OXFORD

Briefings in Bioinformatics, 20(3), 2019, 952–975

doi: 10.1093/bib/bbx158
Advance Access Publication Date: 29 November 2017
Paper

Computer-aided biomarker discovery for precision medicine: data resources, models and applications

Yuxin Lin*, Fuliang Qian*, Li Shen, Feifei Chen, Jiajia Chen and Bairong Shen

Corresponding author: Bairong Shen, Center for Systems Biology, Soochow University, Postbox 206, No1. Shizi Street, Suzhou, Jiangsu 215006, China. Tel.: +86-512-65110951; Fax: +86-512-65110951; E-mail: bairong.shen@suda.edu.cn
*These authors contributed equally to this work.

Abstract

Biomarkers are a class of measurable and evaluable indicators with the potential to predict disease initiation and progression. In contrast to disease-associated factors, biomarkers hold the promise to capture the changeable signatures of biological states. With methodological advances, computer-aided biomarker discovery has now become a burgeoning paradigm in the field of biomedical science. In recent years, the 'big data' term has accumulated for the systematical investigation of complex biological phenomena and promoted the flourishing of computational methods for systems-level biomarker screening. Compared with routine wet-lab experiments, bioinformatics approaches are more efficient to decode disease pathogenesis under a holistic framework, which is propitious to identify biomarkers ranging from single molecules to molecular networks for disease diagnosis, prognosis and therapy. In this review, the concept and characteristics of typical biomarker types, e.g. single molecular biomarkers, module/network biomarkers, cross-level biomarkers, etc., are explicated on the guidance of systems biology. Then, publicly available data resources together with some well-constructed biomarker databases and knowledge bases are introduced. Biomarker identification models using mathematical, network and machine learning theories are sequentially discussed. Based on network substructural and functional evidences, a novel bioinformatics model is particularly highlighted for microRNA biomarker discovery. This article aims to give deep insights into the advantages and challenges of current computational approaches for biomarker detection, and to light up the future wisdom toward precision medicine and nation-wide healthcare.

Key words: molecular biomarkers; databases and knowledge bases; bioinformatics models; precision medicine; systems biology

Introduction

Biological markers, also known as markers or biomarkers, are objectively measurable and evaluable indicators of certain biological states in normal and pathogenic processes, or possible pharmacologic responses to therapeutics [1, 2]. From a medical point of view, biomarkers are traceable substances with the ability to classify binary conditions (e.g. normal and disease states, etc.) or multi-conditions (e.g. disease stages, etc.) of diseases, which are of great significance for organism activity

Yuxin Lin is a PhD candidate at Center for Systems Biology, Soochow University. His research interest is in developing computational models for molecular biomarker discovery.
Fuliang Qian is a PhD candidate at Center for Systems Biology, Soochow University. His research interest is in developing computational models for molecular biomarker discovery.
Li Shen is a research assistant both at Center for Systems Biology of Soochow University and Yale University School of Medicine. His current interest is cancer heterogeneity analysis.
Feifei Chen is a graduate student at Center for Systems Biology, Soochow University. Her main interest is pan-cancer analysis on microRNA-mRNA regulatory mechanisms.
Jiajia Chen is an associate professor at School of Chemistry, Biology and Material Engineering, Suzhou University of Science and Technology. Her research is systems biology.
Bairong Shen is a professor and director of Center for Systems Biology, Soochow University. He is also a part-time professor of Guizhou University School of Medicine. His research interest is bioinformatics and medical systems biology.
Submitted: 2 September 2017; **Received (in revised form):** 17 October 2017

examination. In clinical trials, sensitive and specific biomarkers provide functional insights in disease pathogenesis and facilitate the progress of disease surveillance and health management. Compared with routine disease-associated factors, biomarkers predict disease incidence or progression not only at the expression level, more importantly, they hold the potential to indicate the dynamic change of biological processes or states [3].

According to the systems biology paradigm, in this article, biomarkers are classified as molecular biomarkers, clinical phenotype biomarkers and cross-level biomarkers to cater the research interest from the level of 'molecule/cell' to that of 'individual/population'. As one of the most common types, molecular biomarkers, e.g. key genes, RNAs, proteins and metabolite molecules from tissues, blood as well as other body fluids, are well investigated across different studies. Currently, accumulating evidence convinced that profiles of genetic mutations, epigenetic modifications and aberrant expression of biological molecules at different omics levels are powerful tools for monitoring the change of body states, which are propitious to disease early diagnosis, prognostic tracking, targeted drug design and personalized therapy. For example, mutations in gene *BRCA1* and *BRCA2* are highly connected with the development of inherited breast cancer, especially for women over 50 years of age [4]. The methylation of circulating DNA can be used to identify tissue-specific cell death [5]. For example, oligodendrocyte DNA and exocrine pancreas DNA, respectively, are screened in patients with relapsing multiple sclerosis and pancreatic cancer, which open the window for the diagnosis and monitoring of human pathologies [5]. Moreover, *SPG20* gene promoter methylated DNA in plasma is a functional epigenetic biomarker for colorectal cancer (CRC) diagnosis [6]. *HAND2* methylation occurs commonly in endometrial cancer. It could be a diagnostic biomarker and an indicator for predicting the treatment response [7]. The dysregulation of many typical RNAs, such as messenger RNAs (mRNAs), microRNAs (miRNAs), long noncoding RNAs (lncRNAs), piwi-interacting RNAs (piRNAs), small nucleolar RNAs (snoRNAs) and circular RNAs (circRNAs) were also reported to serve as biomarkers. For example, *miR-21-5p* [8], *miR-139-3p* [9], *MALAT1*, *AFAP1-AS1* and *AL359062* [10] play pathogenic roles in disease development. In the past few years, there have been growing focus on the biomarker potential of piRNAs, snoRNAs and circRNAs. Nogueira *et al.* [11] found that the change of piRNA and snoRNA expression levels caused by tobacco could be used to detect lung adenocarcinoma for women. The piRNAs may regulate mRNAs and work in an epigenetic way. Moreover, the use of piRNAs and *PIWI* proteins had latent clinical implications for cancer diagnosis, prognosis and therapy [12]. As an emerging star in RNA world, circRNA has gradually been recognized as a novel biomarker type for diseases [13, 14]. For instance, circRNA_100876 up-regulation was significantly correlated with the shorter overall survival time of patients with non-small cell lung cancer (NSCLC), which implied its biomarker value for NSCLC prognosis and therapeutics [15].

Nowadays, great efforts have been devoted to screening molecular signatures for disease prevention via computational or bioinformatics approaches. In contrast to traditional wet-lab experiments, computer-aided molecular biomarker discovery integrates various data resources and biological knowledge into a holistic framework. It is more efficient to explore the ability and function of biological molecules as well as their interactions, thereby contributing to the systematical understanding of disease evolution. In the era of big data, increasingly accumulated data resources promote the development of biomedical informatics and translational medicine, and novel models or technologies are constantly designed for capturing the crucial signals associated with health–disease homeostasis. Computational strategies based on mathematics, network knowledge and machine learning are widely used for biomarker discovery. For instance, statistical tests, such as Student's *t*-test, Significance Analysis Microarray (SAM), empirical Bayes (eBayes), etc., are classical ways for differentially expressed (DE) gene screening [16, 17], which is often the first step of data processing in many bioinformatics models. Along with network-based methods which highly depend on network topological and functional features [18, 19], some machine learning algorithms, e.g. support vector machine (SVM), Random forest (RF), Clustering, etc., are often applied to mining key players that affect the stability and function of biological systems from large-scale expression data [20–22]. Moreover, integrated methods, e.g. network-regularized sparse logistic regression [23] and network smoothed *t*-statistic SVMs (stSVMs) [24], showed accurate performance on cancer gene biomarker detection. It is rather remarkable that the identification of molecular biomarkers is not limited to single static molecules. Since the dynamic and personalized nature of disease formation and progression, presently network biomarkers at serial time points or disease stages are gradually acknowledged as effective monitors for predicting the abnormal interplay among different biological components within human system, and can be exploited to accurately detect disease states and make personalized clinical decisions for patients [25].

This article intends to give a comprehensive review on computer-aided biomarker discovery. First, biomarker classification based on systems biology theory is briefly explicated. Here, the concept and function of molecular biomarkers are classified as single molecular biomarkers, edge/module biomarkers and network biomarkers. Data resources available for biomarker discovery are sequentially introduced. Computational models, software tools as well as their translational applications toward precision medicine and personalized healthcare are then presented. Besides well-performed methods using mathematical, network or machine learning theories as principles, a novel evidence-based bioinformatics model, which integrates network vulnerability information and biological knowledge for miRNA biomarker discovery, is systematically discussed. Considering the complexity and heterogeneity of complex diseases, innovations in thinking modes and technical routes are kernels for the development of these approaches, and future perspectives on the improvement of computer-aided biomarker discovery are proposed in the last section.

Systems biology-based biomarker classification

Molecular biomarkers

Single molecular biomarkers (node biomarkers)

Single molecular biomarkers, also termed as node biomarkers according to the network theory, are isolated but sensitive molecules specific to certain diseases. Generally, genomic structural variation (e.g. base insertion, deletion, substitution, etc.), abnormal expression and regulation of these molecules have potentially positive or negative effects on biological activity, thus are functional factors to indicate disease formation or differentiate disease conditions (Figure 1).

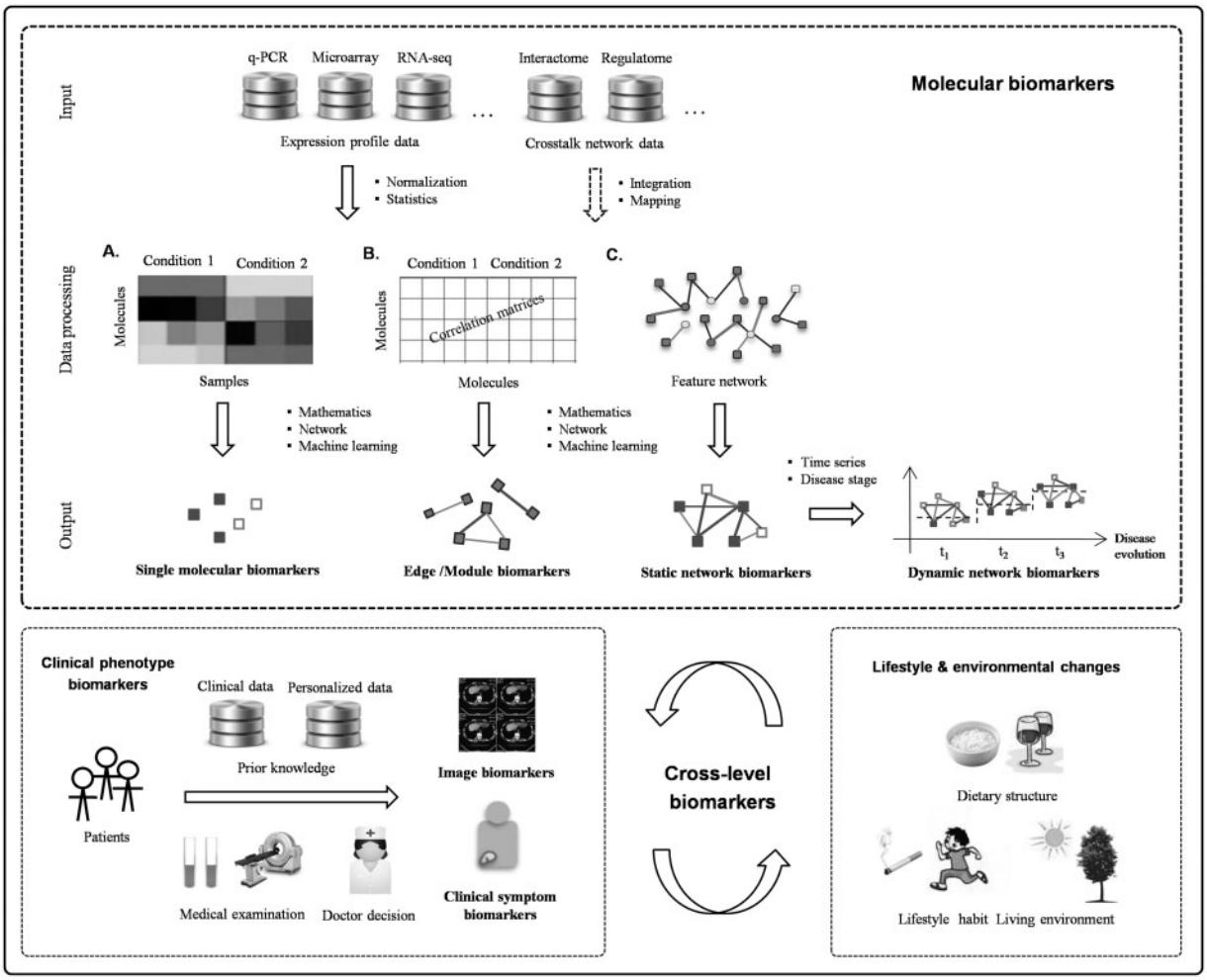

Figure 1. Biomarker classification based on systems biology viewpoints. Molecular biomarkers described here can be categorized into three subtypes: single molecular biomarkers, edge/module biomarkers and network biomarkers. The integration of molecular biomarkers, clinical phenotype biomarkers and lifestyle/environmental factors constitutes the concept of cross-level biomarkers, which are of great significance for precision medicine.
q-PCR: quantitative polymerase chain reaction.

Currently, most of studies are focusing on identifying single molecules as biomarkers for disease pathogenesis tracking as well as clinical decision making because single molecules are easy to be detected from multidimensional experimental data (e.g. microarray, next-generation sequencing, low-throughput experiments, etc.) using technical methods. For example, screening DE molecules between case and control groups and then prioritizing the candidates through statistical or other advanced computational approaches, such as Bayesian reasoning and machine learning. As the expression of these molecules can be directly measured for disease prediction, it would be beneficial to the translation from basic research to clinical practice. For instance, prostate-specific antigen is a single-strand glycoprotein widely used in clinical diagnosis of prostate cancer (PCa) [26, 27]. Serum alanine aminotransferase, an important enzyme, has become a screening tool for the primary detection of live diseases such as acute liver injury [28]. In recent years, a large number of biomarker molecules are identified because of the improvement of both experimental methods and computational techniques. Gupta *et al.* [29] found that lncRNA HOX transcript antisense RNA (*HOTAIR*) could promote cancer metastasis by reprograming chromatin state. In the study by Teschendorff *et al.* [30], this lncRNA or its surrogate DNA methylation could be a biomarker and a target to overcome the resistance of carboplatin in ovarian cancer. Wang *et al.* [31] performed the quantitative real-time polymerase chain reaction experiment on 70 samples from esophageal squamous cell carcinoma (ESCC) patients and healthy volunteers. They found that serum *HOTAIR* also had oncogenic role in ESCC. In contrast with healthy controls, it was significantly overexpressed in ESCC patients and can serve as a potential biomarker for ESCC diagnosis. Dalerba *et al.* [32] ranked genes for subtyping of colon cancer with a new bioinformatics approach integrating both gene-expression array and clinical-grade diagnostic assay data. They demonstrated that the transcription factor (TF) *CDX2* was a functionally prognostic biomarker in Stages II and III colon cancer, which would help the selection of patients who require further adjuvant treatment after surgical intervention [32]. Compared with adjacent liver tissues, *circRNA_0001649* was significantly downregulated in hepatocellular carcinoma (HCC), and it may exert functions in HCC carcinogenesis as well as metastasis [33].

Although single molecular biomarkers play important roles in disease surveillance, they are still unable to capture the dynamic and heterogeneous signature during disease evolution. Complex diseases are always caused by multigenetic factors

[34, 35]; thus, the loss of features from biochemical reactions among multiple molecules may limit the display of a systematical picture toward precision medicine.

Edge, motif and module biomarkers

Traditional node-based biomarkers mostly focused on molecules whose expression levels are differed in variable conditional populations, and ignored the functional importance of those without differential expression. In fact, molecules in living organisms interplay in a network manner. Interactions among different molecules affect the behavior of biological system and create the complex biological phenomena. Accumulating efforts convinced that the rewiring and edges within biological networks are of great significance in response to cellular signal transduction and other vital activities. For instance, Bandyopadhyay *et al.* [36] found that differential interactions among yeast kinases, TFs and phosphatases were important cell responses to DNA damage. Creixell *et al.* [37] indicated that perturbations of interactions in signaling networks were able to create new phenotypic states. For example, network-attacking mutations could induce cancer signaling rewiring, which were helpful for tumor-specific cancer therapy [38]. Moreover, the rewiring of signaling networks may cause the alteration of signal transduction, resulting in cell death, cancer metastasis and other biological behaviors [39, 40]. Recent studies on 'edgetics' also elucidated the linkage role of 'edgotype' between genotype and phenotype [41, 42]. Based on these findings, the concept of edge biomarkers, i.e. the differentially correlated molecular pairs (e.g. gene pairs, RNA pairs, etc.), was presented to better classify biological states for disease early diagnosis [43]. As shown in Figure 1, unlike routine single molecular biomarkers, molecules in edge biomarkers are not always DE, but their expression correlations should be significantly different in case and control groups. Such transformation from node space ('node expression' data) to edge space ('edge expression' data) hold the optimal ability for sample classification and can predict the alternation of biological states in an accurate way [43].

Motifs and modules are substructures with special functions in a network. In biology, key motifs or modules play critical roles in biological processes and are important indicators for disease evaluation. Compared with node- and edge-based biomarkers, in motif/module biomarkers, DE genes and differentially correlated pairs (DCPs) were combined interactively into an systematical framework [25], which provided an integrated window for understanding vital activities, such as gene expression, cell cycle, tumor microenvironment and drug responses to disease therapy [44]. For example, Zhang *et al.* [45] identified the TF and miRNA coregulatory loops as important TF-miRNA regulatory motifs, where transcriptional activation or repression of these regulators was able to alter downstream gene expression in cell proliferation and differentiation process. Moreover, abnormal regulation of TFs and miRNAs in the motifs are strongly related to the development of diseases, such as cancers, schizophrenia and interstitial lung disease [45]. Cui *et al.* [46] extracted nine long intergenic noncoding RNA (lincRNA) modules from PCa RNA sequencing (RNA-seq) data. They found that the principle component of three of the screened modules tended to be highly associated with PCa phenotype, and one module can be used as lincRNA module biomarker for PCa diagnosis [46].

Network biomarkers

The development of diseases is often the consequence of systems-level disorders rather than the breakdown of single molecules. Therefore, the identification of single molecular biomarkers from samples of patients with same diseases often tends to be highly heterogeneous because of the pathogenic complexity of diseases [47]. To address this issue, network biomarkers, which comprise the dysfunctional information of molecules (nodes) as well as their correlations (edges), have found to be powerful for investigating the underlying mechanisms in health–disease and disease–disease transition. Network biomarkers can be regarded as the further expansion of module biomarkers, in which the interactome from multiple metamodules contributes to the systematical understanding of disease etiology and therapeutics, stimulating the creation of novel strategies for analyzing genetic or epigenetic changes across diseases [25, 48].

Network biomarkers can be subclassified into static network biomarkers (SNBs) and dynamical network biomarkers (DNBs). As described in Figure 1, SNBs only focus on the static nature of networks and cannot reveal the dynamic change during disease progression, whereas DNBs pay attention to the disease state at different time points or disease stages, presenting molecules and their relationship in a three-dimensional image for stage-specific or personalized signature discovery [49, 50]. Although derived formally and mathematically for a dynamical system, it should be noticed that the DNB framework is not applied solely to time-course data. In fact, the time-course data, which entail measurements of molecular variables in time for the 'same' individual, are limited. Fortunately, the DNB formalism is still powerful even in the absence of time-course data. For example, it can be applied to cross-sectional cohort data representing different disease stages, i.e. molecular measurements for plenty of individuals representing different disease stages, without the need to measure the same individual across all disease stages.

The method for network biomarker construction is often based on mathematics and machine learning. Usually, a reference network, such as protein–protein interaction (PPI) network, miRNA-mRNA regulatory network, etc., is used for condition-specific network extraction. For instance, Chuang *et al.* [51] identified a network biomarker from integrated PPI and gene expression data using a protein network-based approach. The resulting biomarker networks are enriched in pathways closely associated with tumor growth and have the ability to distinguish metastatic breast cancer from those of nonmetastatic groups. Moreover, they found that genes with known breast cancer mutations were not significantly DE, but were almost kernels in the network through interacting with many DE genes [51]. Chen *et al.* [49] partitioned disease progression into three stages as before-transition state (normal state), pre-transition state (pre-disease state) and after-transition state (disease state), where the first and last states were relatively stable compared with the pre-disease state, indicating the drastic state change of diseases. In contrast to traditional molecular biomarkers which used the expression value to distinguish disease and normal states, DNBs held the power to capture pre-disease signal through molecular fluctuation during different time periods, and showed accurate prediction on both type 1 and type 2 diabetes studies [49, 52].

Network biomarkers cater to the natural law of disease evolution and offer new opportunities for systems-level disease pathogenesis understanding. However, challenges exist in terms of the cost and efficacy for clinical test. Thus, an optimal scoring strategy for evaluating the performance of the

constructed network, i.e. network size, prediction sensitivity and specificity, etc., is urgently needed.

Clinical phenotype biomarkers

The mutation and expression profile data support the molecule-level biomarker discovery. In clinical practice, typical phenotypes are often the most intuitive factors used for disease diagnosis and treatment. As shown in Figure 1, clinical phenotype biomarkers cover a broad range of topics, including image screening (e.g. irregular cell, tissue or organ activities identified by computed tomography, nuclear magnetic resonance, electrocardiogram and electroencephalogram, or some well-performed image analysis techniques, etc.) and individual symptom analysis (e.g. body pain, fever, bleeding, etc.). For example, Galbán *et al*. [53] uncovered the unique signature for phenotype diagnosis of chronic obstructive pulmonary disease (COPD). They used a voxel-wise image analysis technique called parametric response map (PRM) to analyze the whole-lung inspiration and expiration image protocol. After image processing, PRM effectively classified COPD phenotypes based on the density feature of lung parenchyma acquired by computed tomography scans and provided spatial information related to disease distribution and location for COPD personalized diagnosis [53]. Apart from image biomarkers, some individual symptoms are sensitive to disease outcome. For example, body fever is a common symptom in humans; however, prolonged low fever with serious weight loss may be a negative indicator for health [54, 55]. Meanwhile, the states of pain characteristics, including the pain intensity, interference and relief, are able to predict the change in patients undergoing cancer radiotherapy [56].

The change of clinical phenotypes, in certain circumstances, can be caused by molecular disorders under complex pathogenic mechanisms. Thus, a combinative application of phenotype and molecular biomarkers that integrates clinical observation with genotype knowledge would promote the systematic understanding of disease pathobiology and ultimately accelerate the development of precision medicine.

Cross-level biomarkers

Disease occurrence and progression are not solely induced by genetic variation. A combination of extrinsic factors such as lifestyle habits and living environments gives pathogenic insights in disease management, which are of considerable meaning for people healthcare [57]. A recent study indicated that interactions between internal and external environments could result in epigenetic changes, which were solidified and propagated during cell division and finally affected the maintenance of the phenotype [58]. The environment holds the potential to induce individual genotype change. For example, the mutation spectra in multiple synchronous lung tumors between smokers and nonsmokers could be heterogeneous, where C→A and C→T substitutions are, respectively, detected in smoking and nonsmoking patients [47].

In view of systems biology, the development of diseases is determined by a bunch of risk factors. Therefore, it would be better to construct the cross-level biomarker for the systematical prediction of diseases. As illustrated in Figure 1, molecular, image and clinical symptom (unified as clinical phenotypes) biomarkers, respectively, reveal body states at the molecular, cellar and individual level, whereas the change of surrounding environments provides crucial signature for population-based disease analysis [59]. In theory, the cross-level biomarker is macroscopical, which integrates the hallmark from multidimensional biological components and complements the advantages of discrete biomarker types. For instance, the shape and size of endosomatic tumors can be clearly imaged by X-ray, but it would fail to differentiate the grade of malignancy, or even misjudge the malignant tumor and benign diseases in some cases; hence, the further use of molecular test as well as environmental prevention can help break the suspicion and guide the diagnosis and therapy in a systems medicine way.

Databases and knowledge bases

Data resources for biomarker discovery

Molecular data

The advent of high-throughput technologies provides huge molecular data for biomarker discovery. Table 1 lists the publicly available molecular data resources for human interactome, regulatome and diseasome. For example, STRING [60], BioGRID [61], HPRD [62], PINA [63] and TissueNet [64] are five valuable PPI databases. Among them, STRING integrates both experimentally validated and predicted PPI data for cellular function analysis. Here, PPIs derived from coexpression analysis, genomic signals, published literatures and gene orthology knowledge are quality-controlled, which highly enhances the data reliability [60]. BioGRID curates genetic and protein interaction data of model organisms and humans from >20 000 publications, and it highlights the chance for health-related conserved network and pathway studies [61]. HPRD is a comprehensive database for human proteome decoding. It offers protein reference annotation and allows researchers to screen the presence of validated phosphorylation motifs in proteins [62]. PINA unifies PPI associations from six public databases, and its new version (v2.0) encourages the mining of interactome modules from the whole PPI network using clustering approaches [63]. TissueNet considers tissue-associated contexts of experimentally validated PPIs for human protein study [64]. Apart from these PPI-based databases, the database of genotypes and phenotypes (dbGaP) [65] is developed for investigating the interactions between genotypes and phenotypes. It fuses the information associated with the accessioned objects, phenotypes and different molecular assays, such as expression, sequence and epigenetic marks [65]. KEGG [66] and IPA [67], respectively, are well-designed knowledge bases and functional tools for the ingenuity study of genes, genomes and biological pathways.

In contrast to interactome resources, most of the regulatome databases share the information related to biological regulators (e.g. miRNAs, TFs, etc.) as well as their targets. For example, miRTarbase [68], DIANA-TarBase [69], starBase [70] and TransmiR [71] are manually curated databases that can be used for miRNA regulatory mechanism exploration. Besides the interplay among miRNAs and mRNAs, the starBase decodes interactions within noncoding RNAs (e.g. lncRNAs, circRNAs, pseudogenes, etc.) and competing endogenous RNAs (ceRNAs) from large-scale CLIP-Seq data [70], whereas TransmiR supports the retrieval of TF-miRNA regulations for transcriptional-level disease pathology analyses [71]. DIANA-LncBase focuses on miRNA targets on lncRNA transcripts. It indexes miRNA-lncRNA regulation from both experimentally validated and computationally predicted data sets [72]. miRSponge database contains miRNA sponges, such as lncRNAs, circRNAs, coding RNAs and pseudogenes, supported by biological experiments [73]. CircNet is a comprehensive database of circRNAs. In addition to novel circRNAs, the database provides the expression profiles and

Table 1. Publicly available data resources for biomarker discovery

Title/name	Description	URL	PMID
Interactome data			
STRING	• A database integrating quality-controlled PPI data for cellular function analysis	http://string-db.org/	27924014
BioGRID	• The Biological General Repository for Interaction Data sets • A database sharing genetic and protein interaction data	http://www.thebiogrid.org/	21071413
HPRD	• Human Protein Reference Database • A database offering protein reference information for human proteome studies	http://www.hprd.org/	18988627
PINA	• The Protein Interaction Network Analysis • A Web platform integrateing PPI data from six public databases for comprehensive network studies	http://cbg.garvan.unsw.edu.au/pina/	22067443
TissueNet	• Tissue-associated protein–protein network • A database providing PPI data with tissue contexts for human protein study	http://netbio.bgu.ac.il/tissuenet	27899616
dbGaP	• The Database of Genotypes and Phenotypes • A database integrating genotype and phenotype interaction data from public studies	http://www.ncbi.nlm.nih.gov/gap/	24297256
KEGG	• The Kyoto Encyclopedia of Genes and Genomes • An encyclopedia and knowledge base for the functional study of genes and genomes.	http://www.kegg.jp/	27899662
IPA	• Ingenuity Pathway Analysis • A knowledge base and software tool for the ingenuity analysis of biological pathways	http://www.ingenuity.com/	24336805
Regulatome data			
miRTarBase	• A database of experimentally validated miRNA–target regulations	http://miRTarBase.mbc.nctu.edu.tw/	26590260
DIANA-TarBase	• A manually curated database providing a large number of experimentally validated miRNA–gene interactions	http://www.microrna.gr/tarbase/	25416803
starBase	• A comprehensive database for identification of RNA-RNA and protein–RNA interactions from CLIP-Seq data	http://starbase.sysu.edu.cn/	24297251
TransmiR	• Transcription-factor-miRNA regulatory database • A manually curated database containing regulatory data among TFs and miRNAs	http://cmbi.bjmu.edu.cn/transmir/	19786497
DIANA-LncBase	• A database indexing miRNA-lncRNA regulation from both validated and predicted data sets	http://www.microrna.gr/LncBase/	26612864
miRSponge	• A manually curated database for experimentally validated miRNA sponges, including lncRNAs, circRNAs, coding RNAs, pseudogenes, etc.	http://www.biobigdata.net/miRSponge/	26424084
CircNet	• A database of circRNAs, their expression profiles and circRNA–miRNA–gene regulatory networks	http://circnet.mbc.nctu.edu.tw/	26450965
ENCODE	• Encyclopedia of DNA Elements • A comprehensive database providing functional interpretations of the human whole genome	http://genome.ucsc.edu/encode/	22075998
EpiFactors	• Database of Epigenetic Factors • A database of epigenetic regulators, their complexes as well as products	http://epifactors.autosome.ru/	26153137
Diseasome data			
GEO	• Gene Expression Omnibus • A functional repository for the submission, storage and retrieval of gene expression data	http://www.ncbi.nlm.nih.gov/geo/	11752295
TCGA	• The Cancer Genome Atlas • A comprehensive data platform and knowledge base for cancer studies	http://cancergenome.nih.gov/	25691825
miR2Disease	• A manually curated database of dysfunctional miRNAs in human diseases	http://www.miR2Disease.org/	18927107
HMDD	• Human microRNA Disease Database • A database collecting experimentally validated associations between miRNAs and diseases	http://cmbi.bjmu.edu.cn/hmdd/	24194601
CancerNet	• A database decoding miRNA-miRNA functionally synergistic interactions across 33 cancer types	http://bis.zju.edu.cn/CancerNet/	26690544
LncRNADisease	• A database collecting both experimentally supported and computationally predicted lncRNA and disease associations	http://cmbi.bjmu.edu.cn/lncrnadisease/	23175614
Circ2Traits	• The first comprehensive knowledgebase for circRNA and disease associations	http://gyanxet-beta.com/circdb/	24339831
OMIM	• Online Mendelian Inheritance in Man • A knowledge base of human genes and their associated disorders	http://www.ncbi.nlm.nih.gov/omim/	15608251
CIViC	• Clinical Interpretation of Variants in Cancer	http://civicdb.org/	28138153

Continued

Table 1. (continued)

Title/name	Description	URL	PMID
	• A community knowledge base for the accurate interpretation of genetic variants in cancers		
dbEM	• The Database of Epigenetic Modifiers • A database archiving epigenetic modifier data for cancer therapy	http://crdd.osdd.net/raghava/dbem/	26777304
HEDD	• The Human Epigenetic Drug Database • A database of epigenetic drug data set from experiments and literature reports	http://hedds.org/	28025347
DO	• The Human Disease Ontology • A database offering standardized ontology for human diseases	http://www.disease-ontology.org	25348409
Clinical phenotype data			
SEER	• Surveillance, Epidemiology and End Results Program • An authoritative information resource on cancer incidence and survival in the United States	https://seer.cancer.gov/	10613347
TCIA	• The Cancer Imaging Archive • An open-access information resource of advanced medical imaging of cancer	https://www.cancerimagingarchive.net	23884657
PathologyImagebase	• A reference image database providing international standard for pathology studies	www.isupweb.org	28722802
ACTuDB	• Database of array-CGH data for tumors • A public database for the integrated analysis of tumor genomic profiling and clinical data	http://bioinfo.curie.fr/actudb/	17496932
NSNT	• The Neoplasms of the Sinonasal Tract software package • A database and tool for storage and analysis of clinical and image data associated with sinonasal tract neoplasms	http://www.nsntsoftware.com/	15112979
BreCAN-DB	• An integrated repository and web tool for the analysis and visualization of personalized DNA breakpoint profiles of cancer genomes	http://brecandb.igib.res.in/	26586806
G-DOC	• Georgetown Database of Cancer • A platform integrating patient, clinical and high-throughput data for personalized oncology	https://gdoc.georgetown.edu/	21969811
DEER	• A database for interpreting the functional effects of chemical environmental factors on drug responses	http://bsb.kiz.ac.cn: 90/DEER/	23500449
miREnvironment	• A database of information related to miRNAs, environmental factors and their interactions on abnormal phenotypes and diseases	http://cmbi.bjmu.edu.cn/miren/	21984757

PMID, PubMed ID; URL, Uniform Resource Locator.

genomic annotations of circRNA isoforms. More importantly, it integrates network information to illustrate the regulatory pattern among circRNAs, miRNAs and genes [74]. Two other important regulatome databases are ENCODE [75] and EpiFactors [76]. The former is short for Encyclopedia of DNA Elements, which encompasses a diverse set of biological assays and provides visualized functional interpretations of human whole genome [75]. The latter is a famous database of epigenetic factors, where epigenetic regulators, their complexes as well as products are structurally organized [76].

With the advance of biomedical informatics, the amount of diseasome data is increasing [77]. Collaborative efforts for the construction of diseasome repertories accelerate the pace of data sharing and promote disease etiologic studies. GEO, for example, is a gene expression omnibus that allows the submission, storage and retrieval of heterogeneous gene expression data sets. Three entities, i.e. experimental platforms, sample information and data set association, are considered to be central for data organization and guide further disease analysis [78]. The Cancer Genome Atlas (TCGA) is a comprehensive knowledge base especially for cancer studies. It has generated the multidimensional maps relevant to the key genomic changes in 33 cancer types [79]. Some databases, in contrast, focus on gathering interactive connections between molecules and diseases. For instance, miR2Disease [80] and HMDD [81] provide the information of dysfunctional miRNAs in human diseases; CancerNet [82] identifies specific miRNA-miRNA functionally synergistic pairs across different cancer types; LncRNADisease [83] integrates experimentally supported and computationally predicted lncRNA–disease associations. Circ2Traits is the first knowledge base that records disease–circRNA associations based on circRNA-miRNA interactions or disease-associated single-nucleotide polymorphism (SNP) information [84]. In addition to noncoding RNAs, the malfunction of genes is often the leading cause of many diseases. The OMIM database, for example, gains attention to the relationship between genes and genetic disorders and brings burgeoning insights into human genetics [85]. CIViC is a comprehensive community knowledge base that provides the accurate interpretation of genetic variants for cancer study [86]. Besides, several epigenetic databases like dbEM [87] and HEDD [88] archive the epigenetic modifier or drug data set for disease therapy. The DO (Disease Ontology) database [89], importantly, offers standardized concepts of human diseases for biomedical knowledge transition.

Clinical phenotype data

The clinical phenotype data place emphasis on the significance of medical images, clinical indications and environmental effects for personalized biomarker identification. As illustrated in Table 1, most of these data are propitious to the systematic defense against cancer. For example, the Surveillance, Epidemiology, and End Results (SEER) Program is an authoritative resource for evaluating cancer incidence and mortality in the United States. The program provides continuing population-based information on the change of disease

diagnosis, therapy and patient survival, and promotes the detection of prophylactic factors pertaining to the patterns of lifestyle, environment and healthcare that are amenable to cancer intervention [90]. TCIA is an advanced imaging archive for cancers. More than 3.3 million images are recorded in the first year of operation. Data in the database are open-access for high-level analysis [91]. Pathology Imagebase is a reference image database aiming at developing the international standardization to promote the consistency of pathology studies [92]. ACTuDB is a public database integrating both genomic hybridization and clinical data for tumors and cell lines. It is a useful resource for tumor genomics study [93]. NSNT is a database with clinical and image information specific to sinonasal tract neoplasm [94]. Besides, BreCAN-DB [95] and G-DOC [96] are online systems medicine platforms for the personalized analysis of cancer genomes, whereas DEER [97] and miREnvironment [98], respectively, raise interests in the influence of environmental factors on drug responses and clinical phenotypes, which support the goal of cross-level biomarker discovery.

Biomarker databases and knowledge bases

Biomarker repertories for cancers

For decades, plenty of work is being carried out to identify biomarker signals for cancer management. Thus, building a database to integrate the existing findings from scattered studies is of important theoretical and practical value. As shown in Table 2, biomarkers in the databases can be separated into two subtypes, one is related to genetic mutations such as SNP, and the other mainly involves the expression change of molecules, i.e. significantly upregulation or downregulation of genes, miRNAs or other molecules. Ouyang *et al.* [99] created a database of prognostic biomarkers for hepatocellular carcinoma (dbPHCC). Based on the manual collection of citations in PubMed, 323 proteins, 154 genes and 90 miRNAs with expression pattern changes in low-throughput experiments are carefully recorded. Moreover, the database provides online analysis tools for HCC prognostic modeling [99]. CoReCG [100] and dbCPCO [101] are both useful for CRC studies. The CoReCG database contains 268 gene biomarkers with mutations or expression changes in CRC [100], and dbCPCO focuses on the aggregation of CRC-associated genetic mutation biomarkers, e.g. crucial polymorphisms or somatic/germline mutations [101].

Compared with the databases only specific to one cancer type, CGMD [102] and GDKD [103] are designed for pan-cancer analysis. For example, CGMD integrates tumor genes and biomarkers based on their molecular/pathway characteristics for the precision treatment of different human cancers [102]. GDKD is a gene–drug knowledge base that presents the prevalence of gene variants with the possibility of cancer incidence across various solid tumors. It contains genetic biomarkers for the clinical survey of gene–drug targetability, which is a fundamental step for cancer therapy [103].

Biomarker repertories for other complex diseases

Except for cancers, a number of databases are constructed for biomarker collection of complex diseases in respiratory, urinary or immune system. As listed in Table 2, DAAB is a database manually curated for allergy and asthma biomarkers [104]. About 1200 unique genes/proteins with the diagnostic value are recorded based on the comprehensive review of high-throughput studies related to genomics, proteomics and epigenetics [104]. IDBD is a community annotation database, which contains 611 diagnostic or therapeutic biomarkers with structure or expression changes in infectious diseases, including proteins, nucleic acids, carbohydrates as well as small molecules for 66 infectious diseases and 70 pathogens [105]. As an important biomarker source, urine holds the power for disease pathophysiological study. Shao *et al.* [106] collected human and animal urine protein biomarkers mainly associated with urological and nonurological diseases from published literatures and built the database called OPB for biomarker discovery in the urinary proteome. Moreover, they found that biomarkers identified by different proteomic methods or from different sample sources, e.g. human, rat, mouse and cat, tended to indicate high heterogenicity [106]. Exposome-Explorer is a special database that includes biomarker indices of exposure to dietary and other environmental risk factors [107]. A total of 692 dietary/pollutant biomarkers with concentration values in various human biospecimens are compiled from 480 publications. The database is conductive to the performance comparison of different biomarkers in their fields of application, thereby allowing for the exposome-wide understanding of the etiology of chronic diseases [107]. As the cleavage and release of membrane proteins are functional modulators in many biological processes and disease pathologies, the SheddomeDB database, for the first time, gathers the experimentally validated shed membrane protein biomarkers in certain cellular processes and diseases, and offers a valuable resource for the discovery of membrane-bound shed biomarkers [108].

Models and software tools

Mathematical models

Mathematical methods are useful for molecular biomarker detection. As shown in Figure 2, statistical tests, such as *t*-test, SAM and eBayes, are commonly used to extract dysfunctional molecules from large-scale expression data, which are integrated as an important analytical step in many biomarker identification pipelines. On the other hand, linear programming and nonlinear dynamics theories are often chosen for the systems modeling of disease process. Table 3 contains several well-performed mathematical approaches. For example, Agostini *et al.* [16] integrated the advantages of meta-analysis and DE gene prioritization to identify key genes associated with the response to the preoperative chemoradiotherapy of locally advanced rectal cancer. They first merged the raw gene expression data from three discrete studies using an annotation-based microarray data meta-analysis tool, and then applied the SAM method to screen DE genes that were able to distinguish responders from those nonresponders. Kang *et al.* [109] developed a probabilistic method called CancerLocator. It took advantage of cell-free DNA (cfDNA) methylation profiles to predict cfDNA burden and tissue-of-origin. In the first step, CpG clusters with sufficiently large methylation range were selected from the sample data as features. Then, cancer together with its location were detected and predicted based on the selected features along with their beta distributions through maximum-likelihood estimation [109]. Li *et al.* [110] proposed a miRNA–cancer functional consistency method for cancer miRNA identification. They supposed that if miRNAs were involved in a cancer, their targets should have the similar function as the known cancer-related genes. Based on this assumption, a functional consistency score, called FCS, was defined to measure the functional consistency between miRNAs and certain cancer types [110]. Gao *et al.* [111] performed the genome-wide expression analysis on snoRNAs. Through receiver operating

Table 2. Publicly available biomarker databases and knowledge bases

Title/name	Disease	Description	URL	PMID
Biomarker repertories for cancers				
dbPHCC	Hepatocellular carcinoma	• Database of prognostic biomarkers and models for hepatocellular carcinoma • 567 expression biomarkers for prognostic evaluation, including 323 proteins, 154 genes and 90 miRNAs	http://lifecenter.sgst.cn/dbphcc/	26940364
CoReCG	Colorectal cancer	• Colon Rectal Cancer Gene Database • 268 gene biomarkers with mutations or expression changes in colorectal cancer	http://lms.snu.edu.in/corecg/	27114494
dbCPCO	Colorectal cancer	• Database of genetic biomarkers with prognostic and predictive value in colon rectal cancer • Genetic mutation biomarkers, including 778 findings on 456 polymorphisms, somatic/germline mutations associated with 189 genes	http://www.med.mun.ca/cpco/	20506273
CGMD	Human cancers	• Cancer Gene Marker Database • 309 cancer genes and 206 expression biomarkers	http://cgmd.in/	26160459
GDKD	Solid tumors	• Gene-Drug Knowledge Database • Interactions among 130 genes, 90 drugs and 40 types of malignancies	https://www.synapse.org/#! Synapse: syn2370773/	25656898
Biomarker repertories for other complex diseases				
DAAB	Allergy and asthma	• Database of Allergy and Asthma Biomarkers • 1200 unique gene/protein expression biomarkers with diagnostic value	http://bicresources.jcbose.ac.in/ssaha4/daab/	25973645
IDBD	Infectious diseases	• Infectious Disease Biomarker Database • 611 diagnostic or therapeutic biomarkers with structure or expression changes in infectious diseases, including proteins, nucleic acids and carbohydrates associated with 66 infectious diseases and 70 pathogens	http://biomarker.cdc.go.kr/	17982173
UPB	Urological and nonurological diseases	• Urinary Protein Biomarker database • 553 human proteins with differential expression between disease and control or among different disease stages	http://122.70.220.102/biomarker/index.asp/	21876203
Exposome-Explorer	Chronic diseases	• Biomarkers of exposure to dietary and environmental factors • 692 dietary or pollutant biomarkers as well as 10 510 associated concentration values	http://exposome-explorer.iarc.fr/	27924041
SheddomeDB	Human diseases	• The ectodomain shedding database for membrane-bound shed markers • 401 experimentally validated shed membrane proteins	http://bal.ym.edu.tw/SheddomeDB/	28361715

PMID, PubMed ID; URL, Uniform Resource Locator.

characteristic curve (ROC), univariate and multivariate Cox proportional hazards regression analyses, key snoRNAs with prognostic power in lung cancer were uncovered.

Owing to the limited diagnostic power of single molecules, the identification of combinatorial biomarkers or edge biomarkers is becoming a hot spot for research these years. Mazzara *et al.* [112] focused on the combination of multiple biomarkers to improve disease diagnostic accuracy. They developed a computational tool termed as CombiROC to help determine the optimal biomarker combinations from omics data. A user-defined test stringencyD was used for stringency examination and combinatorial analysis, and then sensitivity and specificity filters were applied for gold combination screening [112]. As interactions among molecules are important to biological systems and most of the existing studies ignored the functional significance of nondifferentially expressed (non-DE) genes in disease development, Zhang *et al.* [43] identified the differentially correlated molecular pairs as edge biomarkers for elucidating the biological mechanism of complex phenomena. In methodology, they first computed the Pearson correlation coefficient (PCC) among genes in disease and control groups from the expression profile and selected the DCPs. After mathematical modeling, the expressions of all genes in DCPs were transformed into the edge space, and discriminative edge features could be then extracted. Finally, edge biomarkers with the optimal prediction power were determined through the employment of sequential forward floating algorithm [43].

Systems biology views vital activities as a unified framework; thus, molecules formed as sets or networks are better for the systematic analysis of disease pathogenesis. Ren *et al.* [113] developed a novel ellipsoid-based method for inferring heterogeneous sets of genes as cancer biomarkers from gene expression data. They assumed that the state of cancer and normal samples was stable in the gene expression space, and the differences between cancer and normal samples or between different cancers were caused by sample heterogeneity. Hence, each

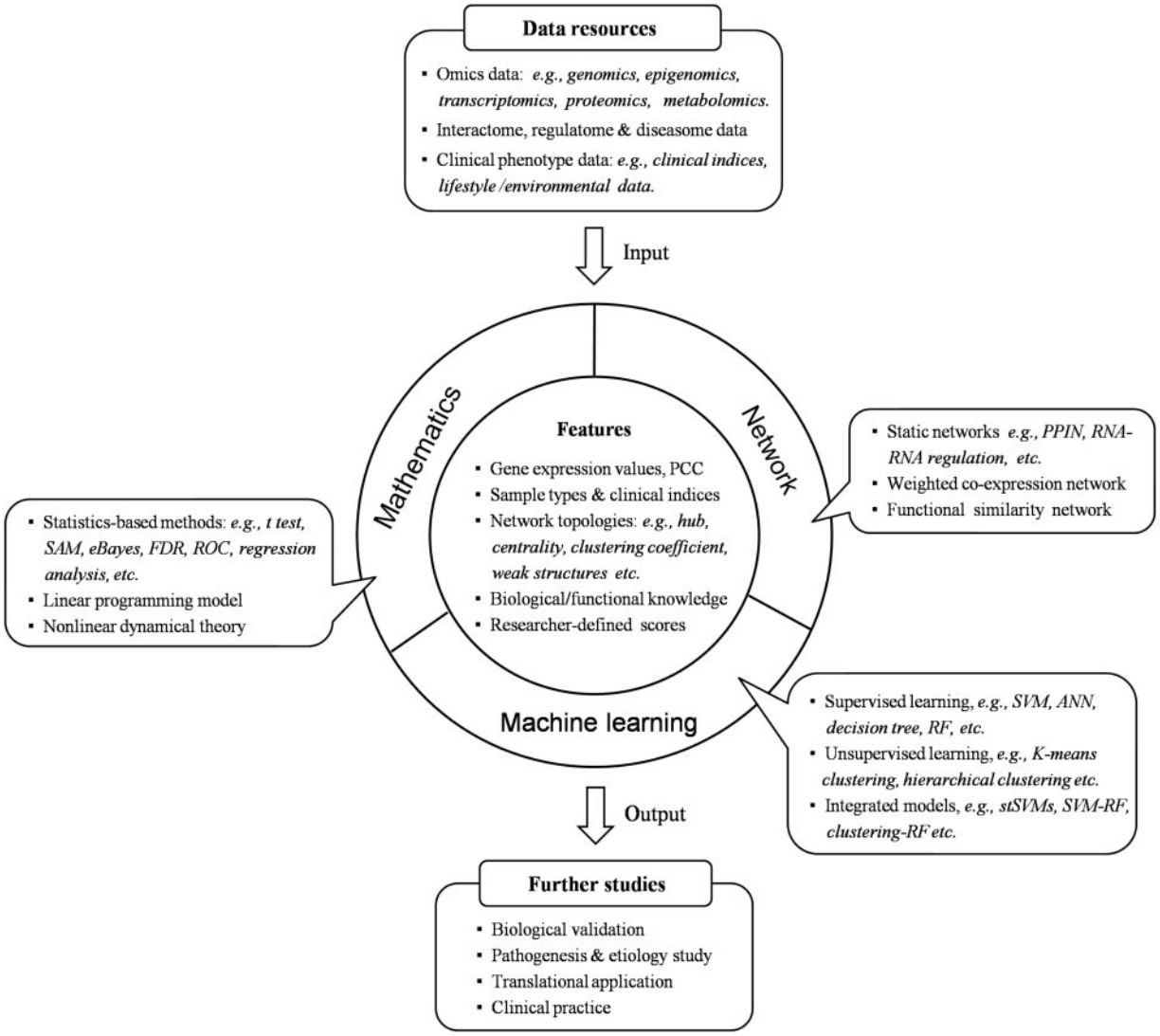

Figure 2. Schematic pipeline of computer-aided biomarker discovery. The bioinformatics models identify functional biomarkers (panels/modules, networks) from various data resources based on feature-guided mathematical, network or machine learning methods. The outputs are chosen for further studies.
FDR, false discovery rate; PPIN, protein–-protein interaction network; ANN, artificial neural network.

cancer type could be represented by an ellipsoid in the gene space, and the key problem was to seek minimal set of genes and maximize the distance between ellipsoids [113]. Min *et al.* [23] integrated the clinical sample information and network knowledge into a network-regularized sparse logistic regression model to predict potential clinical risks as well as biomarker gene sets. In the model, a novel network-regularized term was defined to evaluate the opposite effect of variables with different labels [23]. On the basis of elastic net regularization, Kayano *et al.* [114] proposed a functional logistic model to screen genes with dynamical changes in different conditions. Two steps, i.e. smoothing and estimation, were mathematically designed to rank genes as time-dependent biomarkers from time-course expression data. Similarly, Chen *et al.* [49] also considered the evolutionary trait of disease and supposed that the disease and normal status were more stable than the pre-disease stage, during which the state of the biological system could be sharply changed. DNBs, therefore, tended to be powerful for capturing such changeable signatures in disease progression. Based on the nonlinear dynamical theory, they defined a nonlinear dynamical system equation as $Z(k+1)=f(Z(k); P)$ to describe the state dynamics of a biological organism. In this equation, k and P, respectively, represent the time point and slowly varying genetic or epigenetic factors that can drive the system to change, and the vector $Z(k)=(z_1(k),\ldots, z_n(k))$ denotes the observed data with n variables at the k-th $(k=0, 1, 2,\ldots)$ time point. Moreover, a composite index (CI) was built to quantify the critical transition or pre-disease state as disease early warning signal in biological systems. The index is equivalent to $CI=(SD_d\times PCC_d)/PCC_o$, where SD_d is the average SD of the molecules in DNB, and PCC_d and PCC_o, respectively, are absolute values for the average PCC of the molecules in DNB and the one between DNB and others [49].

Network-based models

Biological networks are important knowledge resources for measuring the interactions among different molecules at the systems level. As described in Figure 2 and Table 3, the PPI network, gene coexpression network and noncoding RNA network are often the

Table 3. Bioinformatics models developed for biomarker discovery

Title/name	Disease	Biomarker type	Principle		Implementation	PMID
			Model and method	Input and feature		
Mathematical models						
Agostini et al.:An integrative approach	• Rectal cancer	• Prognosis • Single genes	• Meta-analysis using A-MADMAN • SAM method	• Gene expression value • Clinical data • Gene functional annotations	• R, localhost • URL: N/A	26359356
Kang et al.:A probabilistic method for tissue-of-origin prediction(CancerLocator)	• Breast cancer • Lung cancer • Liver cancer	• Diagnosis • Single cfDNA with methylation	• Maximum-likelihood estimation	• DNA methylation data • Methylation range • Beta distribution • The methylation level of CpG site in the cfDNA of a cancer patient	• Java, localhost • URL: https://github.com/jasminezhoulab/CancerLocator/	28335812
Li et al.:Cancer miRNA prioritization (CMP/FCS)	• Human cancers	• Diagnosis • Single miRNAs	• MiRNA-cancer functional consistency modeling	• MiRNA/gene functional knowledge • MiRNA–gene associations • Functional consistency score	• Web server • URL: http://bioinfo.hrbmu.edu.cn/CMP/	21976726
Gao et al.:snoRNA biomarker identification	• Lung cancer	• Prognosis • Single snoRNAs	• ROC analysis • Univariate/multivariate Cox proportional hazards regression	• SnoRNA expression data from deep sequencing • DE-snoRNAs • Sensitivity and specificity	• N/A	25159866
Mazzara et al.:Marker Combination Selection (CombiROC)	• Autoimmune hepatitis • Primary central nervous system lymphoma	• Diagnosis • MiRNA or protein combinations	• StringencyD test • ROC analysis	• Multimarkers profiling data • Sensitivity and specificity	• R, Web server • URL: http://CombiROC.eu/	28358118
Zhang et al.:Edge marker identification (EdgeMarker)	• Cholangio carcinoma • Diabetes	• Diagnosis • Gene pairs (edge biomarkers)	• Mathematical transforming • Sequential forward floating selection	• Gene expression data • DE and non-DE genes, PCC, DCPs • Node and edge space features	• N/A	24931676
Ren et al.:Ellipsoid Feature Net (ellipsoidFN)	• Prostate cancer • Breast cancer • Leukemia	• Diagnosis • Gene sets	• Linear programming model	• Gene expression data • Sample information	• Matlab, localhost • URL: http://doc.aporc.org/wiki/EllipsoidFN/	23262226
Min et al.:Network regularized sparse logistic regression model	• Glioblastoma multiforme • NSCLC	• Prognosis • Gene sets	• Network-regularized sparse logistic regression	• Gene expression data • Clinical data • PPI network • A regularized term in logistic regression framework	• R, localhost • URL: http://page.amss.ac.cn/shihua.zhang/	28113328
Kayano et al.:Functional logistic model (F-logistic)	• Multiple sclerosis	• Prognosis • Gene sets	• Elastic net regularization-based functional logistic regression	• Time-course gene expression profiles • Mean function, random-effect term, principal component curves	• N/A	26420796

Continued

Table 3. (continued)

Title/name	Disease	Biomarker type	Principle		Implementation	PMID
			Model and method	Input and feature		
Li et al.:DNBs	• Type 2 diabetes	• Diagnosis • Gene networks	• Nonlinear dynamical system equation • Clustering	• Elastic net regularization parameters • Gene expression data in samples at different time points • PCC, SD of genes	• N/A	23620135
Network-based models						
Wang et al.:Potential gene marker screening	• Carotid atheroma plaque	• Diagnosis • Single genes	• PPI network analysis	• Gene expression data • PPI network • DE genes, hub genes	• R, localhost • URL: N/A	28260035
Liu et al.:A WGCNA-based approach	• Estrogen receptor-positive breast cancer	• Prognosis • Single genes	• Gene coexpression network analysis	• Gene expression data • PCC • Weighted network features, hub genes	• R, localhost • URL: N/A	25756514
Jones et.al:A systems epigenomics approach	• Endometrial cancer	• Diagnosis • Single genes with methylation	• The Functional Epigenetic Modules (FEM) algorithm • Epigenetic Modules (EpiMods) algorithm	• DNA methylation and mRNA expression data • Epigenetically deregulated interactome hotspots • Node and Edge weights in the integrated DNA methylation interactome	• R, localhost • URL: http://code.google. com/p/epimods/downloads/list	24265601
Chen et al.:Random walk with restart for disease miRNA testing (RWRMDA)	• Breast cancer • Colon cancer • Lung cancer	• Diagnosis • Single miRNAs	• MiRNA-miRNA functional similarity network • Random walk with restart	• MiRNA set with seed and candidate miRNAs • MiRNA-miRNA functional similarity score	• Matlab and R, localhost • URL: http://asdcd.amss.ac.cn/Software/Details/3	22875290
Zhou et al.:Dysregulated lncRNA-related ceRNA network	• Pancreatic ductal adenocarcinoma	• Diagnosis • LncRNA panels	• MiRNA-mediated lncRNA-mRNA network analysis	• MiRNA-lncRNA, miRNA-mRNA pairs • Paired miRNA/mRNA/lncRNA expression data • Node degree, hubs • Diagnostic odds ratio	• N/A	27487139
Wang et al.:LncACTs	• Human cancers	• Prognosis • LncRNA–miRNA–gene triplets	• LncRNA-associated ceRNA network analysis	• MiRNA-lncRNA, miRNA-mRNA pairs • TCGA cancer expression data sets • LncACT competing activity score • Betweenness centrality • Biclique modules	• Web server • URL: http://www. biobigdata.net/LncACTdb/	25800746
Cui et al.:LincRNA module biomarker discovery	• Prostate cancer	• Diagnosis • LincRNA modules	• WGCNA method • Gene coexpression network analysis	• RNA-seq lincRNA expression data • PCC	• R, localhost • URL: N/A	26100580

Continued

Table 3. (continued)

Title/name	Disease	Biomarker type	Principle		Implementation	PMID
			Model and method	Input and feature		
Shao et al.:LDCCNet	• Lung adenocarcinoma	• Diagnosis • CeRNA modules	• CeRNA (lncRNA, coding-gene and pseudogene) network analysis • Affinity propagation clustering approach	• Gene expression data • miRNA–target associations • PCC difference of a given ceRNA pair between normal and disease samples • Gain/loss ceRNAs	• R, localhost • URL: N/A	26325208
Yuan et al.:Network biomarker construction	• Leukemia	• Diagnosis • Gene networks	• PPI network analysis • Greedy algorithm	• Gene expression data • PPI network • Leukemia-associated genes • PPI network structure	• R, localhost • URL: N/A	28243332
Wang et al.:Network-based biomarker approach	• Lung cancer	• Diagnosis • Gene/protein networks	• PPI network analysis • Protein association model	• PPI network • Expression level of proteins in samples • Association ability between proteins • Number of proteins interacted with a given protein	• N/A	21211025
Liu et al.:SSN approach	• Breast invasive carcinoma • Glioblastoma multiforme • Kidney renal clear cell carcinoma • Lung squamous cell carcinoma *et al.*	• Diagnosis • Gene networks	• SSN theory	• Gene expression data of n reference samples and the test sample • PCC_{n+1} and PCC_n • Differential correlation coefficient ΔPCC_n • High/low correlation edges	• N/A	27596597
Farahmand et al.:Game theoretic approach (GTA)	• Breast cancer metastasis	• Prognosis • Gene networks	• PPI network analysis • Payoff function-based scoring scheme	• Genome-wide expression profiles • PPI network • Node degree, clustering coefficient • T-test statistic score of log likelihood ratio values	• N/A	26750920
*Chuang et al.*Protein network-based approach	• Breast cancer metastasis	• Prognosis • Gene networks	• PPI network analysis • Greedy algorithm • The random model	• PPI network • Gene expression data • A defined discriminative score MI(a′, c), MI: the mutual information; a': a discretized form of network activity scores; c: class labels	• N/A	17940530
Machine learning models						
Zhao et al.:MiRNA and pathway analysis (miR_Path)	• Lung cancer • Colon cancer	• Diagnosis • Single miRNAs	• Knowledge-based clustering	• Gene expression data • MiRNA–gene network	• R, localhost	25505085

Continued

Table 3. (continued)

Title/name	Disease	Biomarker type	Principle		Implementation	PMID
			Model and method	Input and feature		
	• Breast cancer • Gastric cancer			• Cellular pathway knowledge	• URL: http://comp-sysbio.org/miR_Path/	
Mukhopadhyay et al.:SVM-wrapped multiobjective evolutionary feature selection approach	• Human cancers	• Diagnosis • Single miRNAs	• Genetic algorithm for multi-objective optimization • SVM	• MiRNA expression data set with samples and miRNAs • DE miRNAs	• N/A	24235309
Xu et al.:A network and SVM-based approach	• Prostate cancer	• Diagnosis • Single miRNAs	• MiRNA target-dysregulated network analysis • SVM	• MiRNA and mRNA expression data • MiRNA-mRNA associations • FCs in expression • Network topological features, e.g. the number of dysregulated genes; the number of miRNA coregulators	• N/A	21768329
Yang et al.:MiRNA combinatorial biomarker identification	• Breast cancer	• Diagnosis • MiRNA combinations	• Clustering • Fisher linear discriminant analysis	• MiRNA set filtered from miRNA expression data by *t*-test • Mean squared loss, mean loss rate	• URL: on request	28361698
Li et al.:CircRNA identificantion and characterization	• Hepatocellular carcinoma	• Diagnosis • Single circRNAs and assoicatied genes	• LASSO algorithm • RF	• CircRNAs/genes expression profiles • DE-circRNAs/genes • LASSO-related parameters, coefficient, etc.	• N/A	28346469
Cun et al.:stSVMs	• Breast cancer • Ovarian cancer • Prostate cancer	• Prognosis • Gene sets	• stSVMs	• Gene/miRNA expression data • Biological networks • *T*-statistics of genes, node degrees, routes between pairs of nodes	• R, localhost • URL: http://sourceforge.net/p/netclassr/	24019896
Butti et al.:BioPlat	• Breast cancer	• Prognosis • Gene sets	• *K*-means clustering • Particle swarm optimization	• Gene expression data • A metric of outcome-associated quality	• Java, R, localhost • URL: http://www.cancer-genomics.net/	24574115
Wang et al.:A hybrid approach (HykGene)	• ALL/AML leukemia • MLL leukemia • Colon tumor	• Diagnosis • Gene sets for phenotype classification	• Feature filtering methods, *i.e.* Relief-F, information gain and the χ^2-statistic • Hierarchical clustering	• Gene expression data • 50–100 top-ranked genes selected by feature filtering methods	• URL: http://www.cs.dart-mouth.edu/~wyh/softwarehtml/	15585531
Tremoulet et al.:A data-mining approach (Clustering-RF)	• Kawasaki disease	• Diagnosis • Gene sets	• Clustering • RF • PCA	• Gene expression data • Significant genes identified by statistic methods	• R, localhost • URL: N/A	26237629
Hu et al.:Key marker identification (SVM-RF)	• Neuronal developmental diseases	• Therapy • Gene sets	• SVM-based recursive feature elimination method • RF	• Neural single cell RNA-seq data • Sample information	• R, localhost • URL: N/A	28155657

Continued

Table 3. (continued)

Title/name	Disease	Biomarker type	Principle		Implementation	PMID
			Model and method	Input and feature		
Zou et al.:Detection of multi-biomarker panels (MILP_k)	• Colorectal tumors	• Diagnosis • Gene sets	• Intuitive nearest centroid classifier • SVM	• Gene functional information • DE genes • Gene expression data • Laboratory or clinical features • Distances between the query sample and known malignant or benign samples	• Matlab, localhost • URL: http://doc.aporc.org/wiki/MILP_k/	25980368
Ding et al.:Biomarker network analysis (atBioNet)	• Acute leukemia • Systemic lupus erythematosus • Breast cancer	• Diagnosis and prognosis • Gene modules	• Structural Clustering Algorithm for Networks	• Protein/gene seeds • Network-specific knowledge	• Web server • URL: http://www.fda.gov/ScienceResearch/BioinformaticsTools/ucm285284.html/	22817640
Wen et al.:Multi-class Centroid Feature Selection (MCentridFS)	• Multistages hepatocellular carcinoma • Multitissues breast cancer	• Diagnosis • Gene modules for phenotype classification	• Clustering • Binary integer linear programming model	• High-throughput data with multi-phenotypes • Differential PPI network • Activity matrix of gene modules	• Matlab, localhost • URL: http://www.sysbio.ac.cn/cb/chenlab/images/MCentridFS.rar/	25099602

Note: Models in this table are divided into three subgroups based, respectively, on the principles of mathematics, network and machine learning. In each subgroup, they are sorted by the identified biomarker type from single genes/RNAs to gene/RNA modules and networks.

PMID, PubMed ID;A-MADMAN, the Annotation-based Microarray Data Meta-Analysis tool; FC, fold change; PCA, principal component analysis; AML, acute myeloid leukemia; ALL, acute lymphoid leukemia; MLL, mixed lineage leukemia; URL, Uniform Resource Locator; N/A, not available; CMP: cancer miRNA prioritization; FCS: functional consistency score; CLIP-Seq: crosslinking-immunprecipitation and high-throughput sequencing.

kernels in network-based models, and candidate biomarkers can be screened based on the topological features in the network, e.g. node or edge centrality, vulnerability, the shortest path, etc. For example, Wang *et al.* [115] identified potential gene biomarkers for predicting the formation of carotid atheroma plaque. They constructed the PPI network using genes with DE patterns between high-stage atheroma plaque samples and the controls, and selected hub genes (degree ≥ 20) as seeds for further biomarker signature detection [115]. Liu *et al.* [19] also focused on the hubs in the network. They performed the weighted gene coexpression analyses and found that hub genes in each identified module, e.g. *CDK1*, *DLGAP5*, *MELK*, *NUSAP1* and *RRM2*, were highly associated with the survival of patients with estrogen receptor-positive breast cancer [19]. Jones *et al.* [7] proposed an integrative epigenome–transcriptome–interactome approach to extract epigenetically deregulated interactome hot spots related to endometrial cancer (EC). The bioinformatics strategy first identified EC-associated differential methylation hot spots using epigenetic modules algorithm at DNA methylation level, and then integrated mRNA expression data to infer modules with functional importance. The methylation of hub gene in the top-ranked differential methylation hotspot, i.e. *HAND2*, was further validated as a noninvasive biomarker for EC diagnosis [7]. Chen *et al.* [116] built a miRNA-miRNA functional similarity network and applied an updated random walk strategy to predict latent miRNA–disease associations. In this model, the algorithm starts at a given disease miRNA (seed node) and moves to its neighbors randomly. Compared with traditional methods, they adopted the global network similarity measure for miRNA prioritization, and the top 50 miRNAs were selected for biological validation [116].

Through network analysis, molecular panels or modules with special functions can be easily extracted. Based on 'ceRNA hypothesis', Zhou *et al.* [117] and Wang *et al.* [118] integrated expression data sets with miRNA-lncRNA and miRNA-mRNA regulatory information to construct cancer-associated lncRNA-miRNA-mRNA network for biomarker discovery. In the study by Zhou *et al.* [117], hubs in the network were regarded as key players and a 7-lncRNA signature was screened for pancreatic cancer diagnosis. Wang *et al.* [118] defined the lncRNA-miRNA-mRNA relationship as lncRNA-associated competing triplets (LncACTs). After measuring the activity of LncACT cross talk network, functionally important LncACTs revealing global topological properties were detected as prognostic biomarkers for cancers. Cui *et al.* [46] used the weighted gene coexpression network analysis (WGCNA) algorithm [119] to identify lincRNA coexpression modules for PCa diagnosis. Shao *et al.* [120] built a dysregulated ceRNA–ceRNA network (LDCCNet) for the pathogenesis investigation of lung adenocarcinoma. The network is multivariate, as coding genes, lncRNAs and pseudogenes may competitively combine with miRNAs and form a large-scale interaction network. In this model, the PCC difference of a ceRNA pair between cancer and normal groups was calculated using the criterion $\Delta R = corr_{\text{cancer}}(A, B) - corr_{\text{normal}}(A, B)$, where A and B represented a pair of ceRNAs, and $corr_{\text{cancer}}$ and $corr_{\text{normal}}$ represented the PCC in cancer and normal groups. Hence, the concept of gain and loss ceRNAs was defined as 'significant ΔR, $\Delta R > 0.5$, and significantly positive $corr_{\text{cancer}}(A, B)$' and 'significant ΔR, $\Delta R < -0.5$, and significantly positive $corr_{\text{normal}}(A, B)$', respectively. In terms of network structural features, *e.g.* betweenness, the number of components, etc., both gain and loss ceRNAs, were located at the key sites in LDCCNet, and the dysregulated modules with these ceRNAs were functional for lung adenocarcinoma cancer diagnosis [120].

Yuan *et al.* [18] integrated the gene expression and PPI data to construct the network biomarker for accurate prediction of leukemia. In the study, six individual data sets were downloaded from NCBI GEO database, and human PPI network was chosen as the reference for leukemia-associated PPI network (LPPIN) extraction. Here, the significance of genes in each data set was measured based on its node centrality in LPPIN, and the greedy algorithm was applied to screen top-ranked subnetworks related to each data set. Finally, genes as well as their associations shared by no fewer than two of the six identified subnetworks were incorporated to the biomarker network [18]. Wang *et al.* [121] also used the PPI knowledge for diagnostic protein network identification. The main difference is that they developed protein association models for cancer and noncancer samples, respectively, and selected significant proteins based on the expression level, association ability and structural characteristics of the rough PPI network [121]. Liu *et al.* [122] published a sample-specific network (SSN) method to identify gene networks for disease diagnosis. First, a reference network was constructed based on the PPC of genes in n samples (PCC_n). Then, a new sample d was added and the PPC was calculated again in the $n+1$ samples (PCC_{n+1}) to build the perturbed network. Finally, the difference between the perturbed and reference network could be measured using $\Delta PCC_n = PCC_{n+1} - PCC_n$ for each edge (also known as a gene pair), and the edges with significant ΔPCC_n constituted the sample d-specific SSN for disease characterization. Based on this idea, *TP53* was found to be shared by all of the four individual-specific networks in breast cancer study, which indicated its importance in breast cancer initiation [122]. In respect of prognosis monitoring, Farahmand *et al.* [123] and Chuang *et al.* [51], respectively, proposed the PPI-dependent models for predicting breast cancer metastasis. Among them, Farahmand *et al.* [123] evaluated the biomarker gene network through a payoff function-based scoring scheme, whereas Chuang *et al.* [51] defined a discriminative score to select genes that were able to optimize the classification result.

Machine learning models

Machine learning guides the computer simulating human learning behaviors to require new skills or information from data; meanwhile, it focuses on reorganizing existing knowledge structures so as to improve its prediction performance constantly. As shown in Figure 2, traditional supervised and unsupervised algorithms are now widely used for biomarker feature extraction, and the model integration has also become a research topic in recent years. In the field of miRNA biomarker discovery, as listed in Table 3, Zhao *et al.* [124] proposed a novel approach that inferred cancer miRNAs based solely on gene expression data. In the model, clustering analysis was performed to identify dysfunctional genes and pathways associated with cancer development, and miRNA–gene network was used as a 'bridge' to derive candidate miRNAs from identified gene signatures [124]. Mukhopadhyay *et al.* [125] introduced a SVM-wrapped multiobjective feature selection approach, which integrated Genetic Algorithm and SVM classifier for miRNA biomarker identification. Here, the SVM served as a wrapper for feature evaluation [125]. Xu *et al.* [126] also used the SVM as the training model. They identified disease miRNAs depending on the topological features in the miRNA target-dysregulated network [126]. As combinatorial miRNAs had higher predictive power than single biomarkers, Yang *et al.* [127] designed a clustering-based method to screen the optimal miRNA combinations for breast cancer diagnosis. They removed the

uninformative miRNAs from raw data and applied a hierarchical clustering on the remaining miRNAs. Using a linear discriminate method, representative miRNAs within each cluster were selected and formed biomarker combinations [127]. In addition to miRNA studies, circRNA has been considered as a new star these years. Li *et al.* [128] identified circRNAs in HCC and characterized their roles through functional enrichment, clustering and network analysis. Finally, the RF model was built to differentiate HCC samples from the normal controls based on the profile data of circRNAs and their associated genes.

The machine learning models can also be applied to detect biomarkers in the form of gene sets. For example, Cun *et al.* [24] proposed a novel method that combined the network feature and expression data into an SVM classifier to screen gene signatures for cancer prognosis analyses. The differential expression levels of genes in the network were calculated by *t*-test first, and then the *P*-step random walk kernel was chosen to measure the node similarity and further smooth the raw *t*-statistics for better classification [24]. Among the 10-gene signature identifed from breast cancer data set, *BRCA1* and *TP53* were reported to be critical in breast carcinogenesis [4, 129]. For PCa study, the androgen receptor was found to be functional in metastatic processes through interacting with many other cancer-associated genes [24]. Butti *et al.* [20] developed the software Biomarkers Platform (BioPlat) for biomarker discovery. The workflow includes: first, mining gene signatures from curated repositories such as MSigDB and GeneSigDB, and then importing gene expression profiles to validate the identified signatures using *K*-means clustering method. Finally, optimizing the signatures and selecting compact feature subsets with the best prediction accuracy as biomarkers based on the particle swarm optimization [20]. Wang *et al.* [130] showed a hybrid approach to select gene sets for cancer phenotype recognition. The approach started at the gene ranking procedure using three feature filtering methods, i.e. relief-F, information gain and the χ^2-statistic. Hierarchical clustering was later performed on the 50–100 top-ranked genes. After collapsing dense clusters, the biomarker genes were identified for further verification [130]. Tremoulet *et al.* [21] also used the clustering method to find biomarkers with the potential to distinguish acute Kawasaki disease patients from febrile controls. Moreover, a supervised RF model was integrated to strengthen the biomarker panels [21]. Differed from this work, Hu *et al.* [22] combined the usage of RF with SVM, and constructed the SVM-RF predictive model to screen key genes for classifying developing neocortical cells and neural progenitor cells using single-cell RNA-seq data. In the model, feature genes were extracted through a multistep strategy, i.e. deregulated pathway enrichment, differential expression analysis, SVM-based classification, etc. [22]. Considering the interpretability of clinical data, Zou *et al.* [131] created an intuitive nearest centroid classifier to identify multibiomarker panels based on laboratory/clinical features and the distance between the query sample and known malignant or benign samples. Finally, a small panel containing *CEA*, *IL-10*, *IMA* and *NSE* was obtained to predict the genesis of CRC, and its performance was estimated through an SVM classification.

The construction and application of module biomarkers are also well considered in some computational models. For example, Ding *et al.* [132] built a network-specific knowledge base from public PPI data, and then implemented a fast structural clustering algorithm to mine statistically significant biomarker modules. The model manifested good generalization and was applicable for both diagnostic and prognostic studies [132]. Similarly, Wen *et al.* [133] exploited the PPI data and expression profiles to construct a differential PPI network. The network was then clustered into several modules, and the function of each module was measured using an activity score. The selection of biomarker modules, therefore, could be transferred as a binary integer linear programming problem, where a series of parameters was set to balance the module number and the classification ability [133].

Priori evidence-based model

Many of the existing methods identify biomarkers based solely on the expression difference of molecules at different biological states, i.e. molecules or module members with significantly differential expressions between case and control groups [25]. According to systems biology viewpoints, biomarkers are important indicators to reflect the change between different biological states. The model for biomarker identification should consider the mechanism of the alteration affecting systems stability. Furthermore, with the accumulation of reports on biomarker discovery, the data and information could be collected to conclude and distil general evidences for the characterization of biomarker features.

Zhang *et al.* [134] found the distribution of nodes, i.e. miRNAs and genes/mRNAs, in human miRNA-mRNA network tended to follow the power-law distribution, which implied that miRNAs with more targets were fewer in number. Meanwhile, there were still a proportion of genes uniquely regulated by specific miRNAs. To decode the topological characteristics of miRNA biomarkers in the network, they performed statistical analyses on the reported biomarker miRNAs. The result showed that miRNAs as biomarkers were more likely to regulate genes independently [134]. Based on this observation, Yan *et al.* [135] refined the biological function of miRNA targets and indicated the relationship between biomarker miRNAs and TF genes. Shen *et al.* [136] further extracted the disease-associated genes as the prior knowledge to measure the specificity between miRNAs and the studied disease. Three features, therefore, were defined to measure the propensity of a miRNA as a candidate biomarker for a given biological state change:

Feature 1: Number of single-line regulation (NSR), which represents the number of genes independently regulated by a single miRNA

Feature 2: TF gene percentage (TFP), which represents the percentage of TF genes regulated by a single miRNA

Feature 3: Disease-associated gene percentage (DGP), which represents the percentage of disease-associated genes regulated by a single miRNA

Based on above definitions and the analysis of human miRNA-mRNA network, three evidences were uncovered for miRNA biomarker discovery:

Evidence 1: miRNA biomarkers have high NSR values [134].

Compared with the synergistic regulatory mechanism, single-line regulation sites are more fragile in the miRNA-mRNA network, as the breakdown of such structures may have no substitution and compensation, this can directly affect the function of downstream genes and finally result in the state change of a biological system.

Evidence 2: miRNA biomarkers have high TFP values [135].

TFs are a class of important regulators in living organisms. A wide variety of biological activities, e.g. cell proliferation, cycle, apoptosis, etc., are influenced by TF regulation [137]. If the

expression of TFs is altered, the behaviors of their targeted genes as well as the stability of the gene network will be affected; thus, miRNAs with high TFPs are thought to be important to the pathogenesis of complex diseases.

Evidence 3: miRNA biomarkers have high DGP values [136].

If the targets of a given miRNA are strongly associated with a disease, intuitively, the miRNA would be involved more in the evolutionary process of this disease.

Based on above evidences and network stability, a computational model called MicroRNA Biomarker Discovery (MiRNA-BD) was built. As illustrated in Figure 3, the model integrates miRNA/gene expression data and miRNA-mRNA associations into a reconstructed network, i.e. a condition-specific or disease-specific miRNA-mRNA network. The structural and functional importance of each miRNA in the network can be analyzed via the priori evidences. Finally, miRNAs with significant evidence-based features (default threshold: P-value <0.05, Wilcoxon signed-rank test) are selected as candidate biomarkers for further experimental or bioinformatics validation. Compared with the other methods that are dependent on the training data [125, 126], this model uses statistical evidences as principles for miRNA biomarker discovery, therefore presenting good performance for applicability and generality. Translational applications to cancers [134, 135, 138–141], sepsis [142], acute coronary syndrome [144] and autism spectrum disorder [136] demonstrated its predictive power.

Discussion and perspective

Computer-aided biomarker discovery has now become an emerging paradigm to identify key molecules within biological systems. Over the past years, innovations in both theories and techniques have contributed to the methodological advancement in bioinformatics and other relevant subjects. Unprecedented data resources are continuously shifting the thinking mode from reductionism-based hypotheses to holism-guided views, and constituting abundant material basis for the systematical description of living organisms. Novel computational models, in the meantime, overcome the traditional difficulties in disease etiology simulation, which arise the opportunities for systems-level molecular biomarker identification.

In the biomedical informatics era, data fusion and knowledge merging have been gradually recognized as the major characterization to stand for current bioinformatics wisdoms in biomarker discovery. The accumulated data resources, undoubtedly, drive the progress of intelligent computing and data mining, which benefit the systems learning of biological features for disease state prediction. For example, hub genes with significantly deregulated patterns are always trained as the symbol to characterize the process of disease development [19, 115, 121]. In such cases, the transcriptome data and gene interaction network are essential for the feature extraction, which make great contributions to the knowledge formation. In addition, massive studies focused on the comparison between different computational methods, and unified the advantages of the existing models into a comprehensive framework for the accurate detection of disease signatures. Instead of single molecules, the identified biomarkers have started to form as modules or networks, catering to the holistic understanding of disease pathogenesis. Translational applications to complex diseases, e.g. cancers, diabetes, cardiovascular and neurodevelopmental diseases, uncovered crucial gene and pathway signals related to pathogenesis, e.g. *TP53* [24, 122], *MYC* [18, 51], cell cycle [123, 132] and *p53* signaling [124, 127], which gained solid biological evidences on molecular pathogenicity and provided multiangle decision supports for systems therapeutics in the clinical context.

Although the achievements seem to be encouraging, some shortages are still worth being concerned about. On one hand, the 'big data' term brings valuable treasures for bioinformatics analysis, but the high-dimensional property may severely limit the realizability and extendibility of calculation. On the other hand, the validity of the data is confusing all along, and dirty ingredients greatly detract from the reliability and accuracy of experimental findings. Thus, quality inspections, e.g. data cleaning, noise filtering, nomenclature normalizing, etc., are necessary to be done before data processing. Considering the personal details of patients or donators, e.g. genetic background, medical history and the identity authentication, privacy protection remains to be improved. With regard to the models, the vast majority of published approaches identifying biomarkers mainly concentrated on the expression levels of molecules, which are not always the true essence for disease measurement. Owing to the dynamical nature of disease evolution, alternations in biological states or conditions ought to be the first considerable principle for biomarker characteristic description. The 'training-and-testing' procedure used in most of the machine learning models often gets overfitting results, and few of the methods deliver deeper insights into the hidden structures of the data. It should be noticed that the 'use of data' is differed from the 'analysis of data', while the latter not only advocate the transition from data to information but also highlight the refining of information to various biological knowledge. Bioinformatics studies cannot be performed in isolation from biological rules; therefore, functional annotations associated with specific biological processes should be particularly considered in model construction as well as result demonstration. Moreover, some of the methods obtained hundreds of molecules as biomarkers; however, it might be hard to be applied into the real clinical use because of the bottlenecks in advanced data analysis and the cost.

One of the key issues that need to be dealt with is the clinical use of identified biomarkers toward personalized medicine. Currently, plenty of efforts put particular emphasis on genetic, lifestyle and environmental differences between population groups, and personalized strategies for biomarker analyses are sequentially developed. For example, the records in Exposome-Explorer may help monitor the personal dietary habits related to chronic diseases [107], whereas the SSN model is a new powerful tool to identify individual-specific disease biomarkers for personalized characterization [122]. Biomarkers are the foundation of precision diagnosis and personalized drug design. Based on their profiles, specific disease signals can be captured at the early time, and this will be convenient for the development of proper therapeutic regimens. Moreover, the response of certain individuals or populations to a special kind of medicine can be effectively measured to evaluate whether the patients are favorable for the treatment. To better serve the personalized medicine, biomarkers should be able to indicate the heterogeneity caused by the personalized characteristics of diseases; therefore, large-sample-based statistical analysis and informatics models are preferred for the detection of population-specific signatures. During the identification of biomarker features, it is necessary to choose rational methods that take account of the sample/disease heterogeneity problems [144]. In the near future, personalization-based database with paired genotype–

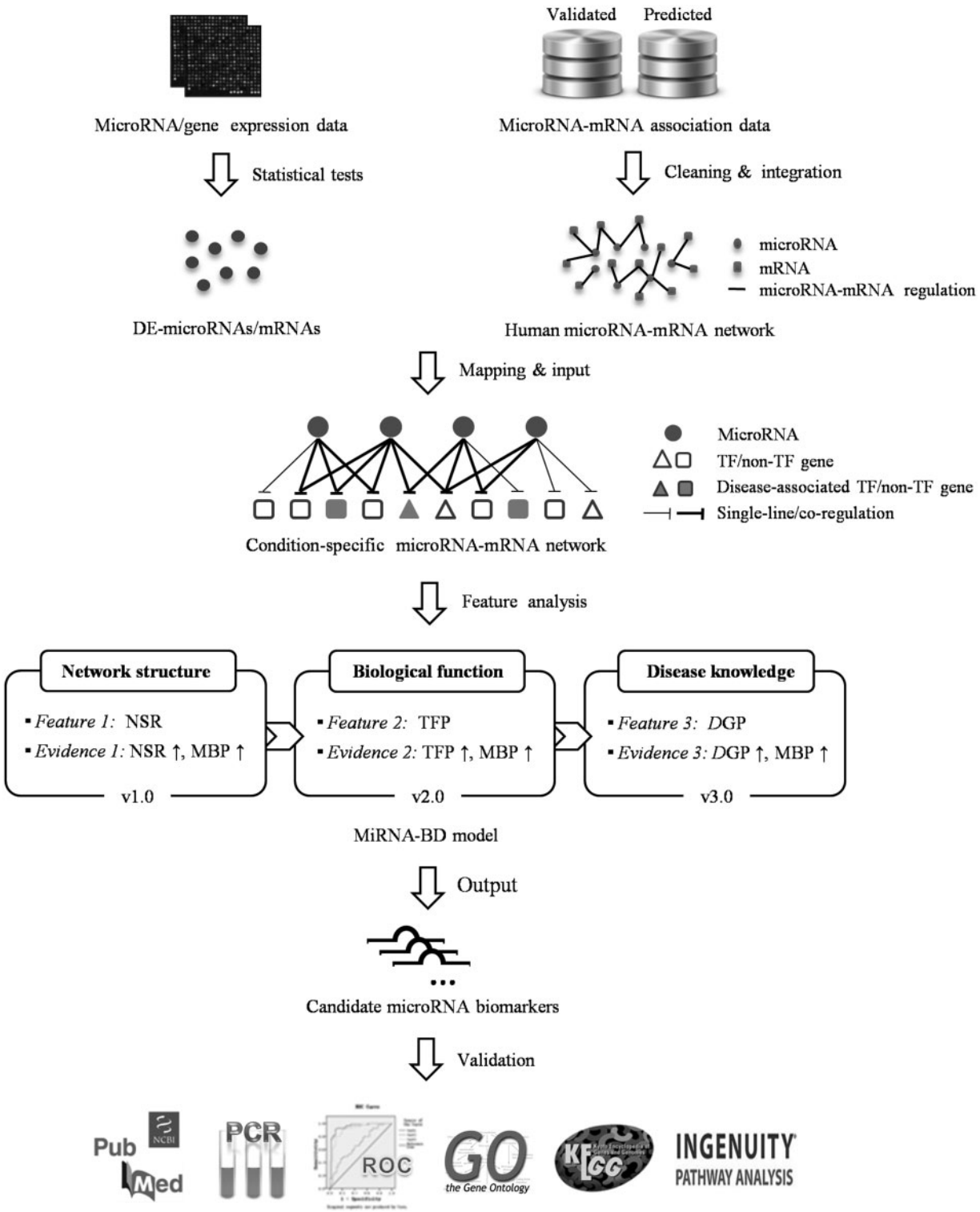

Figure 3. Schematic workflow of the evidence-based MiRNA-BD model. The model constructs condition-specific miRNA-mRNA network from human miRNA-mRNA network (namely, the reference network) and preidentified DE-miRNA/mRNAs. Three feature parameters, i.e. NSR, TFP and DGP, are defined sequentially to measure the possibility of miRNAs as biomarkers based on statistical evidences from network structure, biological function and disease knowledge, respectively. MiRNAs with higher NSR, TFP and DGP values are screened as candidate biomarkers for further validations.

MiRNA-BD: MicroRNA Biomarker Discovery; DGP (here disease is a collective name, which should be specified according to certain case studies, e.g. AGP, autism-associated gene percentage); MBP, microRNA biomarker possibility; ↑, higher; PCR, polymerase chain reaction.

phenotype information of individuals is an attractive direction toward precision medicine. Computational models should not be solely rooted in the single omics-level interpretation of complex biological phenomena. The combination of multiscale molecular knowledge (e.g. genes, RNAs, proteins, etc.) and clinical indices (e.g. images, symptoms, etc.) offers more precise support for medical decision making. It is understandable that the stress from external factors, such as lifestyle and living environments, can also mediate the signal transduction between genotype and phenotype, leading to the heterogeneities on disease incidence and mortality among individuals or populations. Hence, expanding molecular biomarkers into the 'cross-level' form is another considerable way to accelerate the translational application for personalized medicine. In this framework, effects from nongenetic components on molecular functions are thoroughly weighted to quantify the personalized trails in given situations. As the occurrence and progression of diseases are treated more as an evolutionary process, during which a series of physiological activities, e.g. inflammatory reaction, immune regulation, may create the microenvironment for disease growth, to help the early detection of diseases, it would be better to capture the driven signals related to disease initiation and explore possible causalities for the underlying etiology investigation. Such integration and refinement will generate more sensitive and more robust signatures for disease management, thereby opening a systems medicine avenue to promote the flourishing of predictive, preventive, personalized and participatory health spectrum [145].

Key Points

- Biomarkers are sensitive and specific indicators with the ability to predict disease occurrence and progression. Compared with routine disease-associated factors, biomarkers are able to capture the changeable signatures within biological systems.
- Owing to the high efficiency in decoding disease pathogenesis under a holistic framework, computer-aided molecular biomarker discovery has become a burgeoning paradigm nowadays.
- Based on systems biology, in this article, biomarkers are classified as molecular biomarkers, clinical phenotype biomarkers and cross-level biomarkers. Moreover, molecular biomarkers are often formed as single molecules or molecular modules/networks.
- Principles in mathematical, network and machine learning techniques guide the construction of biomarker identification models. The feature values are mined from databases with expression data, network information and biological knowledge. Here, MiRNA-BD is a well-performed model, which integrates network substructural and functional traits for miRNA biomarker discovery.
- Advantages and challenges are coexisting in current bioinformatics studies. The combination of molecular signatures and nongenetic contributors, e.g. lifestyles, environmental factors, etc., will be a future direction toward precision medicine.

Funding

The National Key Research and Development Program of China (grant number 2016YFC1306605), National Natural Science Foundation of China (grant numbers 31670851, 31470821, 91530320, 31400712 and 61602332) and Natural Science Fund for Colleges and Universities in Jiangsu Province (grant number 14KJB520035).

References

1. Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin Pharmacol Ther* 2001;**69**:89–95.
2. Strimbu K, Tavel JA. What are biomarkers? *Curr Opin HIV AIDS* 2010;**5**(6):463–6.
3. Chen J, Sun M, Shen B. Deciphering oncogenic drivers: from single genes to integrated pathways. *Brief Bioinform* 2015; **16**(3):413–28.
4. Ford D, Easton DF, Stratton M, *et al*. Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. The Breast Cancer Linkage Consortium. *Am J Hum Genet* 1998;**62**(3):676–89.
5. Lehmann-Werman R, Neiman D, Zemmour H, *et al*. Identification of tissue-specific cell death using methylation patterns of circulating DNA. *Proc Natl Acad Sci USA* 2016;**113**: E1826–34.
6. Rezvani N, Alibakhshi R, Vaisi-Raygani A, *et al*. Detection of SPG20 gene promoter-methylated DNA, as a novel epigenetic biomarker, in plasma for colorectal cancer diagnosis using the MethyLight method. *Oncol Lett* 2017;**13**:3277–84.
7. Jones A, Teschendorff AE, Li Q, *et al*. Role of DNA methylation and epigenetic silencing of HAND2 in endometrial cancer development. *PLoS Med* 2013;**10**(11):e1001551.
8. Qu K, Zhang X, Lin T, *et al*. Circulating miRNA-21-5p as a diagnostic biomarker for pancreatic cancer: evidence from comprehensive miRNA expression profiling analysis and clinical validation. *Sci Rep* 2017;**7**(1):1692.
9. Ng L, Wan TM, Man JH, *et al*. Identification of serum miR-139-3p as a non-invasive biomarker for colorectal cancer. *Oncotarget* 2017;**8**:27393–400.
10. He B, Zeng J, Chao W, *et al*. Serum long non-coding RNAs MALAT1, AFAP1-AS1 and AL359062 as diagnostic and prognostic biomarkers for nasopharyngeal carcinoma. *Oncotarget* 2017;**8**(25):41166–77.
11. Nogueira Jorge NA, Wajnberg G, Ferreira CG, *et al*. snoRNA and piRNA expression levels modified by tobacco use in women with lung adenocarcinoma. *PLoS One* 2017;**12**(8):e0183410.
12. Assumpcao CB, Calcagno DQ, Araujo TM, *et al*. The role of piRNA and its potential clinical implications in cancer. *Epigenomics* 2015;**7**:975–84.
13. Meng S, Zhou H, Feng Z, *et al*. CircRNA: functions and properties of a novel potential biomarker for cancer. *Mol Cancer* 2017;**16**(1):94.
14. Meng X, Li X, Zhang P, *et al*. Circular RNA: an emerging key player in RNA world. *Brief Bioinform* 2017;**18**(4):547–57.
15. Yao JT, Zhao SH, Liu QP, *et al*. Over-expression of CircRNA_100876 in non-small cell lung cancer and its prognostic value. *Pathol Res Pract* 2017;**213**(5):453–6.
16. Agostini M, Janssen KP, Kim IJ, *et al*. An integrative approach for the identification of prognostic and predictive biomarkers in rectal cancer. *Oncotarget* 2015;**6**(32):32561–74.
17. Datta S, Datta S. Empirical Bayes screening of many p-values with applications to microarray studies. *Bioinformatics* 2005; **21**(9):1987–94.
18. Yuan X, Chen J, Lin Y, *et al*. Network biomarkers constructed from gene expression and protein-protein interaction data for accurate prediction of Leukemia. *J Cancer* 2017;**8**(2): 278–86.

19. Liu R, Guo CX, Zhou HH. Network-based approach to identify prognostic biomarkers for estrogen receptor-positive breast cancer treatment with tamoxifen. *Cancer Biol Ther* 2015;**16**(2):317–24.
20. Butti MD, Chanfreau H, Martinez D, *et al*. BioPlat: a software for human cancer biomarker discovery. *Bioinformatics* 2014; **30**(12):1782–4.
21. Tremoulet AH, Dutkowski J, Sato Y, *et al*. Novel data-mining approach identifies biomarkers for diagnosis of Kawasaki disease. *Pediatr Res* 2015;**78**(5):547–53.
22. Hu Y, Hase T, Li HP, *et al*. A machine learning approach for the identification of key markers involved in brain development from single-cell transcriptomic data. *BMC Genomics* 2016;**17**(S13):1025.
23. Min W, Liu J, Zhang S. Network-regularized sparse logistic regression models for clinical risk prediction and biomarker discovery. *IEEE/ACM Trans Comput Biol Bioinform* 2016, doi: 10.1109/TCBB.2016.2640303.
24. Cun Y, Frohlich H. Network and data integration for biomarker signature discovery via network smoothed T-statistics. *PLoS One* 2013;**8**(9):e73074.
25. Zeng T, Zhang W, Yu X, *et al*. Big-data-based edge biomarkers: study on dynamical drug sensitivity and resistance in individuals. *Brief Bioinform* 2016;**17**(4):576–92.
26. Balk SP, Ko YJ, Bubley GJ. Biology of prostate-specific antigen. *J Clin Oncol* 2003;**21**(2):383–91.
27. Salman JW, Schoots IG, Carlsson SV, *et al*. Prostate specific antigen as a tumor marker in prostate cancer: biochemical and clinical aspects. *Adv Exp Med Biol* 2015;**867**:93–114.
28. Senior JR. Alanine aminotransferase: a clinical and regulatory tool for detecting liver injury-past, present, and future. *Clin Pharmacol Ther* 2012;**92**(3):332–9.
29. Gupta RA, Shah N, Wang KC, *et al*. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* 2010;**464**(7291):1071–6.
30. Teschendorff AE, Lee SH, Jones A, *et al*. HOTAIR and its surrogate DNA methylation signature indicate carboplatin resistance in ovarian cancer. *Genome Med* 2015;**7**(1):108.
31. Wang W, He X, Zheng Z, *et al*. Serum HOTAIR as a novel diagnostic biomarker for esophageal squamous cell carcinoma. *Mol Cancer* 2017;**16**(1):75.
32. Dalerba P, Sahoo D, Paik S, *et al*. CDX2 as a prognostic biomarker in stage II and stage III colon cancer. *N Engl J Med* 2016;**374**(3):211–22.
33. Qin M, Liu G, Huo X, *et al*. Hsa_circ_0001649: a circular RNA and potential novel biomarker for hepatocellular carcinoma. *Cancer Biomark* 2016;**16**(1):161–9.
34. Chen L, Wu J. Systems biology for complex diseases. *J Mol Cell Biol* 2012;**4**(3):125–6.
35. Sudlow C, Gallacher J, Allen N, *et al*. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 2015; **12**(3):e1001779.
36. Bandyopadhyay S, Mehta M, Kuo D, *et al*. Rewiring of genetic networks in response to DNA damage. *Science* 2010; **330**(6009):1385–9.
37. Creixell P, Schoof EM, Erler JT, *et al*. Navigating cancer network attractors for tumor-specific therapy. *Nat Biotechnol* 2012;**30**(9):842–8.
38. Creixell P, Schoof EM, Simpson CD, *et al*. Kinome-wide decoding of network-attacking mutations rewiring cancer signaling. *Cell* 2015;**163**(1):202–17.
39. Lee MJ, Ye AS, Gardino AK, *et al*. Sequential application of anticancer drugs enhances cell death by rewiring apoptotic signaling networks. *Cell* 2012;**149**(4):780–94.
40. Teschendorff AE, Severini S. Increased entropy of signal transduction in the cancer metastasis phenotype. *BMC Syst Biol* 2010;**4**:104.
41. Sahni N, Yi S, Zhong Q, *et al*. Edgotype: a fundamental link between genotype and phenotype. *Curr Opin Genet Dev* 2013; **23**(6):649–57.
42. Yu X, Li G, Chen L. Prediction and early diagnosis of complex diseases by edge-network. *Bioinformatics* 2014;**30**(6):852–9.
43. Zhang W, Zeng T, Chen L. EdgeMarker: identifying differentially correlated molecule pairs as edge-biomarkers. *J Theor Biol* 2014;**362**:35–43.
44. Iida N, Dzutsev A, Stewart CA, *et al*. Commensal bacteria control cancer response to therapy by modulating the tumor microenvironment. *Science* 2013;**342**(6161):967–70.
45. Zhang HM, Kuang S, Xiong X, *et al*. Transcription factor and microRNA co-regulatory loops: important regulatory motifs in biological processes and diseases. *Brief Bioinform* 2015; **16**(1):45–58.
46. Cui W, Qian Y, Zhou X, *et al*. Discovery and characterization of long intergenic non-coding RNAs (lincRNA) module biomarkers in prostate cancer: an integrative analysis of RNA-seq data. *BMC Genomics* 2015;**16**(Suppl 7):S3.
47. Liu Y, Zhang J, Li L, *et al*. Genomic heterogeneity of multiple synchronous lung cancer. *Nat Commun* 2016;**7**:13200.
48. Lin Y, Yuan X, Shen B. Network-based biomedical data analysis. *Adv Exp Med Biol* 2016;**939**:309–32.
49. Li M, Zeng T, Liu R, *et al*. Detecting tissue-specific early warning signals for complex diseases based on dynamical network biomarkers: study of type 2 diabetes by cross-tissue analysis. *Brief Bioinform* 2014;**15**(2):229–43.
50. Liu R, Wang X, Aihara K. Early diagnosis of complex diseases by molecular biomarkers, network biomarkers, and dynamical network biomarkers. *Med Res Rev* 2014;**34**(3):455–78.
51. Chuang HY, Lee E, Liu YT, *et al*. Network-based classification of breast cancer metastasis. *Mol Syst Biol* 2007;**3**:140.
52. Liu X, Liu R, Zhao XM, *et al*. Detecting early-warning signals of type 1 diabetes and its leading biomolecular networks by dynamical network biomarkers. *BMC Med Genomics* 2013; **6**(Suppl 2):S8.
53. Galban CJ, Han MK, Boes JL, *et al*. Computed tomography-based biomarker provides unique signature for diagnosis of COPD phenotypes and disease progression. *Nat Med* 2012;**18**: 1711–15.
54. Pasikhova Y, Ludlow S, Baluch A. Fever in patients with cancer. *Cancer Control* 2017;**24**(2):193–7.
55. Feng F, Tian Y, Yang X, *et al*. Postoperative fever predicts poor prognosis of gastric cancer. *Oncotarget* 2017;**8**(37): 62622–9.
56. Astrup GL, Rustoen T, Miaskowski C, *et al*. Changes in and predictors of pain characteristics in patients with head and neck cancer undergoing radiotherapy. *Pain* 2015;**156**(5): 967–79.
57. Lin Y, Chen J, Shen B. Interactions between genetics, lifestyle, and environmental factors for healthcare. *Adv Exp Med Biol* 2017;**1005**:167–91.
58. Petronis A. Epigenetics as a unifying principle in the aetiology of complex traits and diseases. *Nature* 2010;**465**(7299): 721–7.
59. Wood CL. Environment. Environmental change and the ecology of infectious disease. *Science* 2014;**346**(6214):1192.
60. Szklarczyk D, Morris JH, Cook H, *et al*. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res* 2017; **45**(D1):D362–8.

61. Stark C, Breitkreutz BJ, Chatr-Aryamontri A, *et al*. The BioGRID interaction database: 2011 update. *Nucleic Acids Res* 2011;**39**:D698–704.
62. Keshava Prasad TS, Goel R, Kandasamy K, *et al*. Human protein reference database–2009 update. *Nucleic Acids Res* 2009; **37**:D767–72.
63. Cowley MJ, Pinese M, Kassahn KS, *et al*. PINA v2.0: mining interactome modules. *Nucleic Acids Res* 2012;**40**(D1):D862–5.
64. Basha O, Barshir R, Sharon M, *et al*. The TissueNet v.2 database: a quantitative view of protein-protein interactions across human tissues. *Nucleic Acids Res* 2017;**45**(D1):D427–31.
65. Tryka KA, Hao L, Sturcke A, *et al*. NCBI's database of genotypes and phenotypes: dbGaP. *Nucleic Acids Res* 2014;**42**(D1): D975–9.
66. Kanehisa M, Furumichi M, Tanabe M, *et al*. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* 2017;**45**(D1):D353–61.
67. Kramer A, Green J, Pollard J, Jr, *et al*. Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics* 2014;**30**:523–30.
68. Chou CH, Chang NW, Shrestha S, *et al*. miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. *Nucleic Acids Res* 2016;**44**:D239–47.
69. Vlachos IS, Paraskevopoulou MD, Karagkouni D, *et al*. DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA: mRNA interactions. *Nucleic Acids Res* 2015;**43**:D153–9.
70. Li JH, Liu S, Zhou H, *et al*. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res* 2014; **42**(D1):D92–7.
71. Wang J, Lu M, Qiu C, *et al*. TransmiR: a transcription factor-microRNA regulation database. *Nucleic Acids Res* 2010; **38**(Suppl 1):D119–22.
72. Paraskevopoulou MD, Vlachos IS, Karagkouni D, *et al*. DIANA-LncBase v2: indexing microRNA targets on non-coding transcripts. *Nucleic Acids Res* 2016;**44**(D1):D231–8.
73. Wang P, Zhi H, Zhang Y, *et al*. miRSponge: a manually curated database for experimentally supported miRNA sponges and ceRNAs. *Database* 2015;**2015**:bav098.
74. Liu YC, Li JR, Sun CH, *et al*. CircNet: a database of circular RNAs derived from transcriptome sequencing data. *Nucleic Acids Res* 2016;**44**(D1):D209–15.
75. Rosenbloom KR, Dreszer TR, Long JC, *et al*. ENCODE whole-genome data in the UCSC genome browser: update 2012. *Nucleic Acids Res* 2012;**40**:D912–17.
76. Medvedeva YA, Lennartsson A, Ehsani R, *et al*. EpiFactors: a comprehensive database of human epigenetic factors and complexes. *Database* 2015;**2015**:bav067.
77. Qin G, Zhao XM. A survey on computational approaches to identifying disease biomarkers based on molecular networks. *J Theor Biol* 2014;**362**:9–16.
78. Edgar R, Domrachev M, Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 2002;**30**(1):207–10.
79. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol* 2015;**19**(1A):A68–77.
80. Jiang Q, Wang Y, Hao Y, *et al*. miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res* 2009;**37**:D98–104.
81. Li Y, Qiu C, Tu J, *et al*. HMDD v2.0: a database for experimentally supported human microRNA and disease associations. *Nucleic Acids Res* 2014;**42**(D1):D1070–4.
82. Meng X, Wang J, Yuan C, *et al*. CancerNet: a database for decoding multilevel molecular interactions across diverse cancer types. *Oncogenesis* 2015;**4**:e177.
83. Chen G, Wang Z, Wang D, *et al*. LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res* 2013;**41**:D983–6.
84. Ghosal S, Das S, Sen R, *et al*. Circ2Traits: a comprehensive database for circular RNA potentially associated with disease and traits. *Front Genet* 2013;**4**:283.
85. Hamosh A, Scott AF, Amberger JS, *et al*. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 2005;**33**:D514–17.
86. Griffith M, Spies NC, Krysiak K, *et al*. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nat Genet* 2017;**49**(2):170–4.
87. Singh Nanda J, Kumar R, Raghava GP. dbEM: a database of epigenetic modifiers curated from cancerous and normal genomes. *Sci Rep* 2016;**6**(1):19340.
88. Qi Y, Wang D, Wang D, *et al*. HEDD: the human epigenetic drug database. *Database* 2016;**2016**:baw159.
89. Kibbe WA, Arze C, Felix V, *et al*. Disease ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Res* 2015;**43**:D1071–8.
90. Hankey BF, Ries LA, Edwards BK. The surveillance, epidemiology, and end results program: a national resource. *Cancer Epidemiol Biomarkers Prev* 1999;**8**:1117–21.
91. Clark K, Vendt B, Smith K, *et al*. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J Digit Imaging* 2013;**26**(6):1045–57.
92. Egevad L, Cheville J, Evans AJ, *et al*. Pathology imagebase-a reference image database for standardization of pathology. *Histopathology* 2017;**71**(5):677–85.
93. Hupe P, La Rosa P, Liva S, *et al*. ACTuDB, a new database for the integrated analysis of array-CGH and clinical data for tumors. *Oncogene* 2007;**26**(46):6641–52.
94. Trimarchi M, Lund VJ, Nicolai P, *et al*. Database for the collection and analysis of clinical data and images of neoplasms of the sinonasal tract. *Ann Otol Rhinol Laryngol* 2004;**113**(4):335–7.
95. Narang P, Dhapola P, Chowdhury S. BreCAN-DB: a repository cum browser of personalized DNA breakpoint profiles of cancer genomes. *Nucleic Acids Res* 2016;**44**(D1):D952–8.
96. Madhavan S, Gusev Y, Harris M, *et al*. G-DOC: a systems medicine platform for personalized oncology. *Neoplasia* 2011;**13**(9):771–83.
97. Yu Q, Huang JF. The DEER database: a bridge connecting drugs, environmental effects, and regulations. *Gene* 2013; **520**(2):98–105.
98. Yang Q, Qiu C, Yang J, *et al*. miREnvironment database: providing a bridge for microRNAs, environmental factors and phenotypes. *Bioinformatics* 2011;**27**(23):3329–30.
99. Ouyang J, Sun Y, Li W, *et al*. dbPHCC: a database of prognostic biomarkers for hepatocellular carcinoma that provides online prognostic modeling. *Biochim Biophys Acta* 2016; **1860**(11 Pt B):2688–95.
100. Agarwal R, Kumar B, Jayadev M, *et al*. CoReCG: a comprehensive database of genes associated with colon-rectal cancer. *Database* 2016;**2016**:baw059.
101. Savas S, Younghusband HB. dbCPCO: a database of genetic markers tested for their predictive and prognostic value in colorectal cancer. *Hum Mutat* 2010;**31**(8):901–7.
102. Pradeepkiran JA, Sainath SB, Kumar KK, *et al*. CGMD: an integrated database of cancer genes and markers. *Sci Rep* 2015;**5**: 12035.

103. Dienstmann R, Jang IS, Bot B, *et al*. Database of genomic biomarkers for cancer drugs and clinical targetability in solid tumors. *Cancer Discov* 2015;**5**(2):118–23.
104. Sircar G, Saha B, Jana T, *et al*. DAAB: a manually curated database of allergy and asthma biomarkers. *Clin Exp Allergy* 2015;**45**(7):1259–61.
105. Yang IS, Ryu C, Cho KJ, *et al*. IDBD: infectious disease biomarker database. *Nucleic Acids Res* 2008;**36**:D455–60.
106. 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* 2011; **10**(11):M111.010975.
107. Neveu V, Moussy A, Rouaix H, *et al*. Exposome-explorer: a manually-curated database on biomarkers of exposure to dietary and environmental factors. *Nucleic Acids Res* 2017; **45**(D1):D979–84.
108. Tien WS, Chen JH, Wu KP. SheddomeDB: the ectodomain shedding database for membrane-bound shed markers. *BMC Bioinformatics* 2017;**18**(Suppl 3):42.
109. Kang S, Li Q, Chen Q, *et al*. CancerLocator: non-invasive cancer diagnosis and tissue-of-origin prediction using methylation profiles of cell-free DNA. *Genome Biol* 2017;**18**:53.
110. Li X, Wang Q, Zheng Y, *et al*. Prioritizing human cancer microRNAs based on genes' functional consistency between microRNA and cancer. *Nucleic Acids Res* 2011; **39**(22):e153.
111. Gao L, Ma J, Mannoor K, *et al*. Genome-wide small nucleolar RNA expression analysis of lung cancer by next-generation deep sequencing. *Int J Cancer* 2015;**136**(6):E623–9.
112. Mazzara S, Rossi RL, Grifantini R, *et al*. CombiROC: an interactive web tool for selecting accurate marker combinations of omics data. *Sci Rep* 2017;**7**:45477.
113. Ren X, Wang Y, Chen L, *et al*. ellipsoidFN: a tool for identifying a heterogeneous set of cancer biomarkers based on gene expressions. *Nucleic Acids Res* 2013;**41**(4):e53.
114. Kayano M, Matsui H, Yamaguchi R, *et al*. Gene set differential analysis of time course expression profiles via sparse estimation in functional logistic model with application to time-dependent biomarker detection. *Biostatistics* 2016;**17**(2): 235–48.
115. Wang GM, Kuai D, Yang YD, *et al*. Screening of potential gene markers for predicting carotid atheroma plaque formation using bioinformatics approaches. *Mol Med Rep* 2017;**15**(4): 2039–48.
116. Chen X, Liu MX, Yan GY. RWRMDA: predicting novel human microRNA-disease associations. *Mol Biosyst* 2012;**8**(10): 2792–8.
117. Zhou M, Diao Z, Yue X, *et al*. Construction and analysis of dysregulated lncRNA-associated ceRNA network identified novel lncRNA biomarkers for early diagnosis of human pancreatic cancer. *Oncotarget* 2016;**7**(35):56383–94.
118. Wang P, Ning S, Zhang Y, *et al*. Identification of lncRNA-associated competing triplets reveals global patterns and prognostic markers for cancer. *Nucleic Acids Res* 2015;**43**(7): 3478–89.
119. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 2008;**9**:559.
120. Shao T, Wu A, Chen J, *et al*. Identification of module biomarkers from the dysregulated ceRNA-ceRNA interaction network in lung adenocarcinoma. *Mol Biosyst* 2015;**11**(11): 3048–58.
121. Wang YC, Chen BS. A network-based biomarker approach for molecular investigation and diagnosis of lung cancer. *BMC Med Genomics* 2011;**4**(1):2.
122. Liu X, Wang Y, Ji H, *et al*. Personalized characterization of diseases using sample-specific networks. *Nucleic Acids Res* 2016;**44**(22):e164.
123. Farahmand S, Goliaei S, Ansari-Pour N, *et al*. GTA: a game theoretic approach to identifying cancer subnetwork markers. *Mol Biosyst* 2016;**12**(3):818–25.
124. Zhao XM, Liu KQ, Zhu G, *et al*. Identifying cancer-related microRNAs based on gene expression data. *Bioinformatics* 2015;**31**(8):1226–34.
125. Mukhopadhyay A, Maulik U. An SVM-wrapped multiobjective evolutionary feature selection approach for identifying cancer-microRNA markers. *IEEE Trans Nanobioscience* 2013; **12**(4):275–81.
126. Xu J, Li CX, Lv JY, *et al*. Prioritizing candidate disease miRNAs by topological features in the miRNA target-dysregulated network: case study of prostate cancer. *Mol Cancer Ther* 2011; **10**:1857–66.
127. Yang Y, Huang N, Hao L, *et al*. A clustering-based approach for efficient identification of microRNA combinatorial biomarkers. *BMC Genomics* 2017;**18**(Suppl 2):210.
128. Li Y, Dong Y, Huang Z, *et al*. Computational identifying and characterizing circular RNAs and their associated genes in hepatocellular carcinoma. *PLoS One* 2017;**12**(3):e0174436.
129. Samuel N, Id Said B, Guha T, *et al*. Assessment of TP53 Polymorphisms and MDM2 SNP309 in premenopausal breast cancer risk. *Hum Mutat* 2017;**38**(3):265–8.
130. Wang Y, Makedon FS, Ford JC, *et al*. HykGene: a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data. *Bioinformatics* 2005;**21**(8):1530–7.
131. Zou M, Zhang PJ, Wen XY, *et al*. A novel mixed integer programming for multi-biomarker panel identification by distinguishing malignant from benign colorectal tumors. *Methods* 2015;**83**:3–17.
132. Ding Y, Chen M, Liu Z, *et al*. atBioNet–an integrated network analysis tool for genomics and biomarker discovery. *BMC Genomics* 2012;**13**:325.
133. Wen Z, Zhang W, Zeng T, *et al*. MCentridFS: a tool for identifying module biomarkers for multi-phenotypes from high-throughput data. *Mol Biosyst* 2014;**10**(11):2870–5.
134. Zhang W, Zang J, Jing X, *et al*. Identification of candidate miRNA biomarkers from miRNA regulatory network with application to prostate cancer. *J Transl Med* 2014;**12**(1):66.
135. Yan W, Xu L, Sun Z, *et al*. MicroRNA biomarker identification for pediatric acute myeloid leukemia based on a novel bioinformatics model. *Oncotarget* 2015;**6**(28):26424–36.
136. Shen L, Lin Y, Sun Z, *et al*. Knowledge-guided bioinformatics model for identifying autism spectrum disorder diagnostic MicroRNA biomarkers. *Sci Rep* 2016;**6**(1):39663.
137. Vaquerizas JM, Kummerfeld SK, Teichmann SA, *et al*. A census of human transcription factors: function, expression and evolution. *Nat Rev Genet* 2009;**10**(4):252–63.
138. Chen J, Zhang D, Zhang W, *et al*. Clear cell renal cell carcinoma associated microRNA expression signatures identified by an integrated bioinformatics analysis. *J Transl Med* 2013;**11**(1):169.
139. Zhu J, Wang S, Zhang W, *et al*. Screening key microRNAs for castration-resistant prostate cancer based on miRNA/mRNA functional synergistic network. *Oncotarget* 2015;**6**(41):43819–30.

140. Zhu Y, Peng Q, Lin Y, *et al*. Identification of biomarker microRNAs for predicting the response of colorectal cancer to neoadjuvant chemoradiotherapy based on microRNA regulatory network. *Oncotarget* 2017;**8**:2233–48.
141. Yan W, Wang S, Sun Z, *et al*. Identification of microRNAs as potential biomarker for gastric cancer by system biological analysis. *Biomed Res Int* 2014;**2014**:901428.
142. Huang J, Sun Z, Yan W, *et al*. Identification of microRNA as sepsis biomarker based on miRNAs regulatory network analysis. *Biomed Res Int* 2014;**2014**:594350.
143. Zhu Y, Lin Y, Yan W, *et al*. Novel biomarker MicroRNAs for subtyping of acute coronary syndrome: a bioinformatics approach. *Biomed Res Int* 2016;**2016**:4618323.
144. Tang Y, Yan W, Chen J, *et al*. Identification of novel microRNA regulatory pathways associated with heterogeneous prostate cancer. *BMC Syst Biol* 2013;**7(Suppl 3)**:S6.
145. Sagner M, McNeil A, Puska P, *et al*. The P4 health spectrum—a Predictive, Preventive, Personalized and Participatory Continuum for promoting healthspan. *Prog Cardiovasc Dis* 2017;**59**(5):506–21.