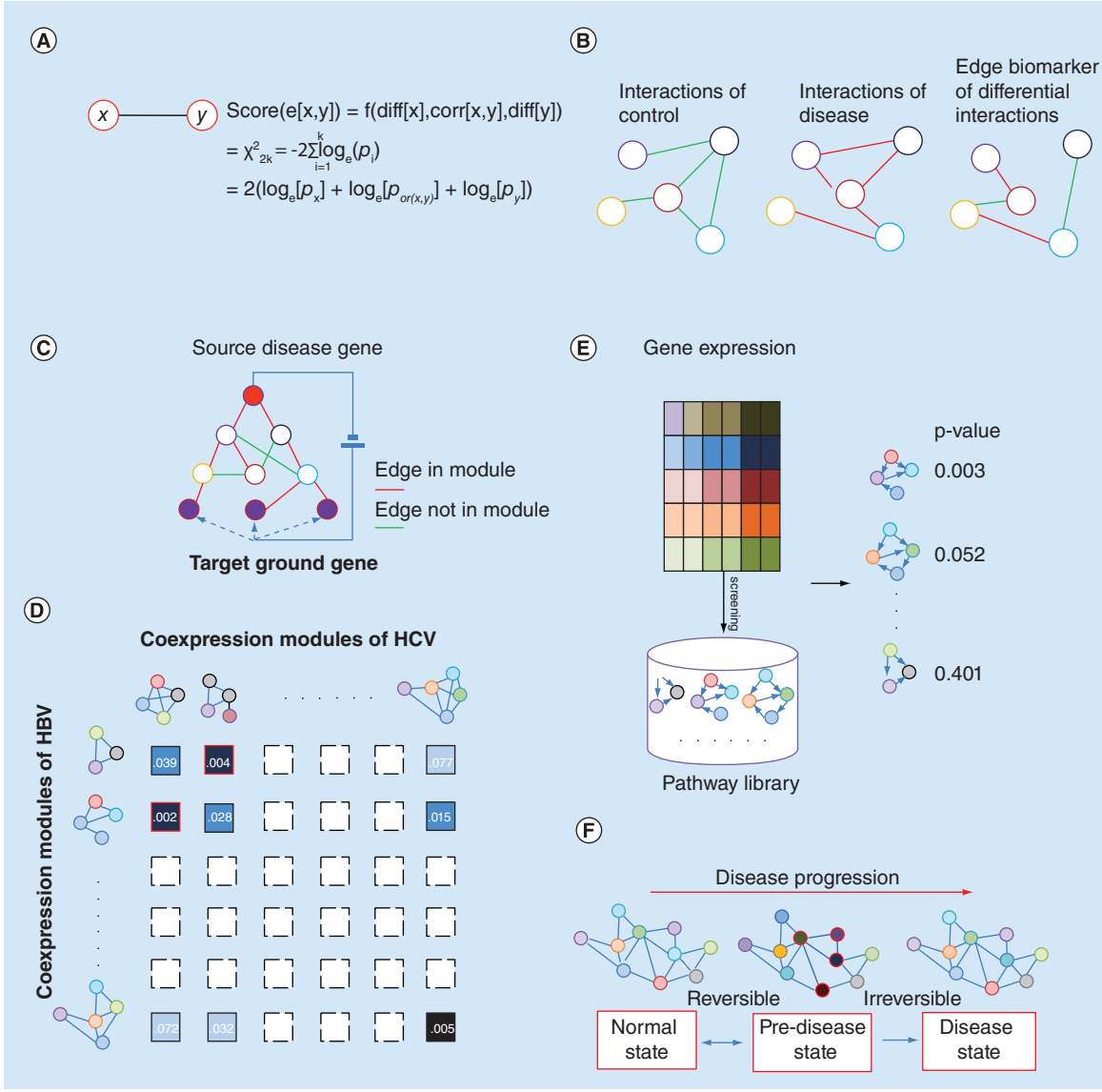

Figure 2. The diagrams of identifying some network-based biomarkers. (A) The edge score between nodes x and y is weighted by Fisher's method. The differential information of two nodes and that of the linkage are combined together [122]. **(B)** The edge-based biomarker identified by rewiring differential interactions in control and disease [87]. **(C)** The information flows between source disease causal genes and target differential genes are modeled by an electronic potential model. The involved nodes and edges construct a module and use it as biomarker after evaluation [91]. **(D)** The correlation matrix between gene coexpression modules of HBV and that of HCV [125]. **(E)** The documented pathways are evaluated for the consistency with phenotypes individually. The enriched pathways are identified as pathway biomarkers for complex diseases [101]. **(F)** The dynamical network biomarker during the disease progression [102]. Disease can be reversible at the critical points of pre-disesae state, while cannot after it. Dynamical network biomarker extracts the early warning signals of disease state transitions. HBV: Hepatitis B virus; HCV: Hepatitis C virus.

words, the unstructured parts of a network, the structured modules and pathways. Unstructured subnetworks refer to the irregular parts without particular properties in the global network. Modules refer to the community structures of a network which contain internally dense connections and sparser connections between groups [74]. These topologically particular network structures have been identified as functional blocks in many networks of biological systems [34,74]. Pathways are often knowledge-based interactions between molecules containing particular functional implications [50].

Systematic identification of disease-related subnetworks and further subnetwork biomarkers from global networks can provide deep insights into the mechanisms of complex diseases and gracious assistance in diagnosis

and prognosis [35–36,93]. Different from the former two methods, subnetwork-based methods evaluate a group of nodes and edges simultaneously in terms of connected subnetworks. From an integrative viewpoint, the signals contained in the locally networked systems and the dysfunctional relationships with phenotypes overwhelm those of individual nodes and edges [36]. The subnetwork-based method of identifying biomarkers often defines a relationship metric of describing the association between phenotypes and subnetworks. To this end, PCC [126], mutual information [127] and Kullback-Leibler divergence [128] are often employed. Then, an algorithm is developed to identify the subnetwork by optimizing its association score with phenotypes. In that way, these subnetworks are evaluated for classifying disease samples and some of them are identified as biomarkers.

There are some specific strategies for different types of subnetworks in biomarker discovery. First, the unstructured subnetwork can be extracted directly from the network. These methods often begin with an interested node (or edge) and gradually increase the solution sets from neighboring nodes and edges by maximizing the predefined association metrics through the phenotypic differences. The searching algorithm will be terminated by a given threshold of distance from the initial network component and the final output is the identified subnetwork-based biomarker. Second, it needs to divide the global network into some subnetworks in the form of modules. These modules are the candidates for selecting biomarkers which can classify the samples according to phenotypes. Third is to refer to certain prior knowledge of pathways in various resources, such as the documented pathways in KEGG [50]. The methods directly screen upon the set of pathways and identify the discriminating pathways as biomarkers.

For the unstructured subnetwork, a seminal work of identifying biomarkers for breast cancer metastasis has been proposed [90]. The method identified protein subnetwork biomarkers for classifying the metastatic and nonmetastatic tumors by calculating the mutual information between subnetworks and phenotypes. The optimal subnetworks indicating different phenotypes are searched by increasing the size of targeting signatures with the justification of their associations with phenotypes. The method resorts to a heuristic optimization strategy to identify biomarkers in the form of subnetworks. The biomarkers are informative of nondiscriminative disease genes [90]. This also proves that the differential genes or causal disease genes might not definitely serve as biomarkers with significant classification accuracy for tumor samples [129]. The networking formation indicates the cooperation among these proteins during cancer metastasis. The authors further applied their pipeline to identify subnetwork biomarkers of chronic lymphocytic leukemia with the validation of immunoblotting techniques [130]. Identifying subnetwork-growing-based biomarkers overcomes the possible classification biases caused by only differentially expressed genes or documented disease casual genes.

In the module-based paradigm of biomarker discovery, we provided a method of identifying dysfunctional module biomarkers in coronary heart disease (CHD) [91] by potential energy as shown in Figure 2C. The information transmitting from the source genes (annotated disease genes) to their targets (differentially expressed genes) is used to decompose the PPI network into modules. These modules are individually evaluated as biomarkers of classifying disease samples via mutual information with phenotypes. The higher accuracy of classification proves the efficiency of our proposed method of identifying module biomarkers of CHD by information transmission on protein networks [36,91].

In the module-based subnetworks of discriminating disease phenotypes, the interrelationship among modules is also very important for discovering the accurate biomarkers by screening all possible module combinations. We proposed a method for identifying module biomarkers for hepatocellular carcinoma (HCC) via gene coexpression networks [125]. The framework is summarized in Figure 2D. The major risk factors of HCC are chronic infection with hepatitis B virus (HBV) and hepatitis C virus (HCV) [131]. The similarities and differences between HBV and HCV are evaluated by comparing the overlap of gene compositions and functional annotations in HBV and HCV modules. We identified distinct patterns of gene coexpression networks and inflammation-related modules from genome-scale microarray data upon viral infection, and further classified them into oncogenic and dysfunctional modules, respectively. These modules perform significant classification powers to distinguish the stages of disease progression. The module biomarkers have also been tested in independent datasets for their classification abilities. We also compared these viral infection modules across HBV- and HCV-induced HCC by module preservation during disease progression. The revealed modules of biomarker properties shed light on the classification of different types of virus-induced HCC, which will highly benefit the diagnosis of liver cancer [125].

So far, many knowledge-based pathways have been documented for describing molecular relationships in functional processes [50]. An alternative to identifying subnetwork-based biomarkers of complex diseases is to screen these pathways [99]. After determining pathway activities in response to specific phenotypic states, the methods evaluate their classification power of distin-

guishing these states and make the apparent pathways serve as biomarkers.

We provided a Gaussian graphical model to detect consistency between subnetwork pathways and high-throughput data by evaluating the significance of their correspondence [101]. Based on knowledge-based pathways, our model is to identify the pathways as the subnetwork-based biomarkers. Figure 2E illustrates the framework of assessing the significance of consistency by screening pathways. The statistical significance of the consistency is evaluated by randomizing the pathway structure and calculating the possibility. An empirical p-value of each pathway is available from the consistency likelihood between pathway architecture and high-throughput data. We have identified the significantly responsive gene regulatory pathways in diabetic development [101,132] and some master regulators of transcription factor [133]. The pathway biomarkers provided deep insights for the diagnosis of diabetic progression [36,132,133].

Identifying dynamical network biomarkers

In personalized medicine, a patient's genetic contents and other molecular or cellular contexts should direct the selection of appropriate treatments and optimal therapies [24,134]. From this regard, different people might contain different patterns of genetics, molecules and cells during disease production and development. And these different patterns would be the ideal biomarkers in individuals for distinguishing the characteristics of disease states [102]. This type of biomarker emphasizes on the specificity rather than the generality underlying these signatures. Moreover, the classification of multiple conditions is strengthened to correspond with the state transition points and signals.

To this end, we proposed a new concept of network-based biomarker, in other words, dynamical network biomarker (DNB) [102]. We provided computational strategies to identify DNBs for grasping the early warning signals of complex diseases during various disease courses [102,135]. From the dynamical viewpoint, we can briefly categorize the development of many diseases into three states, in other words, normal, predisease and disease [102], as shown in Figure 2F. During disease progression, it usually becomes irreversible to the normal state if the system passes the critical point and enters another stable state. If we can identify the early warning signals of the disease and diagnose the predisease state, it is possible to take appropriate intervention actions to prevent qualitative deterioration to the following disease states.

We defined the DNB as a special group of observable molecules to indicate the sudden deterioration of a complex disease, which often involves over thousands of gene, RNAs, proteins and metabolites. In the molecular network of a complex disease, if there is a group of variables of molecules, which satisfy the following three criteria:

- The average PCCs of these molecules drastically increase in absolute value;
- The average PCCs of molecules between this group and any others, in other words, between molecules inside this group and any other molecules outside this group drastically decrease in absolute value;
- The average standard deviations of molecules in this group drastically increase.

This group is thus called a dominant group of the system, whose change will reflect a transition of the system to the disease state. Each of the three conditions represents a criterion, and their combination is naturally expected to be a strong signal or an indicator for the predisease state, a critical state just before the critical transition point. The early warning of critical transition is reflected in the group of molecules. Because the dominant group characterizes dynamical features of the underlying system and the molecules in the group are also strongly and dynamically correlated in the predisease state, the molecules in the group are expected to form a network. We regard it as a DNB of the disease [102]. Compared with the former methods of network components, DNB is often based upon the whole network property.

For the biomarkers identified by the former reviewed methods, their expression and concentration reflect the presence or severity of the disease states. They are required to have consistent values in different disease states for different people. Differently, there are no requirements for such consistency in DNB. The abundance and concentration of molecules in DNB behave dynamically in a strongly collective manner. They tend to increasingly fluctuate when the system approaches the predisease state [102]. Thus, each individual may have a different DNB even for the same complex disease. Therefore, in contrast to traditional biomarkers, a DNB is not necessarily composed of a fixed bunch of molecules and might contain different members depending on individual features in the high-throughput datasets (some people might contain similar features). The early warning signal of a complex disease can be detected by DNB, which is impossible for traditional biomarkers or former methods. Compared with the identified network-based cancer biomarkers of certain molecules [136–138], DNB focuses on the anomaly detection of critical signals. The existence

of DNB implies that the system is in a predisease state. DNB extends the biomarker concept and application, and it is a powerful diagnostic and prognostic tool for precision medicine [139].

Discussion & conclusion

Numerous high-throughput techniques such as genomics, transcriptomics, proteomics, metabolomics and phenomics provide precious opportunities of characterizing the large-scale genetic, molecular and cellular biosystems in comprehensive levels [36]. Many omics projects and efforts have been initialized for complex diseases [39]. The generated big data create the possibility of identifying precise biomarkers of complex diseases for diagnosis, prognosis, therapeutic strategies and treatment assessment. The boom in bioinformatics methods and software meets the urgent need to discover biomarkers from high-throughput data. The new paradigm of biomarker discovery in computational medicine will definitely revolutionize complex disease research.

In this review, we summarized the state-of-the-art computational strategies of identifying biomarkers in the form of network terminologies. The network components such as node, edge, module and pathway are those candidates for screening. From the network-based biomarkers, some widely used computational and statistical methods are reviewed, respectively. We also gave brief introductions to our own works in some categories to further demonstrate the specific and detailed steps in the discoveries. Although the computational methods provided considerable insights of biomarker candidates and achieved great successful identification from data, these methods are still in their developing periods. Different methods often have their own advantages for specific types of available data, respectively. Benchmarks for assessing and comparing these network-based methods are urgently needed for building more powerful high-throughput screening methods of biomarker discovery [140].

The problems underlying current bioinformatics methods can be listed as follows. First is about biomarker validation. Essentially, the computational methods predict the biomarker candidates with great potential of distinguishing diseases from data analyses. The generalization ability and usefulness should be evaluated according to the follow-up *in vivo* experiments and clinical bench trials. Often, the candidates are evaluated by the classification methods obtained from the training datasets. To raise the possibility of a true biomarker, the computational methods should improve the validation steps in multiple independent datasets [97].

Computational methods usually generate a long list of candidates, which are very difficult to validate by traditional *in vitro* and *in vivo* experiments. It is very important to control the false positives on independent datasets and provide ranking scores of these candidates for further wet experiments and clinical validations. The robustness of candidates in the classification of disease samples should be evaluated by designed permutations and tests *in silico* [141]. To resort to more advanced computational strategy is still economic in the biomarker discovery compared with the wet-lab-based validations.

The work to identify biomarkers from data pioneered biomarker discovery and should be extended for further applications. So far, the omics technologies are in their maturation periods. More and more reliable measurements will be generated for describing the states of normal and disease. The computational methods of data processing and mining rely on the central high-throughput techniques. The strategies for handling raw datasets [142], cells [143] or count numbers [144] will affect data exploration. The biomarker representing true conditions might be eliminated or omitted in unsuitable steps. Moreover, these data often describe the biosystems from multiple levels. How to integrate these heterogeneous data for more accurate and robust biomarker identification is a very important research topic [11,145].

Second is biomarker interpretation. The biomarker is expected to be crucial to decipher the disease pathology and inspire optimal therapies. There are high possibilities for these biomarkers to perform dysfunctions as disease drivers and passengers. The causal reasoning of biomarkers from high-throughput data provides very valuable information to dysfunctional implications [146]. The biomedical meaning of these identified biomarkers indicates the disease mechanism, which will direct the treatment and drug design for leading the following therapy. The biomarker contains the generalization ability in various disease samples. While in the personalized medicine era, the DNB should also be emphasized for the specificity of individuals. The genome differences and variations of different people require the precision disease classification and stratification with the specific genetic contexts and backgrounds. The interpretation of the biomarkers assesses the risk of personalized measures identified from data.

In conclusion, we reviewed the computational strategies of identifying network-based biomarkers of complex diseases from high-throughput data. The individual nodes, node sets, edges, unstructured subnetworks, modules and pathways are evaluated as disease biomarkers. The power of computational

methods is due to the information in the big data and the flexibility in identification. Compared with traditional methods, the dry-lab-based methods provide large-scale screening of network components. We summarized the major pipeline of each strategy and commented on their embedded advantages and weaknesses, respectively. Some improvement possibilities are also proposed for further research references. Novel methods are accelerating biomarker discovery and provide valuable information for translational bioinformatics research. Although they are in their infancy, a bright future is assured. In the big data era, this research direction should also be emphasized with the aim of translating the available datasets into biomarkers for diagnosis, prognosis and understanding of complex diseases.

Future perspective

With the availability of multilevel high-throughput data, computational methods of identifying biomarkers from data will become more and more important. We envision that the bioinformatics methods will become routine pipelines in biomarker discovery. The network model is a very powerful framework of organizing the heterogeneous data and characterizing the multiplex molecular relationships in cells. In precision medicine, the combination of intelligent computation and biomedical experiment will definitely accelerate biomarker discovery and validation.

Acknowledgements

The author thanks the four anonymous referees as well as the editors for their helpful comments.

Financial & competing interests disclosure

This work was partially supported by the National Natural Science Foundation of China under grant no. 61572287 and 61533011, Shandong Provincial Natural Science Foundation of China (no. ZR2015FQ001), the Fundamental Research Funds of Shandong University under grant no. 2014TB006 and 2015QY001-04, and the Scientific Research Foundation for the Returned Overseas Chinese Scholars, Ministry of Education of China. The paper was also supported by a fund form the National Center for Mathematics and Interdisciplinary Sciences, Chinese Academy of Sciences. The author has no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Executive summary

- Rational data integration and accurate learning algorithms from data are greatly expected to be designed and developed for biomarker discovery in complex diseases in precision medicine.
- Bioinformatics methods provide powerful tools and valuable pipelines for identifying critical network-based biomarkers from high-throughput data.
- Network components from the individual nodes and edges to integrative modules and pathways can malfunction synergistically and comprehensively to drive disease initiation and development. Multiscale modeling and measuring the disease dynamics from network models generate accurate disease biomarkers.
- Network-based biomarkers propose an insightful alternative to bridge the causal relationship between genotype and phenotype. Not only do they provide the signatures of diagnosis and prognosis of marking complex diseases, but also uncover the pathogenesis of the occurrence, development and progression of complex diseases.

References

1 Aronson JK. Biomarkers and surrogate endpoints. *Br. J. Clin. Pharmacol.* 59(5), 491–494 (2005).

2 Brower V. Biomarkers: portents of malignancy. *Nature* 471(7339), S19–S21 (2011).

3 Ludwig JA, Weinstein JN. Biomarkers in cancer staging, prognosis and treatment selection. *Nat. Rev. Cancer* 5(11), 845–856 (2005).

4 Sawyers CL. The cancer biomarker problem. *Nature* 452(7187), 548–552 (2008).

5 Riaz S. Study of protein biomarkers of diabetes mellitus Type 2 and therapy with vitamin B1. *J. Diabetes Res.* 2015, 150176 (2015).

6 Resnick SM. Preclinical biomarkers in Alzheimer disease: a sum greater than the parts. *JAMA Neurol.* 71(11), 1357–1358 (2014).

7 Leary RJ, Kinde I, Diehl F *et al.* Development of personalized tumor biomarkers using massively parallel sequencing. *Sci. Transl. Med.* 2(20), 20ra14 (2010).

8 Frank R, Hargreaves R. Clinical biomarkers in drug discovery and development. *Nat. Rev. Drug. Discov.* 2(7), 566–580 (2003).

9 Lebo PB, Quehenberger F, Kamolz LP, Lumenta DB. The Angelina effect revisited: exploring a media-related impact on public awareness. *Cancer* 121(22), 3959–3964 (2015).

10 Rifai N, Gillette MA, Carr SA. Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nat. Biotechnol.* 24(8), 971–983 (2006).

11 Aerts S, Lambrechts D, Maity S *et al.* Gene prioritization through genomic data fusion. *Nat. Biotechnol.* 24(5), 537–544 (2006).

12 Van 'T Veer LJ, Dai H, Van De Vijver MJ *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415(6871), 530–536 (2002).

13 Lu J, Getz G, Miska EA *et al.* microRNA expression profiles classify human cancers. *Nature* 435(7043), 834–838 (2005).

14 Huang E, Cheng SH, Dressman H *et al.* Gene expression predictors of breast cancer outcomes. *Lancet* 361(9369), 1590–1596 (2003).

15 Daemen A, Peterson D, Sahu N *et al.* Metabolite profiling stratifies pancreatic ductal adenocarcinomas into subtypes with distinct sensitivities to metabolic inhibitors. *Proc. Natl Acad. Sci. USA* 112(32), E4410–E4417 (2015).

16 Li MJ, Wang P, Liu X *et al.* GWASdb: a database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res.* 40, D1047–D1054 (2011).

17 Calin GA, Croce CM. microRNA signatures in human cancers. *Nat. Rev. Cancer* 6(11), 857–866 (2006).

18 Ghosal S, Das S, Sen R, Basak P, Chakrabarti J. Circ2Traits: a comprehensive database for circular RNA potentially associated with disease and traits. *Front. Genet.* 4, 283 (2013).

19 Massart R, Barnea R, Dikshtein Y *et al.* Role of DNA methylation in the nucleus accumbens in incubation of cocaine craving. *J. Neurosci.* 35(21), 8042–8058 (2015).

20 Manolio TA. Genomewide association studies and assessment of the risk of disease. *N. Engl. J. Med.* 363(2), 166–176 (2010).

21 Jaenisch R, Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat. Genet.* 33(Suppl.), 245–254 (2003).

22 Sun SY, Liu ZP, Zeng T, Wang Y, Chen L. Spatio-temporal analysis of Type 2 diabetes mellitus based on differential expression networks. *Sci. Rep.* 3, 2268 (2013).

23 Zhang X, Gao L, Liu ZP, Chen L. Identifying module biomarker in Type 2 diabetes mellitus by discriminative area of functional activity. *BMC Bioinform.* 16, 92 (2015).

24 Hood L. Systems biology and p4 medicine: past, present, and future. *Rambam Maimonides Med. J.* 4(2), e0012 (2013).

25 Chin L, Andersen JN, Futreal PA. Cancer genomics: from discovery science to personalized medicine. *Nat. Med.* 17(3), 297–303 (2011).

26 Newman JR, Ghaemmaghami S, Ihmels J *et al.* Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. *Nature* 441(7095), 840–846 (2006).

27 Kinney JB, Atwal GS. Equitability, mutual information, and the maximal information coefficient. *Proc. Natl Acad. Sci. USA* 111(9), 3354–3359 (2014).

28 Klein RJ, Zeiss C, Chew EY *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science* 308(5720), 385–389 (2005).

29 Bernstein BE, Stamatoyannopoulos JA, Costello JF *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.* 28(10), 1045–1048 (2010).

30 Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270(5235), 467–470 (1995).

31 Wilkins MR, Pasquali C, Appel RD *et al.* From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. *Biotechnology* 14(1), 61–65 (1996).

32 Fromont-Racine M, Rain JC, Legrain P. Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. *Nat. Genet.* 16(3), 277–282 (1997).

33 Jonsson P, Johansson AI, Gullberg J *et al.* High-throughput data analysis for detecting and identifying differences between samples in GC/MS-based metabolomic analyses. *Anal. Chem.* 77(17), 5635–5642 (2005).

34 Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* 5(2), 101–113 (2004).

35 Barabasi AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* 12(1), 56–68 (2011).

36 Liu ZP, Wang Y, Zhang XS, Chen L. Network-based analysis of complex diseases. *IET Syst. Biol.* 6(1), 22–33 (2012).

37 Ideker T, Sharan R. Protein networks in disease. *Genome Res.* 18(4), 644–652 (2008).

38 Zeng T, Sun SY, Wang Y, Zhu H, Chen L. Network biomarkers reveal dysfunctional gene regulations during disease progression. *FEBS J.* 280(22), 5682–5695 (2013).

39 Network Cgar. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455(7216), 1061–1068 (2008).

40 Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* 43, D789–D798 (2014).

41 Welter D, Macarthur J, Morales J *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* 42, D1001–D1006 (2014).

42 Liu ZP, Wu C, Miao H, Wu H. RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database (Oxford)* 2015, bav095 (2015).

43 Liu ZP, Wu H, Zhu J, Miao H. Systematic identification of transcriptional and post-transcriptional regulations in human respiratory epithelial cells during influenza A virus infection. *BMC Bioinformatics* 15, 336 (2014).

44 Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 30(1), 207–210 (2002).

45 Brazma A, Parkinson H, Sarkans U *et al.* ArrayExpress – a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.* 31(1), 68–71 (2003).

46 Uhlen M, Oksvold P, Fagerberg L *et al.* Towards a knowledge-based Human Protein Atlas. *Nat. Biotechnol.* 28(12), 1248–1250 (2010).

47 Jones P, Cote RG, Martens L *et al.* PRIDE: a public repository of protein and peptide identifications for the proteomics community. *Nucleic Acids Res.* 34, D659–D663 (2006).

48 Von Mering C, Jensen LJ, Kuhn M *et al.* STRING 7 – recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res.* 35, D358–D362 (2007).

49 Stark C, Breitkreutz BJ, Reguly T *et al.* BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* 34 D535–D539 (2006).

50 Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 28(1), 27–30 (2000).

51 Joshi-Tope G, Gillespie M, Vastrik I *et al.* Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.* 33, D428–D432 (2005).

52 Liu CC, Tseng Yt, Li W *et al.* DiseaseConnect: a comprehensive web server for mechanism-based disease-disease connections. *Nucleic Acids Res.* 42, W137–W146 (2014).

53 Jeong SK, Kwon MS, Lee EY *et al.* BiomarkerDigger: a versatile disease proteome database and analysis platform for the identification of plasma cancer biomarkers. *Proteomics* 9(14), 3729–3740 (2009).

54 Shao C, Li M, Li X *et al.* A tool for biomarker discovery in the urinary proteome: a manually curated human and animal urine protein biomarker database. *Mol. Cell Proteomics* 10(11), M111 010975 (2011).

55 Wishart DS, Knox C, Guo AC *et al.* DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* 34, D668–D672 (2006).

56 Consortium EP. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 306(5696), 636–640 (2004).

57 Muers M. Functional genomics: the modENCODE guide to the genome. *Nat. Rev. Genet.* 12(2), 80 (2011).

58 Heng TS, Painter MW. The Immunological Genome Project: networks of gene expression in immune cells. *Nat. Immunol.* 9(10), 1091–1094 (2008).

59 Hudson TJ, Anderson W, Artez A *et al.* International network of cancer genome projects. *Nature* 464(7291), 993–998 (2010).

60 Strausberg RL, Buetow KH, Emmert-Buck MR, Klausner RD. The cancer genome anatomy project: building an annotated gene index. *Trends Genet.* 16(3), 103–106 (2000).

61 Mailman MD, Feolo M, Jin Y *et al.* The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.* 39(10), 1181–1186 (2007).

62 Benson DA, Cavanaugh M, Clark K *et al.* GenBank. *Nucleic Acids Res.* 41, D36–D42 (2012).

63 Ashburner M, Ball CA, Blake JA *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25(1), 25–29 (2000).

64 Wishart DS, Tzur D, Knox C *et al.* HMDB: the Human Metabolome Database. *Nucleic Acids Res.* 35, D521–D526 (2007).

65 Keshava Prasad TS, Goel R, Kandasamy K *et al.* Human Protein Reference Database–2009 update. *Nucleic Acids Res.* 37, D767–D772 (2009).

66 Yang IS, Ryu C, Cho KJ *et al.* IDBD: infectious disease biomarker database. *Nucleic Acids Res.* 36, D455–D460 (2008).

67 Aranda B, Achuthan P, Alam-Faruque Y *et al.* The IntAct molecular interaction database in 2010. *Nucleic Acids Res.* 38, D525–D531 (2010).

68 Kolker E, Higdon R, Haynes W *et al.* MOPED: Model Organism Protein Expression Database. *Nucleic Acids Res.* 40, D1093–D1099 (2012).

69 Cerami EG, Gross BE, Demir E *et al.* Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res.* 39, D685–D690 (2011).

70 Wang M, Weiss M, Simonovic M *et al.* PaxDb, a database of protein abundance averages across all three domains of life. *Mol. Cell. Proteomics* 11(8), 492–500 (2012).

71 Zhao F, Xuan Z, Liu L, Zhang MQ. TRED: a Transcriptional Regulatory Element Database and a platform for in silico gene regulation studies. *Nucleic Acids Res.* 33, D103–D107 (2005).

72 Consortium U. The universal protein resource (UniProt). *Nucleic Acids Res.* 36, D190–D195 (2008).

73 Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. *Science* 349(6245), 255–260 (2015).

74 Newman MEJ. The structure and function of complex networks. *SIAM Rev.* 45(2), 167–256 (2003).

75 Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. *Nature* 323(6088), 533–536 (1986).

76 Wang Y, Chen X, Liu ZP *et al. De novo* prediction of RNA-protein interactions from sequence information. *Mol. Biosyst.* 9(1), 133–142 (2013).

77 Cortes C, Vapnik V. Support-vector networks. *Machine Learning* 20(3), 273–297 (1995).

78 Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. *Machine Learning* 46(1–3), 389–422 (2002).

79 Breiman L. Random forests. *Machine Learning* 45(1), 5–32 (2001).

80 Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10(1), 57–63 (2009).

81 Golub TR, Slonim DK, Tamayo P *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286(5439), 531–537 (1999).

82 Van De Vijver MJ, He YD, Van't Veer LJ *et al.* A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* 347(25), 1999–2009 (2002).

83 Efron B, Tibshirani R. On testing the significance of sets of genes. *Ann. Appl. Stat.* 1(1), 107–129 (2007).

84 Subramanian A, Tamayo P, Mootha VK *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* 102(43), 15545–15550 (2005).

85 Tian L, Greenberg SA, Kong SW *et al.* Discovering statistically significant pathways in expression profiling studies. *Proc. Natl Acad. Sci. USA* 102(38), 13544–13549 (2005).

86 Bandyopadhyay S, Mehta M, Kuo D *et al.* Rewiring of genetic networks in response to DNA damage. *Science* 330(6009), 1385–1389 (2010).

87 Liu X, Liu ZP, Zhao XM, Chen L. Identifying disease genes and module biomarkers by differential interactions. *J. Am. Med. Inform. Assoc.* 19(2), 241–248 (2012).

88 Sahni N, Yi S, Taipale M *et al.* Widespread macromolecular interaction perturbations in human genetic disorders. *Cell* 161(3), 647–660 (2015).

89 Zhang W, Zeng T, Chen L. EdgeMarker: Identifying differentially correlated molecule pairs as edge-biomarkers. *J. Theor. Biol.* 362, 35–43 (2014).

90 Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.* 3, 140 (2007).

91 He D, Liu ZP, Chen L. Identification of dysfunctional modules and disease genes in congenital heart disease by a network-based approach. *BMC Genomics* 12, 592 (2011).

92 He D, Liu ZP, Honda M, Kaneko S, Chen L. Coexpression network analysis in chronic hepatitis B and C hepatic lesions reveals distinct patterns of disease progression to hepatocellular carcinoma. *J. Mol. Cell. Biol.* 4(3), 140–152 (2012).

93 Hofree M, Shen JP, Carter H, Gross A, Ideker T. Network-based stratification of tumor mutations. *Nat. Methods* 10(11), 1108–1115 (2013).

94 Liu ZP, Wang Y, Zhang XS, Xia W, Chen L. Detecting and analyzing differentially activated pathways in brain regions of Alzheimer's disease patients. *Mol. Biosyst.* 7(5), 1441–1452 (2011).

95 Segal E, Friedman N, Koller D, Regev A. A module map showing conditional activity of expression modules in cancer. *Nat. Genet.* 36(10), 1090–1098 (2004).

96 Taylor IW, Linding R, Warde-Farley D *et al.* Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat. Biotechnol.* 27(2), 199–204 (2009).

97 Wen Z, Liu ZP, Liu Z, Zhang Y, Chen L. An integrated approach to identify causal network modules of complex diseases with application to colorectal cancer. *J. Am. Med. Inform. Assoc.* 20(4), 659–667 (2012).

98 Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* 4, Article 17 (2005).

99 Lee E, Chuang HY, Kim JW, Ideker T, Lee D. Inferring pathway activity toward precise disease classification. *PLoS Comput. Biol.* 4(11), e1000217 (2008).

100 Liu KQ, Liu ZP, Hao JK, Chen L, Zhao XM. Identifying dysregulated pathways in cancers from pathway interaction networks. *BMC Bioinformatics* 13(1), 126 (2012).

101 Liu ZP, Zhang W, Horimoto K, Chen L. Gaussian graphical model for identifying significantly responsive regulatory networks from time course high-throughput data. *IET Syst. Biol.* 7(5), 143–152 (2013).

102 Chen L, Liu R, Liu ZP, Li M, Aihara K. Detecting early warning signals for sudden deterioration of complex diseases by dynamical network biomarkers. *Sci. Rep.* 2, 342 (2012).

103 Gosset WS. The probable error of a mean. *Biometrika* 6(1), 1–25 (1908).

104 Wilcoxon F. Individual comparisons by ranking methods. *Biometrics Bulletin* 1(6), 4 (1945).

105 Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA* 98(9), 5116–5121 (2001).

106 Gelman A. Analysis of variance – why it is more important than ever. *Ann. Stat.* 33(1), 54 (2005).

107 Bar-Joseph Z, Gitter A, Simon I. Studying and modelling dynamic biological processes using time-series gene expression data. *Nat. Rev. Genet.* 13(8), 552–564 (2012).

108 Robinson MD, Mccarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26(1), 139–140 (2010).

109 Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol.* 11(10), R106 (2010).

110 Park PJ. ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* 10(10), 669–680 (2009).

111 Wu S, Wu H. More powerful significant testing for time course gene expression data using functional principal component analysis approaches. *BMC Bioinformatics* 14, 6 (2013).

112 Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B* 57(1), 289–300 (1995).

113 Akaike H. A new look at the statistical model identification. *IEEE Transact. Automat. Control* 19(6), 8 (1974).

114 Gideon S. Estimating the dimension of a model. *Ann. Stat.* 6(2), 461–464 (1978).

115 Dreze M, Charloteaux B, Milstein S *et al.* 'Edgetic' perturbation of a *C. elegans* BCL2 ortholog. *Nat. Methods* 6(11), 843–849 (2009).

116 Califano A, Butte AJ, Friend S, Ideker T, Schadt E. Leveraging models of cell regulation and GWAS data in integrative network-based association studies. *Nat. Genet.* 44(8), 841–847 (2012).

117 Sotiriou C, Pusztai L. Gene-expression signatures in breast cancer. *N. Engl. J. Med.* 360(8), 790–800 (2009).

118 Shi X, Sun M, Liu H, Yao Y, Song Y. Long non-coding RNAs: a new frontier in the study of human diseases. *Cancer Lett.* 339(2), 159–166 (2013).

119 Zhong Q, Simonis N, Li QR *et al.* Edgetic perturbation models of human inherited disorders. *Mol. Syst. Biol.* 5, 321 (2009).

120 Zeng T, Zhang W, Yu X *et al.* Edge biomarkers for classification and prediction of phenotypes. *Sci. China Life Sci.* 57(11), 1103–1114 (2014).

121 Sahni N, Yi S, Zhong Q *et al.* Edgotype: a fundamental link between genotype and phenotype. *Curr. Opin. Genet. Dev.* 23(6), 649–657 (2013).

122 Liu Z-P, Wang Y, Wen T *et al.*Dynamically dysfunctional protein interactions in the development of Alzheimer's disease. Presented at: *Systems, Man and Cybernetics, SMC 2009. IEEE International Conference on Systems, Man and Cybernetics*. San Antonio, TX, USA, 11–14 October 2009.

123 Liu ZP, Wang Y, Zhang XS, Chen L. Identifying dysfunctional crosstalk of pathways in various regions of Alzheimer's disease brains. *BMC Syst. Biol.* 4 (Suppl. 2), S11 (2010).

124 Begum T, Ghosh TC. Elucidating the genotype-phenotype relationships and network perturbations of human shared and specific disease genes from an evolutionary perspective. *Genome Biol. Evol.* 6(10), 2741–2753 (2014).

125 He D, Liu ZP, Honda M, Kaneko S, Chen L. Coexpression network analysis in chronic hepatitis B and C hepatic lesion reveals distinct patterns of disease progression to hepatocellular carcinoma. *J. Mol. Cell Biol.* 4(3), 140–152 (2012).

126 Fujita A, Sato JR, Demasi MA *et al.*Comparing Pearson, Spearman and Hoeffding's D measure for gene expression association analysis. *J. Bioinform. Comput. Biol.* 7(4), 663–684 (2009).

127 Zhang X, Zhao XM, He K *et al.* Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. *Bioinformatics* 28(1), 98–104 (2012).

128 Kullback S, Leibler RA. On information and sufficiency. *Ann. Mathemat. Statistics* 21(2), 79–86 (1951).

129 Lo A, Chernoff H, Zheng T, Lo SH. Why significant variables aren't automatically good predictors. *Proc. Natl Acad. Sci. USA* 112(45), 13892–13897 (2015).

130 Chuang HY, Rassenti L, Salcedo M *et al.* Subnetwork-based analysis of chronic lymphocytic leukemia identifies pathways that associate with disease progression. *Blood* 120(13), 2639–2649 (2012).

131 Yang S, Fang X, Huang ZZ *et al.* Can serum glypican-3 be a biomarker for effective diagnosis of hepatocellular carcinoma? A meta-analysis of the literature. *Dis. Markers* 2014, 127831 (2014).

132 Zhou H, Saito S, Piao G *et al.* Network screening of Goto-Kakizaki rat liver microarray data during diabetic progression. *BMC Syst. Biol.* 5(Suppl. 1), S16 (2011).

133 Piao G, Saito S, Sun Y *et al.* A computational procedure for identifying master regulator candidates for diabetes progression in Goto-Kakizaki rat. *BMC Syst. Biol.* 6(Suppl. 1), S2 (2012).

134 Ashley EA. The precision medicine initiative: a new national effort. *JAMA* 313(21), 2119–2120 (2015).

135 Liu R, Li M, Liu ZP, Wu J, Chen L, Aihara K. Identifying critical transitions and their leading biomolecular networks in complex diseases. *Sci. Rep.* 2, 813 (2012).

136 Wang YC, Chen BS. A network-based biomarker approach for molecular investigation and diagnosis of lung cancer. *BMC Med. Genomics* 4, 2 (2011).

137 Wong YH, Chen RH, Chen BS. Core and specific network markers of carcinogenesis from multiple cancer samples. *J. Theor. Biol.* 362, 17–34 (2014).

138 Wong YH, Li CW, Chen BS. Evolution of network biomarkers from early to late stage bladder cancer samples. *Biomed. Res. Int.* 2014, 159078 (2014).

139 Mirnezami R, Nicholson J, Darzi A. Preparing for precision medicine. *N. Engl. J. Med.* 366(6), 489–491 (2012).

140 Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.* 97(457), 77–87 (2002).

141 Dudley JT, Tibshirani R, Deshpande T, Butte AJ. Disease signatures are robust across tissues and experiments. *Mol. Syst. Biol.* 5, 307 (2009).

142 Wu Z, Irizarry RA. Preprocessing of oligonucleotide array data. *Nat. Biotechnol.* 22(6), 656–658; author reply: 658 (2004).

143 Loven J, Orlando DA, Sigova AA *et al.* Revisiting global gene expression analysis. *Cell* 151(3), 476–482 (2012).

144 Robert C, Watson M. Errors in RNA-Seq quantification affect genes of relevance to human disease. *Genome Biol.* 16, 177 (2015).

145 Wang B, Mezlini AM, Demir F *et al.* Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* 11(3), 333–337 (2014).

146 Schadt EE, Lamb J, Yang X *et al.* An integrative genomics approach to infer causal associations between gene expression and disease. *Nat. Genet.* 37(7), 710–717 (2005).