



Big-data-based edge biomarkers: study on dynamical drug sensitivity and resistance in individuals

Tao Zeng, Wanwei Zhang, Xiangtian Yu, Xiaoping Liu, Meiyi Li and Luonan Chen

Corresponding author. Luonan Chen, Key Laboratory of Systems Biology, Innovation Center for Cell Signaling Network, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 320 Yue-Yang Road, Shanghai 200031, China. Tel.: +86 21-5492-0100; Fax: +86 21-5497-2551; E-mail: lncchen@sibs.ac.cn

Abstract

Big-data-based edge biomarker is a new concept to characterize disease features based on biomedical big data in a dynamical and network manner, which also provides alternative strategies to indicate disease status in single samples. This article gives a comprehensive review on big-data-based edge biomarkers for complex diseases in an individual patient, which are defined as biomarkers based on network information and high-dimensional data. Specifically, we firstly introduce the sources and structures of biomedical big data accessible in public for edge biomarker and disease study. We show that biomedical big data are typically 'small-sample size in high-dimension space', i.e. small samples but with high dimensions on features (e.g. omics data) for each individual, in contrast to traditional big data in many other fields characterized as 'large-sample size in low-dimension space', i.e. big samples but with low dimensions on features. Then, we demonstrate the concept, model and algorithm for edge biomarkers and further big-data-based edge biomarkers. Dissimilar to conventional biomarkers, edge biomarkers, e.g. module biomarkers in module network rewiring-analysis, are able to predict the disease state by learning differential associations between molecules rather than differential expressions of molecules during disease progression or treatment in individual patients. In particular, in contrast to using the information of the common molecules or edges (i.e. molecule-pairs) across a population in traditional biomarkers including network and edge biomarkers, big-data-based edge biomarkers are specific for each individual and thus can accurately evaluate the disease state by considering the individual heterogeneity. Therefore, the measurement of big data in a high-dimensional space is required not only in the learning process but also in the diagnosing or predicting process of the tested individual. Finally, we provide a case study on analyzing the temporal expression data from a malaria vaccine trial by big-data-based edge biomarkers from module network rewiring-analysis. The illustrative results show that the identified module biomarkers can accurately distinguish vaccines with or without protection and outperformed previous reported gene signatures in terms of effectiveness and efficiency.

Tao Zeng is an associate professor at Key Laboratory of Systems Biology, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences. His research interest includes network biology and computational biology.

Wanwei Zhang is a PhD student at Key Laboratory of Systems Biology, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences. His main interest is in machine learning and biomarker discovery.

Xiangtian Yu is a PhD student at School of Mathematics, Shandong University. Her main interest is in mathematical model and optimization.

Xiaoping Liu is a postdoctoral fellow at Key Laboratory of Systems Biology, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences. His main interest is in systems biology, dynamical network biomarkers and NGS.

Meiyi Li is a PhD student at Key Laboratory of Systems Biology, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences. Her main interest is in developing computational methods to understand disease progression at a molecular level.

Luonan Chen is professor and executive director of Key Laboratory of Systems Biology, Innovation Center for Cell Signaling Network, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences. His research interest is computational systems biology and bioinformatics.

Submitted: 15 April 2015; Received (in revised form): 26 July 2015

© The Author 2015. Published by Oxford University Press. For Permissions, please email: journals.permissions@oup.com

Key words: big data; edge biomarker; dynamical drug sensitivity and resistance; dynamical network biomarker; module biomarker; personalized medicine

Introduction

Characterizing individual diseases is a key to achieve the precision medicine by indicating disease states of individuals, and also personalized medicine by designing treatments for individuals [1–3]. In general, the specific features of individuals for complex diseases are indicated by the sequence mutations or single nucleotide polymorphisms (SNPs) [4–7]. However, SNPs and mutations generally provide static characteristics on the personalization, e.g. risk of a patient on disease, rather than the current disease state of the patient. In other words, they fail to tell us the dynamical characteristics of a patient on disease development or treatment, e.g. disease state and its critical transition. In contrast, the newly precision medicine promises to ‘deliver the right treatments, at the right time, every time to the right person’ [8–12], which requires not only static but also dynamical features of individuals. Along with the development of translational medicine [13–20], increasingly accumulated new biomedical data as well as the new technologies become available, and can be exploited to accurately diagnose the disease states and also design precise treatment of individual patients. Particularly, to improve the accuracy of diagnosis and prediction, a key issue is to characterize and quantify the dynamical characteristics of disease progression and treatment process [16], e.g. identifying dynamical network biomarkers [21], and drug sensitivity and resistance biomarkers. To address this problem, developing big-data-based novel biomarkers for the diagnosis or prognosis purpose of diseases for individual patients is one important and necessary determinant.

‘Big data’ are characterized by four ‘V’ features emerged in traditional research fields (e.g. society or economy) [22], i.e. Volume, Variety, Value and Velocity. In contrast, in the field of biology or biomedicine, the big data generally have different features, which lie in: (i) small sample size but in high (or big)-dimension space, e.g. omics data for each individual, rather than the traditional big sample size but in low (or small)-dimension space; in other words, biomedical big data are ‘small-sample size in high-dimension space’ or small big-data, comparing with traditional ‘large-sample size in low-dimension space’ or big small-data in many other fields [23, 24]; (ii) the strong diversity or heterogeneity of biomedical big data, which are observed at differential levels, scales and viewpoints [25, 26]; and (iii) the high-value density of biomedical big data with abundant and dynamical information [27–29]. Such features on biomedical big data demand special theoretical and computational methodologies to exploit the information for elucidating essential mechanisms of biological phenomena and complex diseases at the system level, and turning data into meaningful biological applications and knowledge. In particular, to apply such biomedical big data in precision medicine or personalized medicine, there is a pressing need to integrate them on the systematical level for biomarker discovery by unified mathematical methods. Currently, one of such promising methods is attracting wide attention on analyzing biomedical big data, known as the network biology [30, 31] and network biomarkers [32, 33], i.e. exploring network information for biomarker discovery. On the other hand, based on the biological data, many interactions or associations among molecules can be embedded into the biomarker discovery or directly used as novel biomarkers for disease

prediction [32, 34]. Biomarkers are evolving from individual molecules to a network of molecules (Figure 1), e.g. network biomarkers or edge biomarkers [21, 32, 34]. It has been well recognized that a biological function or signal transduction involved in phenotype changes, e.g. disease occurrence or disease recovery [32, 34], is facilitated by the associated interactions (or edges) between molecules, rather than individual molecules. In fact, the concept of ‘edgotype’ has already linked the genotype as interactions to phenotype [35]. Meanwhile, the intensive researches on ‘edgetics’ have also revealed the malfunctions of interactions as the key molecular mechanisms relevant to so-called ‘edgetics’ diseases. In all, edge biomarkers for precision medicine or personalized medicine are expected to achieve the accurate diagnosis and prognosis by combining biomedical big data and network information.

Currently, one colossal difficulty for the biomarker discovery is big samples on population but small samples on each individual [34], i.e. so-called small big-data problem concerning extreme unbalance of sample sizes between population and individual. Because of high heterogeneity of disease on each person, the diagnosis and prognosis have to be conducted on an individual basis, which implies the necessity of the individual biomarkers derived from the limited small samples on each patient. In the academic studies, researchers have the possibility to collect many samples from animal models or clinical patients, and carefully design the measurements of big data required by the disease study. But, in the real clinical applications, there will be many unavoidable economic or clinic problems, e.g. huge cost of advanced disease test or long-term follow-up of patients, so that the clinical data about one patient are much less than that available to a general disease cohort. For example, in the clinical experiment on drug therapy, it is easy to collect many blood samples of one volunteer at consecutive intervals/hours, but, in a hospital, physicians are always required to diagnose a patient by only one or few blood samples. Therefore, the research study by learning from big samples of one individual would be different from the clinical application by predicting from small samples of one individual. To apply big-data-based biomarkers in such a situation, it is necessary to generate dynamical features rather than conventional static features from available samples, and measure or evaluate them on one or a few samples in an individual basis. Thus, to challenge this problem, big-data-based edge biomarkers including dynamical network biomarker (DNB) [21] were developed, and they are those biomarkers derived from the network information and biomedical big data (in high-dimensional space) that are expected to diagnose and predict the disease state on an individual basis.

As a representation of big-data-based biomarkers, big-data edge biomarker is a new concept to characterize disease features from multiple samples with high-dimensional data in a network manner [34], which also provides alternative strategies to indicate disease or predict treatment response in a single sample. An advantage of edge biomarker is to exploit the information of differential correlations or differential associations disregarded in the conventional approaches. Traditionally, molecular biomarkers or node biomarkers are the molecules identified from the differentially expressed genes (DEGs) or molecules, thereby leaving a large amount of the non-differentially expressed genes (NDEGs) or molecules unexamined [36]. But, the ‘edgotype’ (or edgetics) of many

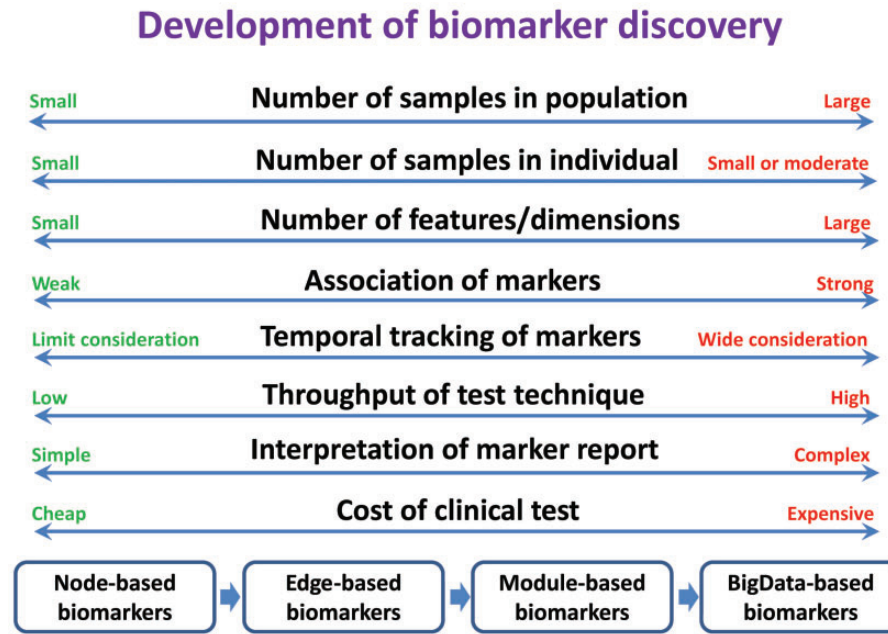


Figure 1. Biomarker discovery.

NDEGs have shown their important roles in the state alterations of biological systems [37, 38]. Dissimilar to individual DEGs to indicate the over/down regulation in different conditions, two interactive NDEGs (as a gene-pair or edge) can exhibit changed correlation from positive to negative regulation, or vice versa. These changes of genes' correlations result in the different states of a biological system (e.g. normal, diseased or treated). Therefore, the networks of biomarkers or edge biomarkers are able to identify more essential features of biological systems than the conventional molecule biomarkers, even from those NDEGs. On the other hand, the rapid advance on high-throughput technologies provide high-dimensional omics data for network-based biomarker discovery [27, 39], where an interaction or edge for a pair of molecules can be represented by their correlation [e.g. Pearson coefficient correlation (PCC) between a pair of molecules], numerically estimated from multiple samples. To some extent, the multiple samples collecting for one person is becoming reality, e.g. an integrative personal omics profile including genomic, transcriptomic, proteomic, metabolomic and autoantibody profiles from a single individual over a 14 month period [40]. By using such data of small samples but with high dimensions, big-data-based edge biomarkers based on module network rewiring-analysis (MNR) [41] were developed and can distinguish consistent gene modules and rewired module interactions and further measure the activities of those modules and interactions as the biomarker indicators. MNR has achieved superior performance on predicting therapy response of individual patients by analyzing temporal transcriptional profiles [41]. In addition, another effective strategy to identify big-data-based edge biomarkers in single samples is an EdgeMarker approach and an EdgeBiomarker approach [36, 42], which were developed to address the difficulty for obtaining correlations or edges from one sample. Specifically, EdgeMarker decomposes PCC into multiple elements so that a new vector embedding correlation-like information instead of conventional vector of raw expression can be used to quantify the edge biomarker even in one sample [36, 42]. Dissimilar to conventional biomarkers, big-data-based edge biomarker (e.g. EdgeMarker, EdgeBiomarker or module biomarker in MNR) is able to predict the disease state by differential associations between molecules

rather than differential expressions of molecules during disease progression or treatment in individual patients. In particular, in contrast to using the information of the common molecules or edges across a population in traditional biomarkers including network and edge biomarkers, big-data-based edge biomarkers are further specific for each individual so as to accurately evaluate the disease state by considering the individual heterogeneity, and thus high-dimensional data are required not only in the learning process but also in the diagnosing or predicting process of the tested patient.

This article intends to give a comprehensive review on biomedical big data and edge biomarkers currently available for diagnosing and predicting complex diseases in an individual patient. The section on 'Big data sources for studying complex diseases and biomarkers' mainly describes the sources and structures of biomedical big data accessible in public for disease study. 'Big-data-based edge biomarkers for diagnosing and predicting disease states for each individual' briefly demonstrates the concept, model and algorithm for identifying edge biomarkers by big data. 'Rationale and hypothesis of big-data-based edge biomarkers for predicting dynamical drug sensitivity and resistance' provides a case study for dynamical drug sensitivity and resistance by edge biomarkers based on MNR.

Big data sources for studying complex diseases and biomarkers

At present, there are many publicly available data sources for biology and medicine, which can be explored to study biomarkers. They can be categorized as database, type and tool platform, which are briefly summarized in [Supplementary Table S1](#). One typical feature of biomedical big data is 'small-sample size in high-dimension space', e.g. omics data for each individual, and it is big samples on populations but small samples on individuals, in contrast to the traditional 'large-sample size in low-dimension space' data in many other fields. Thus, the biomedical big data for individuals are characterized by

small samples with high dimensions, and it is a challenging task to exploit the information from such big data for biomedical applications [43, 44]. The conventional 'big data' emphasizes large samples, typically in low-dimensional space; however, the number of samples in biomedical data is large on populations but small on individuals, and also the data are generally represented in high-dimensional space, i.e. it is high-dimensional and small-sample size data. For instance, gene expression data, protein expression data and medical imaging data for each individual are such biomedical data. In clinical practice, diagnosis, prognosis and treatment are all conducted on an individual basis, and thus typical biomedical data for each individual are considered as 'small-sample size in high-dimension space'. In other words, the number of samples for a population is large, e.g. many blood samples collected from a cohort of patients, but the number of samples for an individual is still small, e.g. few blood samples collected from one patient, and thus 'small-sample size in high-dimension space' characterizes such big biomedical data.

The representative databases include: (i) databases depositing high-throughput data, e.g. TCGA [45, 46], NCBI GEO [47, 48], EMBL-EBI [49, 50] or GigaDB [51, 52]; (ii) databases for biological sequences and elements, e.g. ENCODE [53]; (iii) databases depositing experimental results about cells exposed to a variety of perturbing agents, e.g. LINCS [54, 55]; (iv) databases for physical or associated networks, e.g. KEGG [56, 57] or STRING [58]; and (v) databases depositing prior-known functional annotation of biological elements, e.g. Gene Ontology [59]. The representative data sets include: (i) DREAM Challenges providing potential benchmarks for assessing the cellular network inference and quantitative model [60, 61]; (ii) WTCCC providing well-designed solutions on genome-wide association studies to understand the patterns of human genome sequence variation [62]; and (iii) Cancer Cell Line Encyclopedia characterizing the complete genetic determinations of about 1000 human cancer cell lines, with DNA copy number, messenger RNA expression, mutation data and more [63]. The representative tool platforms are Cytoscape [64–66] and Ingenuity pathway analysis [67]. The representative research projects usually aim to provide large number of samples and data to reveal the system-wide features of living organisms. To provide a comprehensive source on human genetic variations, 1000 Genomes Project [68, 69] is the first project to sequence the genomes of a large number of people, and to find most genetic variants that have frequencies of at least 1% in the populations studied. Along with reduction of sequencing cost, several projects have been proposed to cover more persons. 100K Genomes Project [70] focuses on patients with a rare disease or cancer and their families, and plans to sequence whole genomes of 100 000 patients from National Health Service. 100K Wellness Project [71] provides new concepts and technologies to achieve the fundamentally change on how healthcare is practiced. As its a 10 month pilot study, the Hundred Person Wellness Project is to optimize wellness and minimize disease for 100 'well' individuals by quantifying self measures from each individual [72], whose data are collected and integrated from whole-genome sequencing, gut microbiome and clinical laboratory tests. The final goal of 100K Wellness Project [72] is to provide markers and their multi-parameters, quantifiable wellness metrics to evaluate/predict the early disease transitions for most common diseases, further to achieve earlier disease intervention and, finally, to transit the individual from a disease back to a wellness trajectory as early as possible [72].

On the other hand, although there are so many data sources, they have significant different representative structures. The

typical structures of biological data are briefly shown in [Supplementary Table S2](#). Firstly, annotation data [59] and network data [58] are the common knowledge bases in biology or medicine. The annotation data for biological elements (e.g. genes, proteins, interactions) are usually represented as a topological structure (e.g. tree), and each element is described on the origin, function, mutation and more [59]. The network data are represented as a table: each row represents an interaction or association; commonly the first and second columns give the symbols of genes/proteins involved in the interaction, and the third column points a weight value to indicate the confidence on the this interaction [58]. Obviously, a gene may have multiple functions annotated because of its participation on several interactions. However, these common knowledge bases are nonspecific for biological conditions, and thus cannot provide specific functions of this gene in a particular biological condition, or disease. In contrast, the omics data can provide condition-specific data [73–75]. Generally, there are four kinds of structures of individual-specific omics data: (i) control-case data, e.g. the expression data collected from the tissues or cells at time points before and after perturbation [76]; (ii) multiple-level data, e.g. the expression data, methylation data and sequence data collected for a single disease sample [45, 46]; (iii) multiple-time data, e.g. the expression data of blood samples collected at different time points for a patient receiving treatment [77]; and (iv) multiple-time and level data, as a combination of multiple-level data and multiple-time data, e.g. the expression data, metabolic data and clinical data collected for one person at different life time [40].

Big-data-based edge biomarkers for diagnosing and predicting disease states for each individual

Recently, edge biomarker, network biomarker and DNB [21, 33] have been proposed to study the biological system and its dysfunction in a systematical and dynamical manner, which have demonstrated great potential to predict the disease state in an individual basis of patients, in particular, by exploiting the association and dynamical information from biomedical big data.

Generally, biomarkers are used to evaluate the states of diseases or other changed phenotypes. The discovery of biomarker includes two stages ([Supplementary Figure S1A](#)): (i) identification of biomarkers from the available data of the collected samples from multiple individuals, and (ii) application of the biomarkers on the data of a new test sample from one individual. In terms of mathematical models, the identification of biomarker is known as a process of learning from the available data, which uses machine learning or optimization techniques to identify several features (e.g. molecule markers) to discriminate different phenotypes (e.g. response or nonresponse of treated samples). Meanwhile, the application of biomarkers is just a step of predicting on new data, which uses the identified features to evaluate/predict the phenotype of a test sample (e.g. if a sample or its represented individual has treatment response or not). Recent research works have classified the general biomarkers into four categories ([Supplementary Figure S1B](#)) according to the type of network information used in above two steps [34]. Briefly speaking, (i) node or molecular biomarkers, which exploit the information of differential expressions on a number of individual molecules in both two steps, e.g. differential expression of genes [78] or differential mutation of genes [79]; (ii) network-based biomarkers or network-weighted biomarkers, which exploit correlation or association information

between molecules to identify interactive molecules in the learning step, but only molecules (or a molecule set) without their network information in the predicting step, e.g. PinnacleZ approach extracting discriminative subnetworks rather than individual genes from a protein interaction network [80] or CORGs-based classification method to infer the activity level of a given pathway [81]; (iii) edge biomarkers (or network biomarkers), which exploit correlation or association information between molecules in both two steps for learning and predicting, e.g. PPI-SVM-KNN model proposed to classify time series gene expression via integration of biological networks [82] or EdgeMaker approach to extract differentially correlated gene-pairs (DCPs) from NDEGs [36]; (iv) DNB or dynamical edge biomarker explores dynamical information of data together with network information in two steps, which is able to further detect the critical stage just before the serious deterioration of a disease during the disease progression, e.g. MNR designed based on the temporal expression data of multiple individuals [41] or DNBs identified from dynamic protein-protein interaction networks to analyze the underlying mechanisms of complex diseases [83]. Different from conventional node biomarkers to explore differential expressions of molecules between a disease state and a control state, edge biomarkers are expected to learn and predict by exploring differential associations among molecules, which can provide a systematical and dynamical way to decipher the biological system responding to drug or therapy treatment. Clearly, types (iii) and (iv) could be big-data-based edge biomarkers because they are derived from the network information and big data, and also the prediction on a test sample (or a patient) by them requires the high-dimensional data of such a sample (or patient). In other words, big-data-based biomarkers are generally sample or individual specific either on evaluation of the biomarkers or on composition of the biomarkers, although many of those markers are common across all populations depending on the personal characteristics.

An unavoidable problem for both edge biomarkers and network biomarkers is the requirement of multiple samples in the predicting step. For a particular patient, when multiple samples are available, node biomarkers can make the diagnosis or prediction based on the expressions of those marker molecules, e.g. average expression, or relative value (e.g. average fold change) or even statistic value (e.g. *P* value); meanwhile, edge biomarkers can do it based on the correlations, e.g. PGC from the multiple samples. But, when only one sample is available for an individual, the correlations cannot be evaluated on one sample directly although there is no such problem for node biomarkers. Therefore, the approaches of edge biomarkers can be further grouped to two categories as single sample-based and multiple sample-based methods [34].

Generally, the main procedures of identifying above edge-based biomarkers include the following steps:

- i. Collecting data from experiments or databases, e.g. multiple-level omics data [84] or temporal expression data [85];
- ii. Transforming the original data from a form of node data to a form of edge data, e.g. from raw expression profiles [78] to correlation-like vectors [36];
- iii. Selecting feature nodes or feature edges by a feature selection method [86];
- iv. Quantifying the node biomarkers or edge biomarkers by scores, e.g. based on expression, correlation or activity [81];
- v. Building classification or prediction model based on node biomarkers, or edge biomarkers or their combinations by a machine learning method [87];

- vi. Evaluating the phenotype of a new test individual with one or more samples by the biomarkers and their corresponding measurements on one or more samples.

In summary, the high-dimensional data even with small samples, such as multiple-level data or multiple-time data, provide a great opportunity to achieve the personalized medicine or precision medicine in a systematical and dynamical manner, e.g. construct big-data-based edge biomarkers to predict the disease state and treatment response in an individual basis. Different from conventional biomarkers that use the information of only the common molecules or edges across a population including network and edge biomarkers, big-data-based edge biomarkers are specific for each individual either on evaluation of those common biomarkers or on composition of those biomarkers so as to accurately evaluate the disease state by considering the individual heterogeneity, and thus not only the learning process but also the diagnosing or predicting process requires to measure big data, i.e. high-dimensional data for each individual, not limited to the information of those obtained markers.

Rationale and hypothesis of big-data-based edge biomarkers for predicting dynamical drug sensitivity and resistance

Rationale and hypothesis of biomarkers for predicting treatment response

For conventional node-based biomarkers, their discoveries require the evaluation on the average expressions of markers, e.g. gene expressions are significantly different between control and case populations just like individuals with or without protection in drug treatments, which assumes that the functional gain and loss on gene expression/transcription are the determinants of cellular response to drug sensitivity and resistance (Figure 2). For example, to apply predictive biomarkers to optimal therapy on cancer patients, dynamic BH3 profiling has been used to predict chemotherapy response across many cancer types and many agents, including combinations of chemotherapies by measuring early drug-induced death signaling with differential $\Delta\%$ priming [88]. SHON has been identified as a novel human oncogene with predictive utility in ERp breast cancer, offering a simple biomarker to predict the therapeutic efficacy of antiestrogen therapy in patients with breast cancer [89].

For edge-based biomarkers, their discoveries require that the correlations between two markers, e.g. a gene-pair, are significantly different in control and case populations, which assumes that the functional gain and loss on gene association/interaction determine the response to drug sensitivity and resistance (Figure 2). For example, a framework named as Multivariate Organization of Combinatorial Alterations has been proposed to effectively combine many genomic alterations into biomarkers of drug response by using Boolean set operations coupled with optimization, which is extended from conventional one based on pairwise interactions [90]. An *in silico* approach, as a software tool ExprEssence, has been applied to identify specific mechanisms relevant for TFAC therapy response, from a gene/protein interaction network [91]. Besides, a simple prediction framework has been designed based on the genome-wide and quantitative profiling of cellular responses to individual drugs, whose correlation-based strategy can reveal the synergistic effects of drug combinations [92].

For module-based biomarkers, their discoveries depend on searching a combination of DEGs and DCPs in a form of network


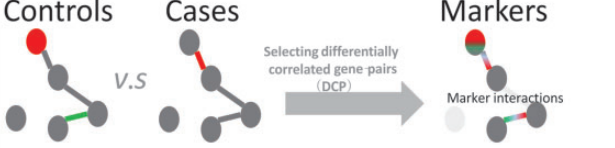
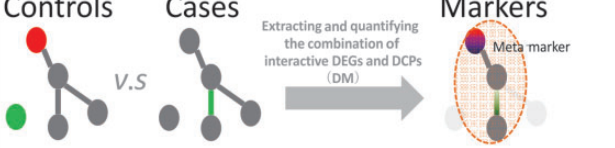
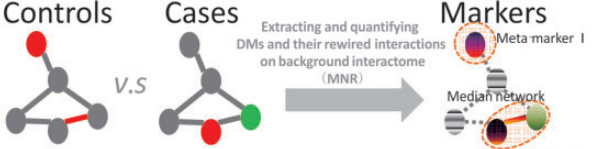
Different biomarkers in predicting drug sensitivity and resistance	Principle of biomarker discovery	Hypothesis of biomarker
 <p>Node based biomarkers</p>	The average expressions of markers, e.g. genes, are significantly different in control and case populations, e.g. individuals with or without protection in drug treatments.	Response to drug sensitivity and resistance should be executed by functional gain and loss on gene expression / transcription.
 <p>Edge based biomarkers</p>	The expression correlations between two markers, e.g. a gene pair, are significantly different in control and case populations.	Response to drug sensitivity and resistance should be executed by functional gain and loss on gene association / interaction.
 <p>Module based biomarkers</p>	An combination of DEGs and DCPs in a form of network (e.g. module), whose quantified measurements based on module members (e.g. activity of meta marker) should have significant differences between control and case populations.	Response to drug sensitivity and resistance should be executed by functional gain and loss on gene network rather than individual genes.
 <p>BigData based biomarkers</p>	An combination of modules in a form of network of networks (e.g. module network), whose quantified measurements dependent on module members (e.g. activity of meta markers) and all other members (e.g. activity of module network connecting meta-markers) should have significant differences between control and case populations.	Response to drug sensitivity and resistance should be executed by consistent functional gain and loss on some gene networks (e.g. population-common variants) and occasional changes among them (e.g. individual-specificity variants).

Figure 2. Summary on the rationale and hypothesis of biomarkers.

(e.g. module), whose quantitative measurements based on the members of the module (e.g. the activity of the corresponding meta-marker) should have significant differences between control and case populations (Figure 2). The hypothesis underlying such kind of biomarkers is that the response to drug sensitivity and resistance should be executed by functional gain and loss on the gene network rather than individual genes. In cases of such biomarkers, an intact commensal microbiota that forms an interacted network as a module biomarker modulates myeloid-derived cell functions in the tumor microenvironment or optimal responses to cancer therapy [93]; and organ transplant recipients treated with a posttransplant therapy that combines immunosuppressive and antiviral drugs, can offer a new window into the effects of immune modulation to measure the health of the immune system, and the connections between immune strength and the viral component of the microbiome [85], which can be viewed as a module biomarker.

Particularly for big-data-based biomarkers, their extractions require to detect a combination of modules in a form of network of networks (e.g. module network), whose quantitative measurements dependent on their members (e.g. the activity of the corresponding meta-markers) and all other members (e.g. the activity of the module network connecting meta-markers) should have significant differences between control and case populations (Figure 2). The systematical assumption of this kind of biomarkers is that the response to drug sensitivity and resistance should be executed by the consistent functional gain and loss on

a group of network modules (e.g. population-common variations) and occasional changes among them (e.g. individual-specificity variations). In fact, several studies have provided evidences on the capability of whole high-throughput data as effective biomarkers, rather than the low-throughput data or dimensionality reduction from high-throughput data. Such researches include the following works: optical metabolic imaging has shown potential as a high-throughput screen technique to test the efficacy of a panel of drugs and to select optimal drug combinations for cancer treatment in individual patients [94]; a ridge regression model has been built for the prediction of chemotherapeutic response in patients using only before-treatment baseline tumor gene expression data, where the models are fitted for whole-genome gene expression against drug sensitivity in a large panel of cell lines by allowing every gene to influence the prediction [95]; a high-complexity barcode library, ClonTracer, has been developed to enable the high-resolution tracking of >1 million cancer cells under drug treatment, and provides quantitative assessment of the ability of combination treatments to suppress resistant clones [96]; rather than disease state, DNB was proposed to detect pre-disease state (or critical state) during disease progression [21, 33], and DNB is a group of molecules characterized with strong correlations but strong variations for those DNB molecules; and MNR [41] was designed based on the temporal expression data of multiple individuals, which has been carried on the analysis of Hepatitis C Virus (HCV) patients receiving interferon treatment.

Typical instances of experimental data and biomarker model for predicting treatment response

Depending on the type and structure of data, conventional mathematical models on characterizing and predicting treatment response can be grouped as static approaches and dynamical approaches.

Static approaches usually used the sequences from host or virus to evaluate the risk or probability of treatment response [97–99]. Taking the treatment of HCV patients as examples, the HCV genotypes of HCV [100], the nucleotide sequence of the hepatitis C virus genome [101], the genomic complexity of hepatitis C virus [102] and the host and virus genome variability [103] all have been reported to be associated with patient response to interferon. Those data were used to analyze the treatment response in many previous works.

Meanwhile, dynamical approaches explore the high-dimensional information of omics data, especially the temporal expression data currently (note, the control-case study can be considered as sampled at two different time points). There are several similar well-designed experiments on the temporal expression profiles of patients with different diseases, e.g. Influenza, HCV or multiple sclerosis (MS). According to the scale of sampling interval, the categories of temporal expression can be short interval (e.g. scale of hours), medium interval (e.g. scale of days or weeks) and long interval (e.g. scale of months or years), respectively. Disregarding the particular type of disease, or drug, there are some common mathematical models used to analyze the signatures of treatment response according to the organization of expression data, e.g. differential expression analysis [104, 105], machine learning analysis [87] and network-based analysis [32, 106, 107].

As an example of short-interval data, the gene expressions of Influenza patterns were collected at different hours. An important determinant of disease progression is known as the host response, especially for differentiating symptomatic and asymptomatic Influenza A infection [77]. To understand the host response on molecule levels, 17 healthy human volunteers received intranasal inoculation of influenza H3N2/Wisconsin, and 9 of them developed mild to severe symptoms according to standardized symptom scoring [77]. A total of 267 gene expression profiles were obtained for all volunteers at 16 time points, as an interval of ~ 8 h post inoculation (hpi) through 108 hpi including baseline (-24 hpi). In a way of bioinformatics analysis [77], Bayesian Linear Unmixing was used to establish an *ab initio* molecular signature strongly correlated to symptomatic clinical disease; EDGE and SOM were also used to investigate the key host factors temporally related to symptomatic and asymptomatic volunteers. Recently, DNB-based approaches (e.g. edge network analysis [108] or single sample-based DNB [109]) have been developed and applied on analyzing this data set, which not only predict the symptomatic or asymptomatic Influenza A infection but also identify the critical time before the symptom of infection appears for a particular volunteer.

As an example of medium-interval data, the gene expressions of HCV patients were measured at different days or weeks. To compare the treatments of HCV patients with different therapeutic regimens, the kinetics of gene expression of peripheral blood mononuclear cells (PBMC) were measured during the first 10 weeks [i.e. before treatment (Day 1) and at Days 3, 6, 10, 13, 27, 42 and 70 days after treatment] of therapy in 20 HCV patients treated with Pegylated-interferon- α 2b and ribavirin [110]. Differential expression analysis was carried on, and the genes with deregulation at given time points were investigated,

where the levels of gene up-regulation or down-regulation were found to be similar to those reported with Pegasys/ribavirin treatment previously [110]. In addition, a new classification algorithm (e.g. a time-dependent diagnostic model [111]) and new edge biomarker (e.g. MNR [41]) were proposed to further improve the discrimination on responders and nonresponders.

As an example of long-interval data, the time-course gene expressions were tested at different months or years. For MS patients treated by recombinant human interferon beta (rIFN β), a relatively large proportion of them are nonresponders [112]. To detect treatment-outcome-associated higher-order predictive patterns on expression of PBMC, a temporal kinetic reverse-transcription polymerase chain reaction data set was generated on 70 genes of 52 MS patients receiving rIFN β treatment [112]. Briefly, in the patients, 33 had good prognosis, and the other 19 had poor prognosis; and for each patient, the expression profiles of 70 genes on seven time points (0, 3, 6, 9, 12, 18 and 24 months after the treatment) were measured. From the viewpoint of data mining and predictive modeling [112], a Bayes' score was defined for each gene triple representing the probability of a patient being a good responder; and then several best predictive gene triplets (i.e. classifiers) were combined into a committee; finally, a majority voting scheme was used to decide which class a new sample would be assigned. Based on this data set, some new gene features (e.g. negative correlation based gene markers [113]) or new classification models (e.g. hidden Markov models [114] or HMM/GMM hybrid model [82]) were further proposed to improve the prediction power.

As described above, biomedical big data (e.g. high-dimensional omics data) indeed have been widely used in biomarker study and application. Conventional node biomarkers extract a few gene signatures from data with high dimensions according to differential expressions of separate molecules, and use only such signatures' low dimensions in prediction. Meanwhile, network and edge biomarkers can detect several gene-pair signatures from high-dimensional data according to differential associations of interactive molecules, and still use such low-dimensional signatures in prediction. Obviously, these biomarkers use the information of the common identified molecules or edges across a population disregarding individual specificity. In contrast, big-data-based edge biomarkers (e.g. module biomarker in MNR [41]) make full use of not only common molecules or edges (e.g. consistent modules in MNR [41]) but also conditional molecules or edges (e.g. module interactions rather than molecule interactions in MNR [41]). Thus, big-data-based edge biomarkers could be specific for each individual so as to accurately evaluate the disease state by considering the individual heterogeneity, in which high-dimensional data are required not only in the learning process but also in the diagnosing or predicting process for the tested patient.

A case study on biomarker discovery of dynamical drug sensitivity and resistance

As an illustration of big-data-based edge biomarkers in analysis of dynamical drug sensitivity and resistance rather than performance comparison, MNR is used in the reanalysis of temporal expression data from a malaria vaccine trial [115] for displaying the routine analysis and feature outcomes of big-data-based edge biomarkers.

MNR firstly builds time-specific network from individuals, then extracts the common/stable modules (termed consistent modules [41]) across those time-specific networks. Next, it uses

these modules as nodes, the interactions between modules as edges, to reconstruct the module network for each individual at particular time point or time period. Thirdly, an activity score of module or module interaction measured by molecular expressions and molecular associations was developed. Obviously, in the learning step, MNR can learn any features of the molecule network rather than individual molecules corresponding to a particular individual at a specific time, which could provide more clues (e.g. quantitative state of network by activity score) about the personal features at a molecular level, e.g. responses of treatment. In the predicting step, with at least three samples available for each individual, its molecule network can be estimated and the activity score can be calculated, and then these activity measurements can be applied to predict the state of biological system transition, e.g. outcome of treatment. Not limited to the measurement of those obtained markers, MNR requires high-dimensional data (omics data) or the measurement of other molecules (e.g. interactions between consistent modules) to evaluate the edge biomarkers (e.g. consistent modules) of the new patient because of the heterogeneity, and thus it is a method of big-data-based edge biomarkers. MNR was used to analyze dynamical drug sensitivity and resistance. The main results are summarized as follows.

- i. **Data description.** MNR [41] has been applied to reanalyze the temporal expression data from a malaria vaccine trial [115]. In this study, samples were collected at study entry, the day of third vaccination, 24 h and 72 h post the third vaccination and 2 weeks post the third vaccination. Thirteen of 39 vaccines were protected, and 26 of 39 were not protected. To guarantee each vaccine has expression data at the first five time points, the pre-processing gives gene expression profiles with 13 436 genes, and 120 samples (24 vaccines at corresponding five time points, where 11 are protected and 13 are not protected without delay). Obviously, these data would be multiple-time data from medium-interval experiment.
- ii. **Brief results of MNR.** Based on all patients' five temporal networks, we found 47 consistent modules and 44 appearance-consistent modules during the whole procedure of malaria vaccine trial, whose members are module network genes (MNGs). Using the consistent modules with significant KEGG pathway enrichment, module network can be reconstructed for a group of protected or non-protected at five time points, respectively, or reconstructed for each vaccine at three time windows, respectively. Each time window contains consecutive three time points, noted as $w_1 = [T_0, T_1, T_2]$, $w_2 = [T_1, T_2, T_3]$, $w_3 = [T_2, T_3, T_4]$, where, T_0 is at study entry; T_1 is on the day of the third vaccination; T_2 is 24 h after the third vaccination; T_3 is 72 h after the third vaccination; and T_4 is 2 weeks after the third vaccination.
- iii. **Comparison with conventional DEGs.** The previous study [115] gave a list of DEGs indicating the response of vaccine at 24 h post the third vaccination for all vaccines (Figure 3A and B). In contrast, MNGs show different expression patterns for protected vaccines and non-protected ones, respectively (Figure 3C and D). These genes would have functional associations during the process of vaccination, but have different association network for particular vaccine, which might induce the successful protection or not. They would have power to predict the final outcome of challenge, as protected or failed in the following discussion.
- iv. **Modules and their functional enrichments related to disease and treatment.** The consistent modules have significant enrichments on diverse biological functions. Many essential pathways were found to be enriched in several appearance-consistent modules (see in Table 1):
 - a. Relation to essential functions, e.g. chemokine signaling pathway and NOD-like receptor signaling pathway. There was a study on the physiologic role of the duffy blood group antigen, which serves as a receptor on the red blood cells for the malarial parasite [116]. NOD- and Toll-like agonists are important to instruct an appropriate adaptive immune response, whose ligands have possibility to generate new vaccine combinations [117].
 - b. Relation to HBV, e.g. Hepatitis B. There are some candidate malaria vaccines that have been produced based on hepatitis B virus core [118, 119].
 - c. Relation to other diseases, e.g. HTLV-I infection and cancer. There is a well-known phenomenon of biological false seropositivity with reactive EIAs and indeterminate western blot patterns, which has been attributed to possible cross-reactivity with malaria antigens [120]. Although there are few studies of the relationship of cancer to malaria, analogies have been reported at the cellular level for the two diseases; the antimalarial artesunate is possibly active against cancer; Epstein Barr Virus-related Burkitt's lymphoma is believed to require cofactors, such as malarial infection, for tumor development; even there is a relationship between malaria in the United States and brain tumor incidence [121].
- v. **Module-based prediction of vaccine protection/response in an individual patient.** Firstly, for groups of vaccines with or without protection, their modules' evaluation on differential expression (E-score) and evaluation on differential expression and correlation (EC-score) were calculated, respectively as defined in [41]. Seeing Figure 4, these activity measurements can discriminate the groups of protected or not well, especially, EC-score have better discrimination than E-score. This supports again differential correlation rather than differential expression would be discriminative features as novel biomarkers. Thus, the consistent modules and their temporal activities can reflect the difference between vaccines with or without protection on the level of networks. Secondly, the activity of module interaction (i.e. modules' connection weights defined in [41]) for each vaccine at a time period (i.e. consecutive three time points) were calculated. This new activity profile was used to build classification model by SVM, and predict the protection or non-protection for each vaccine. At different time periods/windows, the prediction accuracy (evaluated by average area under curve (AUC) on $100\times$ of 10-fold cross-validation) is high, especially approaching to 100% at the days after the third vaccination (Table 2). As control, the average expressions of DEGs reported in previous study [115] were also used, and their performance is no better than MNGs from consistent modules (Table 2). Besides, a few probes/biomarkers were identified to sort samples into protected and non-protected of disease categories with 100% accuracy at Day 5 after challenge [115], but they did not make prediction at the days before challenge. These results strongly support consistent modules, and their activities have high ability to distinguish protected and non-protected vaccines ahead of the previous biomarkers.

Finally, to quantitatively illustrate the efficiency of different categories of biomarkers in the context of dynamical drug

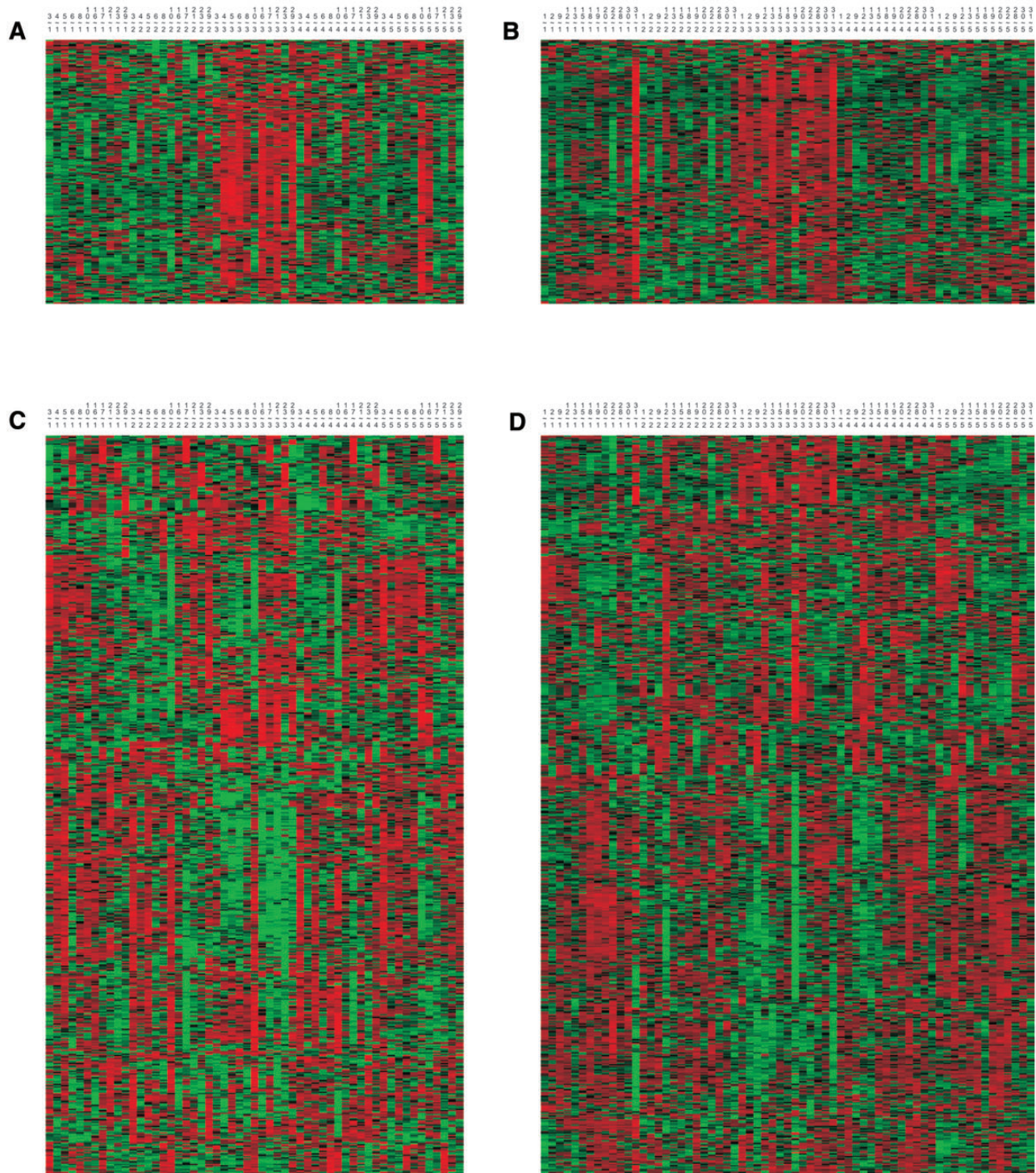


Figure 3. Expression profiles of DEGs and MNGs of individuals protected or non-protected. In each expression heat map, each row represents a gene and each column represents an individual. (A) Expression profiles of DEGs of individuals protected. (B) Expression profiles of DEGs of individuals non-protected. (C) Expression profiles of MNGs of individuals protected. (D) Expression profiles of MNGs of individuals non-protected.

sensitivity and resistance, a number of state-of-the-art methods have also been applied to this case study. Specifically, they include: DEG (conventional node biomarkers), AEP [122, 123] and frSVM [123, 124] (network-based biomarkers), PAC [81, 123], GSA [125] and Pathifier [126] (module-based biomarkers, or pathway-based biomarkers) and MNR (edge-based biomarkers, or big-data-based biomarkers). Similar to above evaluation on

DEG and MNG, five additional methods were also adopted to extract biomarkers from all data, meanwhile discriminating the protected and non-protected groups by using the average expressions or scores in each time window. Particularly, four methods, i.e. DEG, AEP, frSVM and MNR, possibly using biological networks rather than pathways, can provide feature genes in details. Thus, four groups of feature genes represented by

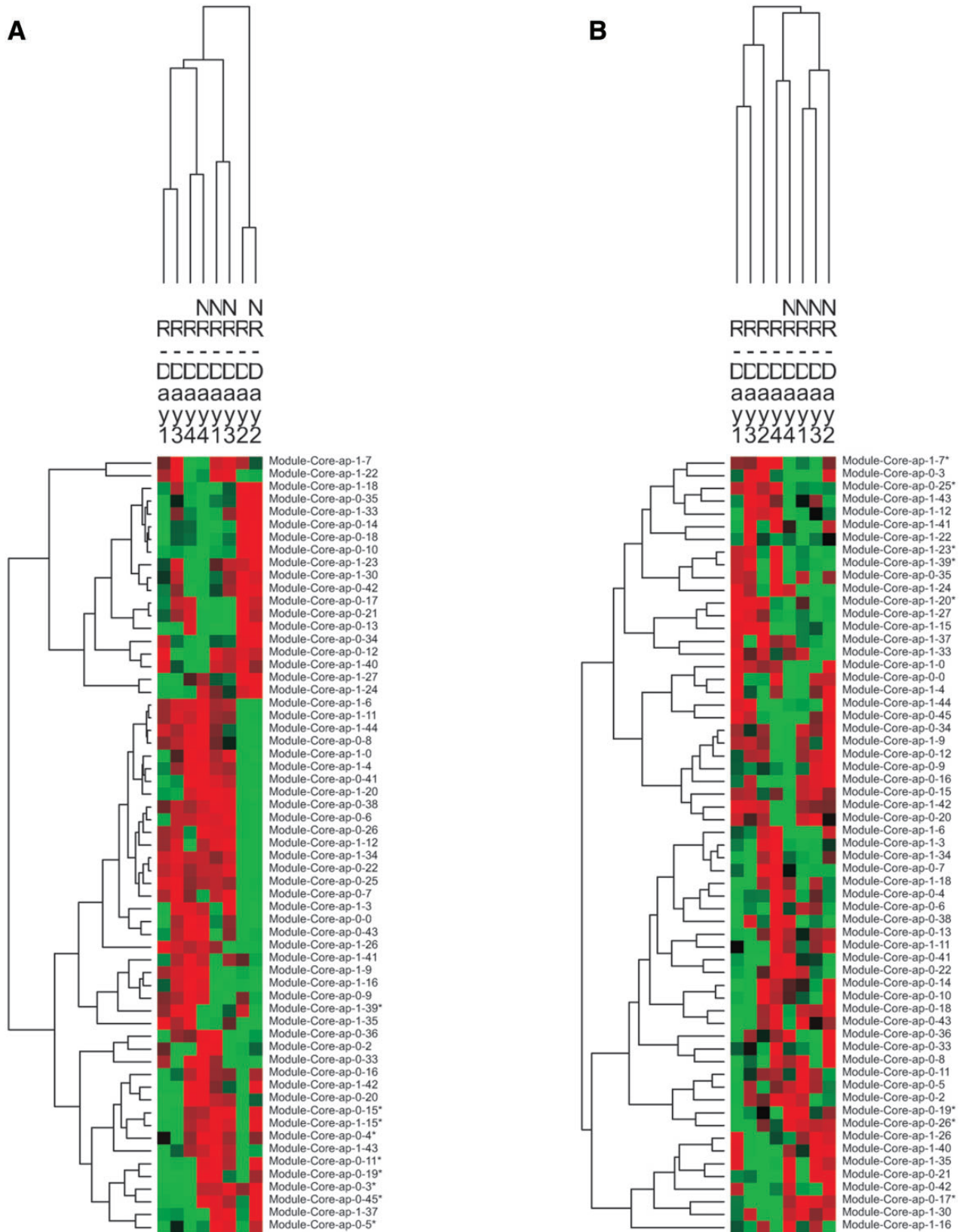


Figure 4. Different activity scores of consistent modules at different time points for two groups of vaccines. **(A)** E-score of consistent modules for groups of protected and non-protected at different time points. **(B)** EC-score of consistent modules for groups of protected and non-protected at different time points. Here, Day 1 represents the day of the third vaccination; Day 2 represents 24 h after the third vaccination; Day 3 represents 72 h after the third vaccination; and Day 4 represents 2 weeks after the third vaccination. Label R points toward group of protected, and NR points toward group of non-protected.

Table 1. Appearance-consistent modules with significant enrichments on KEGG pathways

Id	Gene members	KEGG pathways
Module-Core-ap-1-0	BCL2L1,HSF1,AKT1,PTK2B,POLR3C,CAPRIN1,APLP1,FLAD1,PRPF4,ZYX,OGDH,EIF3M,RBM34,RAN,FNBP4,HSP90AB1,PSMD6,CNPFY2,NAPA,BSG,RELA,WDR61,MAGOH,PNPLA6,ZNF207,YKT6,SHC1,WDR1,P4HB,SPAG5,RANGAP1,CENPA,UTP14A,DHX15,SMARCD1,SNRPD3,FKBP1A,RPL19,CSNK2A1,TWF1,SYNCRIP,NME6,ATIC,GTTF2F1,DNAJA4,PHB,GSK3A,RBBP4,MRPL42,TMEM189-UBE2V1,POP5,UBC,NCAPG,ACTB,ABCE1,IKBKG,CAPZB,STMN1,ELAVL1,MRPL23,METAP2,KHDRBS1,TM9SF4,EDC3,RC3H2,KRR1,UBE2L3,MATR3,MANEA,STIP1,ARHGAP1,MRPL15,SREBF2,MRPL16,GPR172A,NONO,CHCHD8,CCNE1,WDR18,FTSJ1,PAICS,RAB11B,TIMM13,PIK3R1,TRAF3IP3,PDHB,NUDT9,MRPL12,MRPL13,MYO19,TRO,DLD,POP4,TFE3,POLR2E,BCCIP,MMACHC,MRPL3,GART,ARPC4,ANAPC5,PBK,GINS2,PDAP1,PLSCR4,MRT04,UBE2M,CSNK1A1,ESR1,PSMD14,RPAP3,DSCR3,CDC20,MELK,RSL1D1,POLE3,FLOT2,RSL24D1,HNRNPK,IPO5,EIF4A1,EEF1E1,METTL2A,TGFB1,KIAA0020,PRKDC,CNBP,ARRB2,FBXO5,NVL,RPL36AL,IKZF1,RHOA,DCTN5,PKLR,TEK,RBM12,POLDIP3,ASPM,TOMM22,E2F4,USF2,PSMA3,GTPBP4,PSMA1,S100A9,PSMA7,PML,DDX19A,NFKB2,HNRNPC,RANBP1,TGFB2,WTAP,RPL5,TTC27,HNRNPA1,GNB1,DDX21,RXR, TXN2,GGA3,GGA2,RELB,UCHL5,PFN1,WAS,PTBP1,TARDBP,API5,UTP18,SHCBP1,IL2RG,AK2,MTF2,STOML2,HMG20A,NXT2,WDR12,RPL10A,RAD51,GABPA,CCT2,SMARCE1,MKI67,RPS16,VARS,DHX9,CEP57,RPS18,SF3B1,EIF6,PSMD10,MRPS17,CHAF1A,NIP7,SMARCA2,ATAD2,EIF2S1,RNGTT,CPSF1,CAD,PTPN22,NFX1,PRICKLE4,RBM42,IDH3G,PRPS1,RPL24,NRAS,DDX18,MAPKAPK2,APEX1,ETS1,UBE2G2,ATG5,EPRS,EXOSC5,DAPK3,ASF1A,COPA,RAC2,MCTS1,LSM5,NOLC1,LSM7,PMPCB,HGS,MRPS2,BUB3,MCM6,ARHGDI, CFL1,SSBP1,SNRPF,DEDD,ABCF2,CCNB1,DIS3,RPL26L1,EXOC7,SRRM1,RPL34,RPL32,NUP37,RBMX,TYMS,PRMT1,CUL4B,DCAF7,BCL3,EMG1,DKC1,ATF6,WDR77,M6PR	Chemokine signaling pathway Spliceosome Pancreatic cancer Ribosome biogenesis in eukaryotes Chronic myeloid leukemia HTLV-1 infection Ribosome
Module-Core-ap-1-3	PCNA,CCNA2,NCAPD3,FANCG,MCM4,KIAA0101,KIF20A	Cell cycle DNA replication
Module-Core-ap-1-4	BRCA1,RAD21,SMC3,H2AFX,ESPL1,MYST4	Cell cycle
Module-Core-ap-1-6	SKP2,SUPT16H,RFC4,MSH2,H2AFZ,RAD51AP1	Mismatch repair
Module-Core-ap-1-7	ATP5C1,ATP6V0C,ATP5A1,NDUFS1,MDH1,VDAC2	Parkinson's disease Alzheimer's disease Metabolic pathways Huntington's disease Oxidative phosphorylation
Module-Core-ap-1-11	THUMPD1,RPF1,RPL18A,RPL30	Ribosome
Module-Core-ap-1-12	IL7R,TUBA1C,MYC,TUBA1B	Gap junction
Module-Core-ap-1-15	MAP2K2,POLE,ARAF,PKM2	Pathogenic <i>Escherichia coli</i> infection Long-term potentiation ErbB signaling pathway Prostate cancer Glioma Non-small cell lung cancer Melanoma Renal cell carcinoma Endometrial cancer Acute myeloid leukemia Chronic myeloid leukemia Bladder cancer Long-term depression Central carbon metabolism in cancer Choline metabolism in cancer
Module-Core-ap-1-16	PCYT1A,MAPK1,KPNA4,EXOC5	Tuberculosis
Module-Core-ap-1-18	SYK,CARD9,RIPK2,ARHGAP25	NOD-like receptor signaling pathway T-cell receptor signaling pathway Measles Bacterial invasion of epithelial cells Chagas' disease (American trypanosomiasis) Fc gamma R-mediated phagocytosis
Module-Core-ap-1-20	PIK3CD,CD3E,DNM2	

(continued)

Table 1. Continued

Id	Gene members	KEGG pathways
Module-Core-ap-1-23	CES2,CDA,GUSB	Drug metabolism—other enzymes
Module-Core-ap-1-24	GSTA3, CYP26A1, CYP2A6	Metabolism of xenobiotics by cytochrome P450 Retinol metabolism Chemical carcinogenesis Drug metabolism—cytochrome P450
Module-Core-ap-1-27	GPX2,GSTT2,GGT5	Cyanoamino acid metabolism Glutathione metabolism Arachidonic acid metabolism Taurine and hypotaurine metabolism
Module-Core-ap-1-30	PSMC2,PSMD8,RRAGC	Proteasome
Module-Core-ap-1-34	FYN,NCK1,PTPN4	Axon guidance T-cell receptor signaling pathway Pathogenic <i>Escherichia coli</i> infection
Module-Core-ap-1-35	ARID4A,MYBL2,E2F1	HTLV-I infection
Module-Core-ap-1-37	EXO1,AURKB,HMGA1	Mismatch repair
Module-Core-ap-1-39	TAF5,POLR2B,CDK7	Basal transcription factors
Module-Core-ap-1-40	COL14A1,BGN,COL1A1	Protein digestion and absorption
Module-Core-ap-1-41	RPP30,RPP40,POP7	Ribosome biogenesis in eukaryotes RNA transport
Module-Core-ap-1-42	STAT6,TP53,SH3BGL3	Hepatitis B
Module-Core-ap-1-43	PA2G4,WBP11,PCBP1	Spliceosome
Module-Core-ap-1-44	ATF2,EZR,PRKACB	Insulin secretion Thyroid hormone synthesis Dopaminergic synapse Gastric acid secretion Estrogen signaling pathway Cocaine addiction Amphetamine addiction

Table 2. Prediction accuracy of individuals at different time windows

Time windows*	$w_1 = [T_0, T_1, T_2]$	$w_2 = [T_1, T_2, T_3]$	$w_3 = [T_2, T_3, T_4]$
MNGs	0.955 ± 0.0189	0.979 ± 0.0304	0.968 ± 0.0401
DEGs	0.930 ± 0.0526	0.929 ± 0.0627	0.789 ± 0.0718
AEPs	0.807 ± 0.092	0.812 ± 0.073	0.845 ± 0.079
frSVMs	0.648 ± 0.0995	0.643 ± 0.0754	0.530 ± 0.0689
Pathifiers	0.990 ± 1.11e-16	0.990 ± 1.11e-16	0.990 ± 1.11e-16
PACs	0.553 ± 0.0701	0.518 ± 0.099	0.533 ± 0.097
GSVAs	0.990 ± 1.11e-16	0.837 ± 0.0627	0.904 ± 0.073

* T_0 : at study entry; T_1 : on the day of the third vaccination; T_2 : 24 h after the third vaccination; T_3 : 72 h after the third vaccination; T_4 : 2 weeks after the third vaccination. The prediction accuracy is evaluated by the average AUC on 100× of 10-fold cross-validation and its variance.

DEGs, AEPs, frSVMs and MNGs were further compared on their network characteristics, so as to evaluate the biological interpretation of biomarkers.

On one hand, Table 2 shows the efficiency of different methods or biomarkers to discriminate the protected and non-protected groups, which are evaluated by AUC values. MNGs have been shown to be better than DEGs. The performance of DEGs decreases in the third time period. And AEPs display consistent performance in different time periods, although their AUC is not the highest. The AUC of frSVMs is not high, and decreases in the third time period as DEGs, which might be caused by frSVMs; this implies the importance of the association/network among feature genes. Meanwhile, among other three pathway-based methods or biomarkers, Pathifiers have the best

performance. MNGs and Pathifiers achieve comparable performances from two aspects: MNR is an unsupervised method to extract the gene modules and quantify the scores of modules; in contrast, Pathifier is a supervised method to quantify the scores of prior-known pathways. Besides, GSVA estimates the enrichment score of each pathway in one sample, and its performance is also high although it decreases in the second time period. The performance of PAC is not satisfactory in this case study, which may result from the small-sample size.

On the other hand, Figure 5 shows the comparison results among four groups of feature genes represented by DEGs, AEPs, frSVMs and MNGs.

- i. The edge biomarkers (i.e. MNGs) is better than the conventional node biomarkers (i.e. DEGs) and network-based biomarkers (i.e. AEPs and frSVMs), which are evaluated by the average efficiency (AUC) for discriminating the protected and non-protected groups as shown in Figure 5A.
- ii. The edge biomarkers and network-based biomarkers both have clear biological interpretation than node biomarkers, which is evaluated by the degree of genes in the biological network. Figure 5B shows the absolute degree of genes, where the absolute degree for a gene is the number of edges linked to this gene among all feature genes on the biological network. Figure 5C shows the relative degree of genes, where the relative degree of a gene is the ratio of numbers of edges linked to this gene among all feature genes and among all background genes. Obviously, the larger degree of feature genes means closer associations/interactions among the selected feature genes, which would indicate potential biological functions enriched in the feature genes.

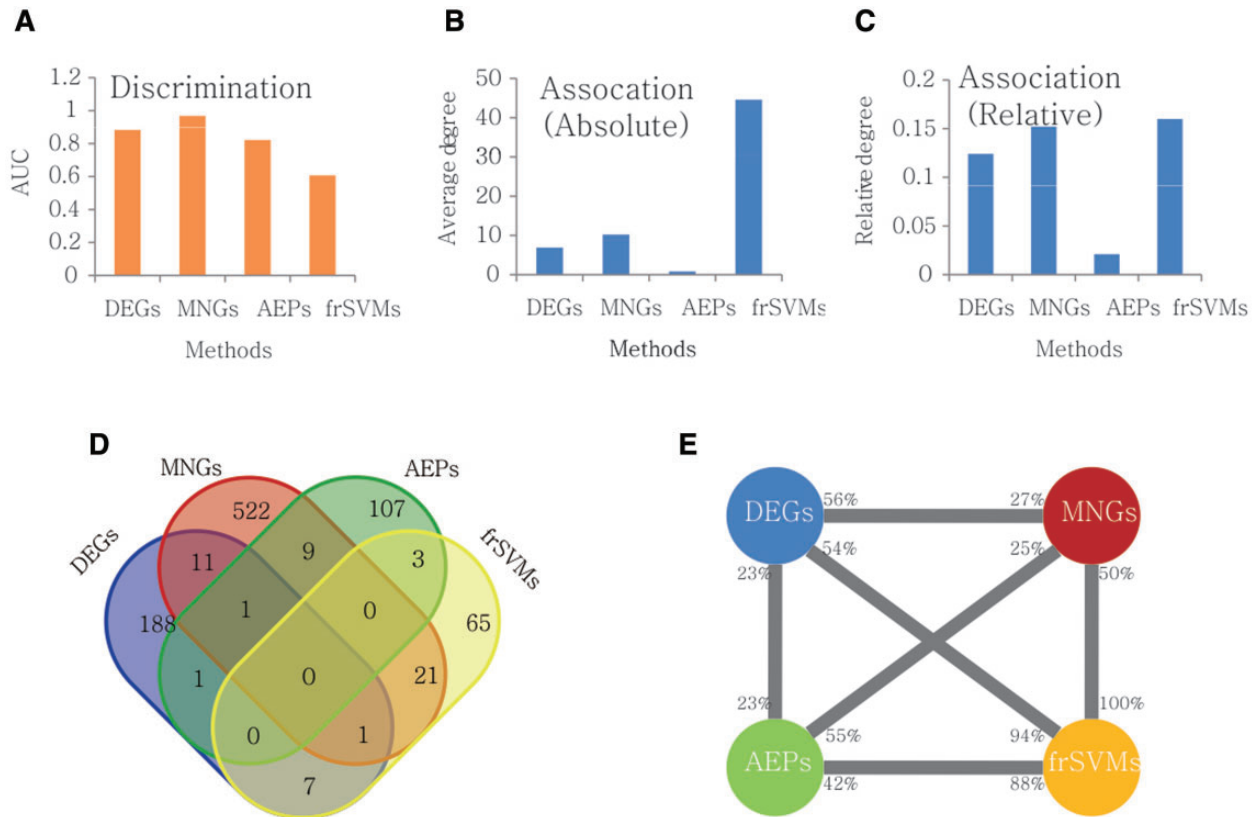


Figure 5. Comparison of feature genes selected by different methods in the case study. (A) The average AUC of four methods using biological networks rather than pathways, which evaluate the average efficiency for discriminating the protected and non-protected groups at different time points. (B) The average absolute degree of feature genes selected by different methods correspondingly, where the absolute degree for a gene is the number of edges linked to this gene among all feature genes, which evaluate the association/network among feature genes. (C) The average relative degree of feature genes selected by different methods correspondingly, where the relative degree of a gene is the ratio of numbers of edges linked to this gene among all feature genes and among all background genes, which evaluate the association/network among feature genes compared with backgrounds. (D) The number of overlapping genes among feature genes selected by different methods correspondingly. (E) The number of associations/edges among feature genes selected by different methods, where the values indicate the percentage of genes of one group feature genes with interactions to other group genes. For example, 56% DEGs have interactions with MNGs, and in contrast, only 27% MNGs have interactions with DEGs.

iii. As shown in Figure 5D, the numbers of overlapped genes among feature genes selected by different methods are small, which implies that these methods detect different global or local biomarkers on the biological network. Meanwhile, the feature genes from different methods have many potential interactions on the biological network (Figure 5E). Especially, 50% genes in MNGs can interact with frSVMs, and 100% genes in frSVMs can interact with MNGs, which means that MNG and frSVM would find biomarkers neighboring on the biological network. Together with above findings, MNGs show better ability on both discrimination and interpretation of biomarkers than DEGs, AEPs and frSVMs in this case study.

These results not only provided the biological insights into dynamical drug sensitivity and resistance but also demonstrated the superiority of big-data-based edge biomarkers for evaluating disease states in terms of both effectiveness and efficiency. The integration of high-confidence prior knowledge is expected to provide additional information on the discovery of edge biomarkers or big data biomarkers. And the future systematical evaluation of biomarkers in different application scenarios, including the study of drug sensitivity and resistance, will inspire more powerful models and technologies in biomarker discovery and application.

Key Points

- A key to achieve the precision medicine or personalized medicine is to characterize individual diseases by their systematical and dynamical features (e.g. network and fluctuation) rather than static features (e.g. sequence mutations or SNPs).
- Big-data-based edge biomarker is a new concept to characterize disease features based on biomedical big data in a dynamical and network manner, which also provides alternative strategies to indicate disease status in single samples. DNB (dynamical network biomarker) is such a biomarker to indicate the critical state during disease progression by using dynamical information.
- There are many sources and structures of biomedical big data accessible in public for edge biomarker and disease study. The biomedical big data are typically 'small-sample size in high-dimension space', i.e. small samples but with high dimensions on feature for each individual, in contrast to traditional big data in other fields, i.e. big samples but with low dimensions.
- In contrast to using the information of the common molecules or edges across a population in traditional

biomarkers including network and edge biomarkers, big-data-based edge biomarkers are further specific for each individual and thus can accurately evaluate the disease state by considering the individual heterogeneity, i.e. high-dimensional data are required not only in the learning process but also in the diagnosing or predicting process of the tested individual.

- As a representation of big-data-based edge biomarker, MNR (module network rewiring-analysis) is able to predict the disease state by learning differential associations between molecules rather than differential expressions of molecules during disease progression or treatment in individual patients. MNR makes full use of not only common molecules or edges (e.g. consistent modules) but also conditional molecules or edges (e.g. module interactions rather than molecule interactions). A deep case study shows that the identified module biomarkers from MNR can accurately distinguish vaccines with or without protection and outperformed over previous reported gene signatures in terms of effectiveness and efficiency.

Supplementary data

Supplementary data are available online at <http://bib.oxfordjournals.org/>.

Funding

This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (No. XDB13040700), the National Program on Key Basic Research Project (No. 2014CB910504), the National Natural Science Foundation of China (NSFC) (Nos. 61134013, 91439103, 31200987), the Knowledge Innovation Program of SIBS of CAS (2013KIP218) and JST's Super Highway.

References

- Auffray C, Charron D, Hood L. Predictive, preventive, personalized and participatory medicine: back to the future. *Genome Med* 2010;**2**:57.
- Hood L, Friend SH. Predictive, personalized, preventive, participatory (P4) cancer medicine. *Nat Rev Clin Oncol* 2011;**8**:184–7.
- Hood L, Flores M. A personal view on systems medicine and the emergence of proactive P4 medicine: predictive, preventive, personalized and participatory. *N Biotechnol* 2012;**29**:613–24.
- Highnam G, Mittelman D. Personal genomes and precision medicine. *Genome Biol* 2012;**13**:324.
- Jamal-Hanjani M, Hackshaw A, Ngai Y, et al. Tracking genomic cancer evolution for precision medicine: the lung TRACERx study. *PLoS Biol* 2014;**12**:e1001906.
- Van Allen EM, Wagle N, Stojanov P, et al. Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. *Nat Med* 2014;**20**:682–8.
- Stover DG, Wagle N. Precision medicine in breast cancer: genes, genomes, and the future of genomically driven treatments. *Curr Oncol Rep* 2015;**17**:438.
- Terry SF. Obama's Precision Medicine Initiative. *Genet Test Mol Biomarkers* 2015;**19**:113–14.
- McCarthy M. Obama seeks \$213m to fund "precision medicine". *BMJ* 2015;**350**:h587.
- Collins FS, Varmus H. A new initiative on precision medicine. *N Engl J Med* 2015;**372**:793–5.
- Obama 2016 budget calls for precision medicine. *Nat Biotechnol* 2015;**33**:216.
- UK catapults precision medicine. *Nat Biotechnol* 2015;**33**:119.
- Kaiser J. National Institutes of Health. A government niche for translational medicine and drug development. *Science* 2010;**330**:1462–3.
- Milne CP, Kaitin KI. Translational medicine: an engine of change for bringing new technology to community health. *Sci Transl Med* 2009;**1**:5cm5.
- Marko NF, Weil RJ. Mathematical modeling of molecular data in translational medicine: theoretical considerations. *Sci Transl Med* 2010;**2**:56rv54.
- Goodsaid FM, Mendrick DL. Translational medicine and the value of biomarker qualification. *Sci Transl Med* 2010;**2**:47ps44.
- Zhang J. Translational medicine in China. *Sci China Life Sci* 2012;**55**:834–6.
- Howells DW, Sena ES, Macleod MR. Bringing rigour to translational medicine. *Nat Rev Neurol* 2014;**10**:37–43.
- Bornstein SR, Licinio J. Improving the efficacy of translational medicine by optimally integrating health care, academia and industry. *Nat Med* 2011;**17**:1567–9.
- Gordon LB, Rothman FG, Lopez-Otin C, et al. Progeria: a paradigm for translational medicine. *Cell* 2014;**156**:400–7.
- Chen L, Liu R, Liu ZP, et al. Detecting early-warning signals for sudden deterioration of complex diseases by dynamical network biomarkers. *Sci Rep* 2012;**2**:342.
- Buxton B, Hayward V, Pearson I, et al. Big data: the next Google. Interview by Duncan Graham-Rowe. *Nature* 2008;**455**:8–9.
- Jiang B, Song K, Ren J, et al. Comparison of metagenomic samples using sequence signatures. *BMC Genomics* 2012;**13**:730.
- Feng H, Qin Z, Zhang X. Opportunities and methods for studying alternative splicing in cancer with RNA-Seq. *Cancer Lett* 2013;**340**:179–91.
- Engemann K, Enquist BJ, Sandel B, et al. Limited sampling hampers "big data" estimation of species richness in a tropical biodiversity hotspot. *Ecol Evol* 2015;**5**:807–20.
- Altman RB, Ashley EA. Using "big data" to dissect clinical heterogeneity. *Circulation* 2015;**131**:232–3.
- Li Y, Chen L. Big biological data: challenges and opportunities. *Genomics Proteomics Bioinformatics* 2014;**12**:187–9.
- Wang Y, Zhang XS, Chen L. Computational systems biology in the big data era. *BMC Syst Biol* 2013;**7** (Suppl 2):S1.
- Liu ZP, Zhang W, Horimoto K, et al. Gaussian graphical model for identifying significantly responsive regulatory networks from time course high-throughput data. *IET Syst Biol* 2013;**7**:143–52.
- Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet* 2004;**5**:101–13.
- Ideker T, Krogan NJ. Differential network biology. *Mol Syst Biol* 2012;**8**:565.
- Zeng T, Sun SY, Wang Y, et al. Network biomarkers reveal dysfunctional gene regulations during disease progression. *FEBS J* 2013;**280**:5682–95.
- Liu R, Wang X, Aihara K, et al. Early diagnosis of complex diseases by molecular biomarkers, network biomarkers, and dynamical network biomarkers. *Med Res Rev* 2014;**34**:455–78. doi:10.1002/med.21293.

34. Zeng T, Zhang W, Yu X, et al. Edge biomarkers for classification and prediction of phenotypes. *Sci China Life Sci* 2014;**57**:1103–14.
35. Sahni N, Yi S, Zhong Q, et al. Edgotype: a fundamental link between genotype and phenotype. *Curr Opin Genet Dev* 2013;**23**:649–57.
36. Zhang W, Zeng T, Chen L. EdgeMarker: identifying differentially correlated molecule pairs as edge-biomarkers. *J Theor Biol* 2014;**362**:35–43.
37. Sun SY, Liu ZP, Zeng T, et al. Spatio-temporal analysis of type 2 diabetes mellitus based on differential expression networks. *Sci Rep* 2013;**3**:2268.
38. Lai Y, Wu B, Chen L, et al. A statistical method for identifying differential gene-gene co-expression patterns. *Bioinformatics* 2004;**20**:3146–55.
39. Chen L. Systems biology with omics data. *Methods* 2014;**67**:267–8.
40. Chen R, Mias GI, Li-Pook-Than J, et al. Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* 2012;**148**:1293–307.
41. Zeng T, Wang DC, Wang X, et al. Prediction of dynamical drug sensitivity and resistance by module network rewiring-analysis based on transcriptional profiling. *Drug Resist Updat* 2014;**17**:64–76.
42. Zhang W, Zeng T, Liu X, et al. Diagnosing phenotypes of single-sample individuals by edge biomarkers. *J Mol Cell Biol* 2015;**7**:231–41.
43. Wang W, Baggerly KA, Knudsen S, et al. Independent validation of a model using cell line chemosensitivity to predict response to therapy. *J Natl Cancer Inst* 2013;**105**:1284–91.
44. Reinhold WC, Varma S, Rajapakse VN, et al. Using drug response data to identify molecular effectors, and molecular “omic” data to identify candidate drugs in cancer. *Hum Genet* 2015;**134**:3–11.
45. Kim HS, Minna JD, White MA. GWAS meets TCGA to illuminate mechanisms of cancer predisposition. *Cell* 2013;**152**:387–9.
46. Cline MS, Craft B, Swatloski T, et al. Exploring TCGA Pan-Cancer data at the UCSC Cancer Genomics Browser. *Sci Rep* 2013;**3**:2652.
47. Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res* 2013;**41**:D991–5.
48. Barrett T, Troup DB, Wilhite SE, et al. NCBI GEO: archive for functional genomics data sets—10 years on. *Nucleic Acids Res* 2011;**39**:D1005–10.
49. McWilliam H, Li W, Uludag M, et al. Analysis Tool Web Services from the EMBL-EBI. *Nucleic Acids Res* 2013;**41**:W597–600.
50. Goujon M, McWilliam H, Li W, et al. A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Res* 2010;**38**:W695–9.
51. Sneddon TP, Zhe XS, Edmunds SC, et al. GigaDB: promoting data dissemination and reproducibility. *Database (Oxford)* 2014;**2014**:bau018.
52. Sneddon TP, Li P, Edmunds SC. GigaDB: announcing the GigaScience database. *Gigascience* 2012;**1**:11.
53. Gerstein M. Genomics: ENCODE leads the way on big data. *Nature* 2012;**489**:208.
54. Liu C, Su J, Yang F, et al. Compound signature detection on LINCS L1000 big data. *Mol Biosyst* 2015;**11**:714–22.
55. Duan Q, Flynn C, Niepel M, et al. LINCS Canvas Browser: interactive web app to query, browse and interrogate LINCS L1000 gene expression signatures. *Nucleic Acids Res* 2014;**42**:W449–60.
56. Kanehisa M, Goto S, Sato Y, et al. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* 2012;**40**:D109–14.
57. Ogata H, Goto S, Sato K, et al. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 1999;**27**:29–34.
58. Szklarczyk D, Franceschini A, Kuhn M, et al. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res* 2011;**39**:D561–8.
59. Blake JA, Dolan M, Drabkin H, et al. Gene Ontology annotations and resources. *Nucleic Acids Res* 2013;**41**:D530–5.
60. Kellis M, Califano A, Bar-Joseph Z. Preface: RECOMB Conference on Systems Biology, Regulatory Genomics, and DREAM Challenges 2010 special issue. *J Comput Biol* 2011;**18**:131.
61. Altman RB. Predicting cancer drug response: advancing the DREAM. *Cancer Discov* 2015;**5**:237–8.
62. Feng T, Zhu X. Genome-wide searching of rare genetic variants in WTCCC data. *Hum Genet* 2010;**128**:269–80.
63. Barretina J, Caponigro G, Stransky N, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 2012;**483**:603–7.
64. Smoot ME, Ono K, Ruscheinski J, et al. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 2011;**27**:431–2.
65. Lopes CT, Franz M, Kazi F, et al. Cytoscape Web: an interactive web-based network browser. *Bioinformatics* 2010;**26**:2347–8.
66. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;**13**:2498–504.
67. Kramer A, Green J, Pollard J, Jr, et al. Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics* 2014;**30**:523–30.
68. Abecasis GR, Altshuler D, Auton A, et al. A map of human genome variation from population-scale sequencing. *Nature* 2010;**467**:1061–73.
69. Abecasis GR, Auton A, Brooks LD, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012;**491**:56–65.
70. The 100,000 Genomes Project. <http://www.genomicsengland.co.uk/the-100000-genomes-project/>
71. 100K Wellness Project. <http://research.systemsbio.net/100k/>
72. Hood L, Lovejoy JC, Price ND. Integrating big data and actionable health coaching to optimize wellness. *BMC Med* 2015;**13**:4.
73. Zeng T, Zhang CC, Zhang W, et al. Deciphering early development of complex diseases by progressive module network. *Methods* 2014;**67**:334–43.
74. Liu ZP, Wu H, Zhu J, et al. Systematic identification of transcriptional and post-transcriptional regulations in human respiratory epithelial cells during influenza A virus infection. *BMC Bioinformatics* 2014;**15**:336.
75. Bar-Joseph Z, Gitter A, Simon I. Studying and modelling dynamic biological processes using time-series gene expression data. *Nat Rev Genet* 2012;**13**:552–64.
76. Wang J, Sun Y, Zheng S, et al. APG: an Active Protein-Gene network model to quantify regulatory signals in complex biological systems. *Sci Rep* 2013;**3**:1097.
77. Huang Y, Zaas AK, Rao A, et al. Temporal dynamics of host molecular responses differentiate symptomatic and asymptomatic influenza a infection. *PLoS Genet* 2011;**7**:e1002234.

78. van 't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002;415:530–6.
79. Akavia UD, Litvin O, Kim J, et al. An integrated approach to uncover drivers of cancer. *Cell* 2010;143:1005–17.
80. Chuang HY, Lee E, Liu YT, et al. Network-based classification of breast cancer metastasis. *Mol Syst Biol* 2007;3:140.
81. Lee E, Chuang HY, Kim JW, et al. Inferring pathway activity toward precise disease classification. *PLoS Comput Biol* 2008;4:e1000217.
82. Qian L, Zheng H, Zhou H, et al. Classification of time series gene expression in clinical studies via integration of biological network. *PLoS One* 2013;8:e58383.
83. Li Y, Jin S, Lei L, et al. Deciphering deterioration mechanisms of complex diseases based on the construction of dynamic networks and systems analysis. *Sci Rep* 2015;5:9283.
84. Zhang S, Liu CC, Li W, et al. Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Res* 2012;40:9379–91.
85. De Vlaminck I, Khush KK, Strehl C, et al. Temporal response of the human virome to immunosuppression and antiviral therapy. *Cell* 2013;155:1178–87.
86. Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics* 2007;23:2507–17.
87. Larranaga P, Calvo B, Santana R, et al. Machine learning in bioinformatics. *Brief Bioinform* 2006;7:86–112.
88. Montero J, Sarosiek KA, DeAngelo JD, et al. Drug-induced death signaling strategy rapidly predicts cancer response to chemotherapy. *Cell* 2015;160:977–89.
89. Jung Y, Abdel-Fatah TM, Chan SY, et al. SHON is a novel estrogen-regulated oncogene in mammary carcinoma that predicts patient response to endocrine therapy. *Cancer Res* 2013;73:6951–62.
90. Masica DL, Karchin R. Collections of simultaneously altered genes as biomarkers of cancer cell drug response. *Cancer Res* 2013;73:1699–708.
91. Warsow G, Struckmann S, Kerkhoff C, et al. Differential network analysis applied to preoperative breast cancer chemotherapy response. *PLoS One* 2013;8:e81784.
92. Zhao J, Zhang XS, Zhang S. Predicting cooperative drug effects through the quantitative cellular profiling of response to individual drugs. *CPT Pharmacometrics Syst Pharmacol* 2014;3:e102.
93. Iida N, Dzutsev A, Stewart CA, et al. Commensal bacteria control cancer response to therapy by modulating the tumor microenvironment. *Science* 2013;342:967–70.
94. Walsh AJ, Cook RS, Sanders ME, et al. Quantitative optical imaging of primary tumor organoid metabolism predicts drug response in breast cancer. *Cancer Res* 2014;74:5184–94.
95. Geeleher P, Cox NJ, Huang RS. Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines. *Genome Biol* 2014;15:R47.
96. Bhang HE, Ruddy DA, Krishnamurthy Radhakrishna V, et al. Studying clonal dynamics in response to cancer therapy using high-complexity barcoding. *Nat Med* 2015;21:440–8.
97. Li L, Fridley BL, Kalari K, et al. Discovery of genetic biomarkers contributing to variation in drug response of cytidine analogues using human lymphoblastoid cell lines. *BMC Genomics* 2014;15:93.
98. Chen CH, Lee CS, Lee MT, et al. Variant GADL1 and response to lithium therapy in bipolar I disorder. *N Engl J Med* 2014;370:119–28.
99. Cui J, Stahl EA, Saevarsdottir S, et al. Genome-wide association study and gene expression analysis identifies CD84 as a predictor of response to etanercept therapy in rheumatoid arthritis. *PLoS Genet* 2013;9:e1003394.
100. Kanai K, Kako M, Okamoto H. HCV genotypes in chronic hepatitis C and response to interferon. *Lancet* 1992;339:1543.
101. Sarashina T, Sakurai T, Watanabe Y, et al. Nucleotide sequence of the hepatitis C virus genome from a patient negative for anti-HCV by the first generation antibody assay. *Nucleic Acids Res* 1993;21:1037.
102. Lopez-Labrador FX, Ampurdanes S, Gimenez-Barcons M, et al. Relationship of the genomic complexity of hepatitis C virus with liver disease severity and response to interferon in patients with chronic HCV genotype 1b infection [correction of interferon]. *Hepatology* 1999;29:897–903.
103. Schweitzer CJ, Liang TJ. Impact of host and virus genome variability on HCV replication and response to interferon. *Curr Opin Virol* 2013;3:501–7.
104. Jörnsten R, Wang HY, Welsh WJ, et al. DNA microarray data imputation and significance analysis of differential expression. *Bioinformatics* 2005;21:4155–61.
105. Rapaport F, Khanin R, Liang Y, et al. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol* 2013;14:R95.
106. Villoslada P, Baranzini S. Data integration and systems biology approaches for biomarker discovery: challenges and opportunities for multiple sclerosis. *J Neuroimmunol* 2012;248:58–65.
107. Baranzini SE, Galwey NW, Wang J, et al. Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. *Hum Mol Genet* 2009;18:2078–90.
108. Yu X, Li G, Chen L. Prediction and early diagnosis of complex diseases by edge-network. *Bioinformatics* 2014;30:852–9.
109. Liu R, Yu X, Liu X, et al. Identifying critical transitions of complex diseases based on a single sample. *Bioinformatics* 2014;30:1579–86.
110. Taylor MW, Tsukahara T, McClintick JN, et al. Cyclic changes in gene expression induced by Peg-interferon alfa-2b plus ribavirin in peripheral blood monocytes (PBMC) of hepatitis C patients during the first 10 weeks of treatment. *J Transl Med* 2008;6:66.
111. Huang T, Tu K, Shyr Y, et al. The prediction of interferon treatment effects based on time series microarray gene expression profiles. *J Transl Med* 2008;6:44.
112. Baranzini SE, Mousavi P, Rio J, et al. Transcription-based prediction of response to IFNbeta using supervised computational methods. *PLoS Biol* 2005;3:e2.
113. Zeng T, Guo X, Liu J. Negative correlation based gene markers identification in integrative gene expression data. *Int J Data Min Bioinform* 2014;10:1–17.
114. Lin TH, Kaminski N, Bar-Joseph Z. Alignment and classification of time series gene expression in clinical studies. *Bioinformatics* 2008;24:i147–55.
115. Vahey MT, Wang Z, Kester KE, et al. Expression of genes associated with immunoproteasome processing of major histocompatibility complex peptides is indicative of protection with adjuvanted RTS,S malaria vaccine. *J Infect Dis* 2010;201:580–9.
116. Hadley TJ, Peiper SC. From malaria to chemokine receptor: the emerging physiologic role of the Duffy blood group antigen. *Blood* 1997;89:3077–91.

117. Maisonneuve C, Bertholet S, Philpott DJ, et al. Unleashing the potential of NOD- and Toll-like agonists as vaccine adjuvants. *Proc Natl Acad Sci USA* 2014;**111**:12294–9.
118. Nardin EH, Oliveira GA, Calvo-Calle JM, et al. Phase I testing of a malaria vaccine composed of hepatitis B virus core particles expressing *Plasmodium falciparum* circumsporozoite epitopes. *Infect Immun* 2004;**72**:6519–27.
119. Sallberg M, Hughes J, Jones J, et al. A malaria vaccine candidate based on a hepatitis B virus core platform. *Intervirology* 2002;**45**:350–61.
120. Proietti FA, Carneiro-Proietti AB, Catalan-Soares BC, et al. Global epidemiology of HTLV-I infection and associated diseases. *Oncogene* 2005;**24**:6058–68.
121. Lehrer S. Association between malaria incidence and all cancer mortality in fifty U.S. States and the District of Columbia. *Anticancer Res* 2010;**30**:1371–3.
122. Guo Z, Zhang T, Li X, et al. Towards precise classification of cancers based on robust gene functional expression profiles. *BMC Bioinformatics* 2005;**6**:58.
123. Cun Y, Frohlich H. netClass: an R-package for network based, integrative biomarker signature discovery. *Bioinformatics* 2014;**30**:1325–6.
124. Winter C, Kristiansen G, Kersting S, et al. Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes. *PLoS Comput Biol* 2012;**8**:e1002511.
125. Hanzelmann S, Castelo R, Guinney J. GSEA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* 2013;**14**:7.
126. Drier Y, Sheffer M, Domany E. Pathway-based personalized analysis of cancer. *Proc Natl Acad Sci USA* 2013;**110**:6388–93.