Approaches to uncovering cancer diagnostic and prognostic molecular signatures

Shengjun Hong†, Yi Huang†, Yaqiang Cao†, Xingwei Chen†, and Jing-Dong J Han*
Chinese Academy of Sciences Key Laboratory of Computational Biology; Chinese Academy of Sciences-Max Planck Partner Institute for Computational Biology; Shanghai Institutes for Biological Sciences; Chinese Academy of Sciences; Shanghai, China

†These authors contributed equally to this work.

Keywords: Bayesian network, cancer diagnosis, cancer prognosis, molecular signature, subtype

Abbreviations: BIC, Bayesian information criterion; BN, Bayesian network; CNV, copy number variation; COPA, cancer outlier profile analysis; DEG, differentially expressed genes; DMV, DNA methylation valley; GSEA, gene set enrichment analysis; GO, gene ontology; GWAS, genome-wide association studies; IncRNA, long non-coding RNA; PAGE, parametric analysis of gene set enrichment; PheWAS, phenome-wide association studies; SOM, self-organizing map; SNP, single-nucleotide polymorphism; SVM-RFE, Support Vector Machine-Recursive Feature Elimination; TCGA, The Cancer Genome Atlas; TCIA, The Cancer Imaging Archive.

The recent rapid development of high-throughput technology enables the study of molecular signatures for cancer diagnosis and prognosis at multiple levels, from genomic and epigenomic to transcriptomic. These unbiased large-scale scans provide important insights into the detection of cancer-related signatures. In addition to single-layer signatures, such as gene expression and somatic mutations, integrating data from multiple heterogeneous platforms using a systematic approach has been proven to be particularly effective for the identification of classification markers. This approach not only helps to uncover essential driver genes and pathways in the cancer network that are responsible for the mechanisms of cancer development, but will also lead us closer to the ultimate goal of personalized cancer therapy.

Introduction

Cancer is a major human health problem worldwide and is related to one-fourth of deaths in the United States.1 Decades of cancer research have revealed many specific details, as well as some general features shared among different cancers. Although the rate of cancer deaths in the United States has declined in recent decades,1,2 the rate of acquisition of cancer is continuously increasing.1,3 Early detection of cancer diagnosis signatures decreases both morbidity and mortality. Moreover, studying the signatures associated with cancer prognosis (quantified by 5-year survival time in clinical trials) not only helps to predict patient outcome, but also holds a key to understanding the genetic mechanisms of cancer development.4,5

With the development of high-throughput technology, signatures existing at multiple levels have been identified for cancer diagnosis and prognosis, including genomic, epigenomic, and transcriptomic signatures. For example, genome-wide single nucleotide polymorphism (SNP) profiling and array-based comparative genomic hybridization have been applied to identify germline and somatic lesions in several cancers. Additionally, hundreds of SNPs or haplotypes have been reported to be significantly associated with cancers. In a study of 1,599 cases and 11,546 controls, Stacey et al.6 found that rs3803662, which is associated with the TOX3 gene, is also significantly associated with breast cancer. The DNA methylome is also under intense study, providing global pictures of epigenetic changes in cancers.7 Transcriptome analyses have successfully demonstrated that the expression of multiple genes, rather than single genes, can serve as effective subtype or prognosis classifiers for many cancers, such as leukemia8 and breast cancer.9

Although different layers of genome-wide analysis have revealed global features of cancers, integration of multilayer information facilitates more accurate cancer subtyping and more comprehensive mechanistic insights. Within such a panorama, the systematic approach has led to identification of the hallmarks of cancers.10

In this review, we aim to provide an insight into the data and methods available for systems level analysis of cancer subtypes and their characterization. We have organized the
review into 3 parts: first, we introduce general features and web-based resources for molecular signatures of cancer diagnosis and prognosis; second, we summarize existing methods for detecting such signatures; and, finally, we discuss potential methods for interpreting these signatures, such as network and module analysis.

Cancer-related high-throughput data types and web resources

Although it is feasible to collect raw cancer-related high-throughput data, such as from GEO\textsuperscript{11} and ArrayExpress,\textsuperscript{12} several databases and web services provide rich cancer-related data in a curated or integrated manner.

The Cancer Genome Atlas (TCGA) project has generated a myriad of cancer "omic" data. To date, more than 8,913 tumor samples across 30 types of cancer have been collected and sequenced. The TCGA provides raw and processed data covering layers of genome, epigenome, and transcriptome data, together with clinical information. The recently established cBioPortal\textsuperscript{13} provides not only downloadable large-scale cancer genomic data, but also online visualization and analysis services for TCGA datasets.

In addition to these comprehensive resources, there are several databases focusing on 1 or 2 specific areas. For example, COSMIC\textsuperscript{14} stores somatic mutations. Its latest version presents a cancer mutation landscape of 132 known cancer genes and 208 fusion gene pairs, based on nearly 8,000 cancer genomes. With a convenient interface, COSMICMart\textsuperscript{15} helps to filter COSMIC data sets into categories. Oncomine\textsuperscript{16} is a database for target identification and validation, drug development, and clinical research. Oncotator (http://www.broadinstitute.org/oncotator/) provides annotation for cancer genes, mutations, and amplification or deletion regions. Tumorscape\textsuperscript{17} provides both a portal to query copy number alterations across multiple cancer types, and a web interface visualizing the results based on the GISTIC\textsuperscript{18} algorithm. IntOgen\textsuperscript{19} integrates somatic mutations, copy number changes, and expression in cancer into 3 query-and-download modules, in addition to providing an interface to TCGA.

These valuable resources have facilitated various efforts in cancer studies and have broadened our perspectives of cancer (Table 1). At the present time most resources are based on isolated samples but we expect that there will be an increase in replicated data for advanced analyses, such as multiple (and even time series) samples from the same patient or from a homogeneous population.

Over the past few years, multiple types of signatures have been reported to associate with cancer diagnosis and prognosis. Most of these studies focused on common somatic mutations,\textsuperscript{20} mRNA\textsuperscript{21,22} and microRNA expression,\textsuperscript{23,24} and protein level changes.\textsuperscript{25,26} With the popularization of high-throughput sequencing technologies and the refinement of bioinformatics pipelines, these features have helped to identify credible markers.

In addition to general genomic and transcriptomic features, various other features could be used to improve the confidence. For example, long non-coding RNA (IncRNA) is an emerging new paradigm in cancer research that can be either oncogenic or tumor suppressive, indicating its possible application for diagnostics and prognosis. One typical example for practical diagnostics is PCA3,\textsuperscript{27} which is widely used in urine testing to determine prostate cancer risk. HOTAIR, which has a chromatin-remodeling effect, serves as an oncogenic biomarker and has been validated in a variety of cancers, such as lung cancer\textsuperscript{28} and liver cancer.\textsuperscript{29} MEG3 acts as a tumor suppressor that is frequently downregulated in pituitary cancer\textsuperscript{30} and glioma.\textsuperscript{31} IncRNAs have advantages over protein-coding RNAs in cancer diagnosis and prognosis because of their expression specificity and direct molecular function.

The methylome, or DNA methylation status on genome-wide CpG sites, has been intensively studied in developmental biology. However, despite the fact that cancer shares key properties with development, albeit inversely, the methylome was only recently applied to cancer diagnosis or prognosis. It has been reported that GSTP1 gains hypermethylation in prostate cancer, indicating its role as a diagnostic marker.\textsuperscript{32} A recent study showed that DNA methylation valleys (DMVs) in stem cells are hypermethylated in cancer\textsuperscript{33} and therefore provide novel aberrant signatures. Nearly 8,000 cancer methylomes are available through public databases,\textsuperscript{34} facilitating future studies to reveal the specific DNA methylation signatures that drive carcinogenesis.

Clinical features are useful tools for diagnosis and prognosis. In particular, imaging has been widely applied in cancer diagnosis using various systems\textsuperscript{35} such as X-ray, computed tomography scan, magnetic resonance imaging, tumor biopsy, and endoscopic examination. The Cancer Imaging Archive (TCIA)\textsuperscript{36} contains medical images of both the National Lung Screening Trial

| Web Resource          | Raw data | Preprocessed data | Features                                              | Clinical information |
|-----------------------|----------|------------------|-------------------------------------------------------|----------------------|
| TCGA and cBioPortal   | Yes      | Yes              | Genomic, transcriptomic, epigenomic, proteomic        | Yes                  |
| COSMIC & COSMICMart   | No       | Yes              | Mutation                                              | No                   |
| Oncomine              | Yes      | Yes              | Microarray-based gene expression and copy number variation | Yes                  |
| Oncotator             | No       | Yes              | Gene, mutation, cancer amplification, and deletion region | No                   |
| Tumorscape            | No       | Yes              | Copy number variation                                 | No                   |
| TCIA                  | Yes      | Yes              | Medical images                                        | Yes                  |

Table 1. Summary of cancer data resources
project\textsuperscript{37} and the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial project.\textsuperscript{38} These clinical features reflect obvious biological outcomes and should also assist molecular signature identification.

**Approaches**

As resources and signature types of cancers keep emerging, so do various bioinformatics methods for analyzing such data. We will introduce the approaches that are most frequently used for either associative or mechanistic inferences.

**Associative inference**

Cancer diagnosis and prognosis have benefitted from the application of multiple molecular markers. However, given the heterogeneous nature of cancers, prognosis-related subtype classification is more, or at least equally, important for personalized treatment.\textsuperscript{39} Methods for subtype detection and classification can be generally grouped into supervised and unsupervised approaches, by which subtypes are characterized by certain signatures (Fig. 1). When differentiating clinically well-defined subtypes, supervised approaches are often used. However, for identification of unknown subtypes or when classifying clinically less well-defined subtypes, unsupervised methods are required.\textsuperscript{40}

**Supervised learning of molecular signatures**

Supervised learning is quite straightforward when the samples are well categorized, for example between cancer samples and controls. In this approach, known subtype-related signals are extracted from noise or confounding factors.

Genetic variation in the human genome provides an abundant resource for cancer research. The study of genetic variation can reveal susceptibility to cancer associated with distinct variations.\textsuperscript{41} Genome-wide association studies (GWAS) have identified many genetic risk factors for different common human cancers, which are available in the NHGRI GWAS catalog.\textsuperscript{42} The most frequently used statistics in GWAS are the Chi-square test and Fisher’s exact test, although the Chi-square test is not suitable when the sample size is small (fewer than 10 samples). As the Pearson Chi square test cannot handle a variable with more than 2 categories, the Cochran-Armitage test for trend may be applied in such cases. All of these statistics tests are integrated in PLINK,\textsuperscript{43} a popular tool for GWAS. Not all SNPs can be accurately genotyped, and SNPTEST\textsuperscript{44} uses Frequentist and Bayesian tests to solve this problem. As GWAS datasets generally involve more than 100,000 SNPs, multiple testing correction is needed. The commonly used correction method is the Bonferroni correction but this may be too strong in some cases, for which the Holm–Bonferroni method might be more suitable. There is no generic bridge that links SNPs to cancers; only when integrated with functional data can a cancer-related SNP be deemed to be responsible for the cancer process.\textsuperscript{45} Phenome-wide association studies (PhenWAS)\textsuperscript{46} can be viewed as a variant of GWAS to investigate the associations between SNPs and phenotypes. This method, especially when accompanied by imaging, is a promising approach to explore cancer genome–phenome associations.

Unlike genomic data, the most general approach to transcriptome data is to identify differentially expressed genes (DEGs). The Student’s t test is frequently used for analysis of DEGs. The t statistic assumes homogeneity of examined samples; however, this does not always hold true in cancer samples. For example, an unstable genome may lead to the same translocations (and subsequent abnormal expression) in some, but not all, cancer samples, and not in the control samples.\textsuperscript{47} To solve this problem, methods that are sensitive to “outliers” have been developed. Cancer outlier profile analysis (COPA) was proposed and applied to prostate cancer,\textsuperscript{47,48} and then implemented as a part of the abovementioned Oncomine. In addition to COPA, several statistics have been sequentially proposed, including OS,\textsuperscript{49} ORT,\textsuperscript{50} MOST,\textsuperscript{51} and GTI.\textsuperscript{52} These statistics may work differently depending on the type of data and thus should be carefully compared.\textsuperscript{53}

Alternatively, sometimes the data may itself contain homogeneity (e.g., after proper subtyping), or researchers may be

---

**Figure 1.** Overview of strategies to detect cancer subtypes related to cancer diagnosis and prognosis.
interested in only the signatures with high penetrance. In these cases, in addition to the Student’s t test there are many well-developed tools that could be applied, such as SAM, limma, edgeR, DESeq, and Cuffdiff. If researchers are unsure about the homogeneity, machine-learning approaches may be more robust. For example, we have applied Support Vector Machine-Recursive Feature Elimination (SVM-RFE) to identify markers of pediatric acute lymphoblastic leukemia.

When data come from multiple resources, either generated in different batches or compared across multiple layers, it is essential to properly preprocess the pooled data and extract signals from noise and confounding factors using so-called “meta-analysis”. Based on the extent of noise reduction, preprocessing methods can be divided into 3 categories, as follows: (1) Moderate approaches smooth data with different relative intensities either by z-score normalization or quantile normalization. Z-score normalization requires the original data to have an approximately Gaussian distribution whereas quantile normalization is non-parametric and especially efficient in microarray data analysis (implemented in RMA). Although both work on sample noise reduction, z-score normalization could also be applied to features, transforming expression intensities to expression patterns. This is often a necessary preprocessing step to integrate time-series or multilayer data. (2) Known confounding factors, typically batch effects, can be reduced after being incorporated into the null model using a generalized linear model such as in DESeq, or an empirical Bayes method such as in Combat. (3) It is also possible to further exclude unknown confounding factors, as implemented in SVA and ISVA.

Unsupervised learning of molecular signatures

Unsupervised methods for subtyping may be more intrinsic and more robust for experimental designs. Clustering is widely used for omics data, to divide features or samples into subgroups. Since the first application of hierarchical clustering in microarray gene expression datasets, the approach has rapidly been applied to cancer gene expression datasets. K-means clustering and Self-Organizing Maps (SOMs) are often applied in similar ways. At the present time hierarchical clustering remains the typical choice; however, it is challenged by multilayer data. Efforts have therefore been made to improve clustering for integrative analysis, for example classic hierarchical clustering of the correlation matrix between mRNA and microRNA expression and biclustering on the correlation matrix between mRNA expression and DNA copy number. More sophisticated tools were also developed, such as iCluster and its extended version iClusterPlus, PSDF, MDI, JIVE, and SNF. iCluster is intensely used with TCGA to discover subtypes with distinct clinical outcomes, whereas iClusterPlus is able to handle both discrete (somatic mutations) and continuous variables. Most cluster methods cannot automatically determine the optimal number of sample and feature clusters. We proposed an adaptive clustering algorithm incorporating the Bayesian information criterion (BIC) and an unsupervised “Super k-means” method. With this approach, we detected 7 subtypes for ovarian cancer with significantly different clinical outcomes based on the combination of mRNA and miRNA expression, DNA methylation, and copy number variation. Accordingly, we also developed a useful framework to detect modular signatures for prognosis.

Mechanism inference

As mentioned above, there are various features and resources that provide multiple dimensions of signatures for cancer diagnosis and prognosis. However, association only provides a pool of molecules, including false positives, unimportant candidates that should be ignored, and false negatives. Moreover, cancer arises from a complex origin composed of genomic, transcriptomic, and epigenomic variations. Disease-specific genes do not function independently; instead, they usually act as a network module that associates with a certain biological function. In several cases, a single type of molecular signature cannot uncover the molecular mechanism of cancer prognosis to predict clinical outcome. In order to either accurately diagnose disease at an early stage or pursue the cause of prognosis, careful refinement of these candidate signatures is required, especially approaches based on networks with modularity analysis or causality inference.

Modularity analysis

Modularity is an important feature of networks. A network module is a context-coherent sub-network with conditionally comparable temporal and spatial profiles, and ideally with defined inputs and outputs. In practice, modules are often described without the exact inputs and outputs defined (but instead assumed) and evaluated by relative functional homogeneity within the module. Based on the static structure of a network, a module can be defined and dissected as a sub-network whose nodes are more densely connected within the sub-network than toward the outside, or that has more than random expectation as measured by the modularity metric or the clustering coefficient.

The cancer signaling network, for example, can be topologically divided into 12 blocks or modules. Edge type consistency has been used to find epistatic modules, although gene expression profile similarity and dissimilarity are most frequently used to define dynamic modules that are active under a certain biological context.

With no exception, classification markers are modularly organized and reflect dysregulation or driver mutations through network analysis. An unbiased search for markers of pediatric acute lymphoblastic leukemia subtype classification yielded a group of 62 genes that segregate into several network modules for different subtypes and are potentially controlled by subtype-specific transcriptional regulators. Such modularity in cancer marker genes has been used to identify network-based classifiers that have been demonstrated to outperform classical single gene or non-network module-based classifiers. Compared with gene expression networks, markers identified as sub-networks from protein-protein interaction networks are more reproducible and achieve significantly higher accuracy for classifying metastatic versus non-metastatic breast tumors. Martin et al. have found that high-
quality breast cancer prognosis markers can only be identified within subtypes, and that combinations of various markers can optimize the performance of the marker gene set. Most surprisingly, they found that each marker gene signature forms a network module, within which the marker genes interact intensively with genes that are frequently mutated in breast cancers, although the marker genes themselves are mostly not mutated. Moreover, the mutated interacting genes in the modules can also distinguish metastatic versus non-metastatic samples, implying that these might be driver mutations within each module. Therefore, the modularity of disease-associated genes in molecular interaction networks allows prediction of new disease-associated genes through their direct or indirect interactions with known disease-associated genes.

Uncovering the cancer regulatory network without a predefined reference set will reduce study bias and facilitate more objective analysis of all potential features involved in the network properties. To this end, Andreas Califano’s group has developed an algorithm called “ARACNe” which can ab initio infer regulatory interactions based on mutual information between 2 genes across a set of measurements. In particular, they have successfully applied the algorithm to predict transcriptional interactions specific for high-grade glioma. In combination with searches for transcription factors whose targets overlap significantly with genes that are overexpressed in mesenchymal cells, they narrowed down a key regulatory module. Also using an approach based on mutual information, Mani et al. developed a new algorithm that can identify dysregulated interactions in B-cell lymphoma using a Bayesian analysis that predicted a B-cell specific interactome as the backbone. The dysregulation is defined as loss/gain of a correlation in gene expression comparing lymphoma versus normal B cells. Compared with candidate gene-based reverse engineering approaches, such de novo network reverse engineering can identify not only dysregulated interactions, but also coherently dysregulated network modules that arise from the interactions, in an unbiased manner.

Most of the abovementioned methods of network modularity analysis have been implemented in Cytoscape and its rich plugin apps. Cytoscape is an expanding computational platform for the integration, visualization, statistical modeling, and annotation of biological networks. The apps are well organized and categorized within http://apps.cytoscape.org/, which makes it convenient for users to access and use to compile an analysis pipeline. After modularity inference the general network modules are always compared to specific biology “pathways,” such as widely used Gene Ontology (GO) terms and KEGG pathways. This comparison is termed enrichment analysis, and often uses Fisher’s exact test or hypergeometric test to draw statistical significance for the selected enriched terms. For GO analysis only, AmiGO and BiNGO implemented in Cytoscape are good choices. DAVID is highly recommended for multisource integration. The IntOGen web application allows evaluation of the contribution of biological modules such as KEGG pathways to a cancer by testing the significance of overlap between genes that are changed in the cancer and genes in a defined module.

More general-purpose applications such as Gene set enrichment analysis (GSEA) or its modified version Parametric analysis of gene set enrichment (PAGE) can also reveal whether a pathway, module, or signature set is significantly changed based on the rank or average expression intensity of genes within a gene set.

Bayesian networks and causality inference

Compared with mutual information or correlation-based methods, Bayesian network (BN) inference as a network reverse engineering approach has higher theoretical consistency, is able to distinguish direct and indirect interactions, and can identify both strong and weak and linear and non-linear dependencies as well as potential causal relationships. BN is a network or graph representation of the joint probability distribution over a set of variables (nodes) or conditional dependencies between variables. The BN structural learning algorithm searches for the network structure that has the best fit of joint probability distribution to the data using a scoring function such as the BIC. BIC contains 2 terms: one to evaluate the likelihood that the data are generated by the model, and another to penalize the complexity of the model. Recently, BN has been used for the diagnosis and prognosis of several cancer types including breast cancer and lung cancer. Olivier et al. applied BN to integrate clinical and microarray data. Evaluation of the performance of BN showed that this method performed well in predicting prognosis of breast cancer patients.

One restriction of BN learning is that the graph must be acyclic; that is, no loops are allowed even though they truly exist. Such feedback relationships can sometimes be resolved by additional temporal information, for example by the so-called “dynamic BN” approach. Potential causal relationships can also be identified from the consistently directed edges (irreversible edges) within the whole set of equivalent BN structures. Data from gene perturbation experiments can provide more direct evidence for inferring causal relationships. For example, a directed signaling network of 11 molecules can be reverse engineered by BN learning on thousands of single-cell flow cytometry measurements of the level of the molecules in human primary T cells after gene perturbations in the network.

The requirement for a large number of data points for BN inference has been a limiting factor in directly inferring gene regulatory networks from gene expression measurements. The recent rapid accumulation of microarray and deep sequencing data has made such approaches more practical. In pursuing key signatures, “early” changes may arise from system instability and thus have low penetrance, whereas a few function-related “late” changes are causal to cancer development. Therefore, the “intermediate” mediators and potential causality inferred by BNs will facilitate both early diagnoses and accurate prognoses when the core of the network is found to be affected by
The approaches discussed here are summarized in Table 2.

### Discussion and Perspectives

With the huge amount of high-throughput data that is already available or is being generated at an accelerated rate for different layers of the cancer molecular interaction network, obtaining a global picture of the full cancer molecular network for each cancer type, or even each individual tumor, will be completely feasible in the near future. The genomic, transcriptomic, epigenomic, and even proteomic and metabolomic changes in various cancers can be viewed as heterogeneous molecular phenotypes of the cancer cells. Many of these changes might be by-products resulting from genome instability or transcriptional and metabolic dysregulation and thus reflect the state of the underlying molecular and metabolic networks. Among these changes, some are necessary for the cancer cells to overcome multiple checkpoints and surveillance mechanisms and expand through clonal selection and expansion, and therefore ultimately enable invasive growth.10

There are 2 key points regarding cancer diagnosis and prognosis that should be addressed in the near future: (1) How to sift through numerous multilayer changes within the molecular network of cancer and find critical steps driving the cancer development and metastasis; and (2) How to find essential controllers, better interpret the hallmarks of cancer, and design successful treatment strategies. Toward these goals, both experimental and computational approaches should be

---

**Table 2. Methods for the detection of molecular signatures of cancer diagnosis and prognosis**

| Approach                                      | Summary                                                                 |
|-----------------------------------------------|-------------------------------------------------------------------------|
| **Associative inference—Supervised learning** |                                                                         |
| GWAS                                          |                                                                        |
| PLINK                                         | An open-source whole genome association tool, including statistics such  |
|                                               | as Chi-square test, Cochran-Armitage test, and Fisher’s exact test.    |
| SNPTES                                        | Incorporates imputation methods for genotype association test           |
|                                               |                                                                        |
| PheWAS                                        | Investigates the association between SNP and phenotypes                |
| DEGs                                          | Identifies DEGs, assuming homogeneity of examined samples               |
|                                               | Identifies DEGs, robust to heterogeneity of examined samples            |
| **Noise reduction**                            |                                                                        |
| Z-score normalization                         | Preprocess expression data with relative intensities                   |
| Quantile normalization                        |                                                                        |
| Combat                                        | Handles known confounding factors such as batch effects                |
| SVA and ISVA                                  | Excludes unknown confounding factors                                   |
| **Associative inference—Unsupervised learning**|                                                                        |
| Clustering analysis                           |                                                                        |
| Hierarchical clustering, Kmeans clustering,  | Partitions features or samples into subgroups                          |
| SOM                                           |                                                                        |
| Biclustering, iCluster, iClusterPlus,         | Discovers subtypes with clinical outcomes, integrating multiple types  |
| PSDF, MDI, JIVE, SNF, Super k-means           | of data                                                               |
| **Mechanism inference—Modularity analysis**   |                                                                        |
| IntOGen                                       | Evaluates the contribution of biological modules to a cancer           |
| DAVID, GSEA, PAGE                             | Reveals whether a cancer-related module is significantly enriched for a |
|                                               | known pathway                                                         |
| ARACNe                                        | Infers regulatory interactions based on mutual information between genes|
| Cytoscape                                     | Network and modularity analysis                                       |
| **Mechanism inference—Bayesian network and causality inference** |                                     |
| De novo network                               | Infers a network by detecting potential causal relationships between genes |
| Bayesian network                              | Allows feedback relationship compared to regular Bayesian network      |
investigated to annotate the multicancer layer network. By taking advantage of the ever-increasing rate and reduced cost of accumulating data, more efforts should be made to achieve the ultimate goal of personal cancer genomics and individualized cancer treatment.

Recent research based on TCGA projects, such as the Pan Cancer Project, have started to integrate cancer types and offer a comprehensive set of cancer systems biology data and new tools for cancer genomics and bioinformatics analysis. In addition to classical classification of different tumors, this will help to repurpose targeted therapies for cancers under the direction of their molecular pathologies.

References

1. Siegel R, Naishadham D, Jemal A. Cancer statistics, 2013. CA Cancer J Clin 2013; 63:11-30; PMID: 23353087
2. Edwards BK, Noone AM, Mariotto AB, Simard EP, Boncace FMP, Henley SJ, Jemal A, Cho H, Anderson RN, Kohler BA, et al. Annual Report to the Nation on the status of cancer, 1975–2010, featuring prevalence of comorbidity and impact on survival among persons with lung, colorectal, breast, or prostate cancer. Cancer 2014; 120:1290-314; PMID:24343171; http://dx.doi.org/10.1002/cncr.28509
3. Smith BD, Smith GL, Hurria A, Hortobagyi GN, Black RM, Gershenwald JE, Antman KH, et al. Genetic pathologies. Nature 2008; 454:841-3; PMID:18585097; http://dx.doi.org/10.1038/nature07358
4. Duque M, Modlin IM, Gupta A, Saif MW. Bio-markers in neuroendocrine tumors. JOP J Pancreas 2013; 14:372-6; PMID:23846930
5. Chibon F. Cancer gene expression signatures—the rise and fall! Eur J Cancer 2013; 49:2000-9; PMID:23498875; http://dx.doi.org/10.1016/j.ejca.2013.02.021
6. Stacey SN, Manouselis A, Sulem P, Rafnar T, Gudmundsson J, Gudjonsson SA, Mason G, Jakobsson M, Thorlacius S, Helgason A, et al. Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. Nat Genet 2007; 39:85-8; PMID:17529974; http://dx.doi.org/10.1038/ng.2064
7. Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration across human cancers. Nature Genet 2010; 42:543-9; PMID:20164920; http://dx.doi.org/10.1038/ng.1004
8. Merkel CH, Schumacher SE, Hill B, Meyerson ML, Beroudh P, Gerst G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol 2011; 12:R41; PMID:21527027; http://dx.doi.org/10.1186/gb-2011-12-4-r41
9. Zhou Y, Zhang X, Velbanskis A. MEG3 noncoding RNA: a tumor suppressor. J Mol Endocrinol 2012; 48:R45-53; PMID:22393162; http://dx.doi.org/10.1530/JME-12-0008
10. Wang P, Ren Z, Sun P. Overexpression of the long non-coding RNA MEG3 impairs in vitro glioma cell proliferation. J Cell Biochem 2012; 113:886-74; PMID:22366474; http://dx.doi.org/10.1002/jcb.24055
11. Heyn H, Esteller M. DNA methylation profiling in the clinic: applications and challenges. Nat Rev Genet 2012; 13:679-92; PMID:22945394; http://dx.doi.org/10.1038/nrg3270
12. Xie W, Schultz MD, Lister R, Hou Z, Rajagopal N, Ray P, Whittaker JW, Tian S, Hawkins RD, Leung D. Epigenomic analysis of multilocus differentiation of human embryonic stem cells. Cell 2013; 153:1134-48; PMID:23664764; http://dx.doi.org/10.1016/j.cell.2013.04.022
13. Stutzaker C, Taberly PC, Statham AL, Clark SJ. Mining cancer methylomes: prospects and challenges.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Funding

This work was supported by grants from the China Natural National Science Foundation (Grant #91329302, 3210103916, and 91019019), Chinese Ministry of Science and Technology (Grant #2011CB504206), Chinese Academy of Sciences (Grant #KSCX2-EW-R-02 and KSCX2-EW-J-15), and stem cell leading project XDA01010303 to J.D.J.H and by the China Postdoctoral Science Foundation (Grant #2014M551466) to S.J.H.
