Chapter 5
From Gene Expression to Disease Phenotypes: Network-Based Approaches to Study Complex Human Diseases

Quanwei Zhang, Wen Zhang, Rubén Nogales-Cadenas, Jhin-Rong Lin, Ying Cai and Zhengdong D. Zhang

Abstract Gene expression is a fundamental biological process under tight regulation at all levels in normal cells. Its dysregulation can cause abnormal cell behaviors and result in diseases, and thus gene expression profiling and analysis have been widely used to provide the first clue about the molecular mechanisms of human diseases. Because genes and their products interact with and regulate one another, it is essential to analyze gene expression data and understand the genetics of disease in a biological network context. In this chapter, we first introduce the state-of-the-art gene expression analysis (GEA) with network integration and the joint analysis of mRNA and miRNA expression to understand disease regulatory mechanisms and then discuss how disease genes are predicted by incorporating knowledge of gene regulation and characterized in biological networks.

Keywords Gene expression · Disease phenotypes · Biological networks

5.1 Introduction

In the central dogma of biology, gene expression is the intermediate, critical step at which genetic information flows from DNA to functional gene products such as proteins and noncoding RNAs through RNA transcription and translation in each cell. It is the key step where various types of gene regulation—including DNA modification, transcriptional regulation, and posttranscriptional modification—take place. Gene regulation receives and spreads signals in the form of gene regulatory networks (GRNs), in which a group of genes interact with each other and control certain cell functions. Dysregulated gene expression in the network due to promoter

Q. Zhang · W. Zhang · R. Nogales-Cadenas · J.-R. Lin · Y. Cai · Z.D. Zhang (✉)
Department of Genetics, Albert Einstein College of Medicine, Michael F. Price Center for Genetic and Translational Medicine, 1301 Morris Park Avenue, Bronx, NY 10461, USA
e-mail: zhengdong.zhang@einstien.yu.edu

J. Wu, *Transcriptomics and Gene Regulation*,
Translational Bioinformatics 9, DOI 10.1007/978-94-017-7450-5_5

mis-methylation [1, 2], changed transcription factor levels [3, 4], mutated transcriptional regulatory elements (TREs) [5], and miRNA deregulations [6] can result in abnormal cell behaviors and have all been observed in human diseases. Considered as intermediate phenotypes, mRNA expression profiles have been analyzed in biological networks to identify causal genes of human diseases in many studies [7, 8]. In particular, among gene products, microRNAs (miRNAs) are small noncoding RNAs overrepresented in GRNs [9, 10]. Recent studies have revealed their striking gene regulatory activities at the posttranscriptional level [11] and their profound involvement in human diseases [12].

The prevailing assumption about human diseases is that the disease phenotypes are the outcome of interactions between genes and environment [13]. Linking disease phenotypes to genotypes is thus fundamental to understanding human diseases. Linkage analysis has been effective to study disorders with Mendelian inheritance patterns. To date, over 3000 genes with mutations linked to disease phenotypes are cataloged in the Online Mendelian Inheritance in Man (OMIM) database [14]. However, in contrast to Mendelian diseases with simple genetic architectures, complex diseases are characterized by the multifactorial nature and epistasis, in which the causal effects of many risk genes are obscure and cannot be effectively detected by traditional approaches [15, 16]. Furthermore, unlike Mendelian disorders where mutations usually occur within protein coding regions, the majority of mutations of complex diseases occur in noncoding regions associated with gene expression regulation [17, 18]. Deciphering the relationship between genotypes and phenotypes for complex diseases thus requires incorporating the knowledge of gene expression regulation.

Over the last decade, the Encyclopedia of DNA Elements (ENCODE) Consortium has been exploring the functional elements in the human genome and has generated comprehensive data for gene regulation such as transcription factor binding sites and gene–locus interactions [19]. This knowledge provides important basis for analyzing genetic factors of complex diseases. On the other hand, newly developed high-throughput technologies can generate genomic data with an increasingly large sample size and will certainly improve the statistical power to detect subtle associations in complex diseases. This shift has made it possible to tackle the challenges of deciphering complex diseases. With the abundance of genomic data and knowledge of gene regulation, nevertheless, new approaches are needed to integrate genomic data and knowledge of gene regulation to connect genotypes and phenotypes of complex diseases.

Most proteins exert their functions through interactions with other proteins. Such inter- and intracellular interconnectivity implies that the impact of a specific genetic variation is not restricted to the activity of the gene product that carries it, but can spread along the links of the network and alter the activity of other related gene products that otherwise carry no changes. Therefore, an understanding of gene/protein network context is essential to understand the genetics of disease. With the advent of next-generation sequencing, the throughput and the resolution of gene expression profiling have both been increased to an unprecedented level. In addition to traditional methods of gene expression analysis (GEA), network-based

approaches to GEA have also been developed [20–23]. Incorporation of network information into the estimation procedure of the regression model not only encourages smoothness in the estimate of contributions of candidate genes but also integrates into its calculation a priori biological information from the network, which is ignored in conventional methods. A network-based method for gene set enrichment analysis has been developed. Combining a graph-based statistic with an interactive sub-network visualization, EnrichNet takes into account the network structure of physical interactions between the gene sets of interest and improves the prioritization of putative gene set associations as well as exploits information from molecular interaction networks and gene expression data [24]. NetworkAnalyst, another software tool, can perform network analysis and visualization given a gene list. It can also consider multiple meta-data parameters to perform a meta-analysis of multiple gene expression datasets [25].

Not only can disease genes be identified with network-integrated methods, but also they can be studied as a whole in the context of biological networks. Most biological networks are scale-free networks whose degree distribution follows a power law: $P(X = x) = x^{-\alpha}$, in which x is the node degree and α is a constant. In a scale-free network, a small number of nodes tend to have higher degree (such nodes are called hubs), while a large number of nodes have low degrees. Generally, we can divide commonly used network characteristics into different levels. On the gene (protein) level, degree, closeness centrality, and betweenness centrality are often used. They measure, respectively, the number of its interactions, its centeredness in the network, and its importance in communication between genes. On the neighborhood level, clustering coefficient is widely used to measure the probability that the neighbors of a node are connected with one another. On the gene pairs level, one of the most used characteristics is the shortest path between two nodes. Studies of the network characteristics of a group of related disease genes can provide us insights into the molecular mechanisms of the disease.

5.2 Gene Expression Analysis with Network Integration

Gene expression analysis (GEA) has been widely used in human disease studies. High-throughput technologies to profile gene expression include DNA microarrays, serial analysis of gene expression, quantitative RT-PCT, differential-display RT-PCR, and parallel signature sequencing [26]. Network-based GEA is an efficient way to analyze gene expression data because it takes advantage of the functional relationship among genes or their products.

Networks are particularly valuable for modeling large-scale biological systems and have been used with increasing frequency to analyze such complex systems. Graph theory provides useful mathematical tools for general network analysis [27], which can be easily adapted to study genes and pathways. Here, we introduce a class of regression methods with network integration, focusing on the difference between their approaches and applications. We first introduce linear regression with

network regularization. We then present a network-regularized logistic regression method. We next describe a network-regularized Cox model. And finally, we summarize the application results.

5.2.1 Linear Regression Methods with Network Regularization

One issue in GEA is the high dimensionality of the transcriptomic data, e.g., the number of covariates (genes) is much larger than that of observations (samples) [28]. Providing a straightforward mathematical framework for variation indications, linear models have been widely used in data analysis [28]. The biological network can be described as a graph by its adjacency or Laplacian matrix and provides crucial and complementary biological information to gene expression data. A novel linear regression method governed by Laplacian network-deduced matrix has been proposed to identify molecular pathways from gene expression data [20]. In this method, a network-constrained penalty function is used to penalize the L_1-norm of regression coefficients [20]. The method is in essence a mathematical programming problem whose solution criterion is $\hat{\theta} = \arg\min_{\theta} \mathbb{C}(\theta, \lambda, \alpha)$, in which $\hat{\theta}$ is the estimated contribution coefficient of each gene, $\mathbb{C}(\theta, \lambda, \alpha)$ is the network-constrained regularization criterion defined in [20], λ and α are the two parameters to be defined through a leave-one-out cross-validation (CV) process.

5.2.2 Network-Regularized Logistic Regression Method

For classification problems with gene expression data, Logit-Lapnet was put forward to identify molecular pathways associated with breast cancer [21]. It is a regression method combining logistic models and network regularization with the graphical Laplacian matrix. The data matrix is derived from gene expression profiles. The L_1-normed regularization and the corresponding extensions, elastic net and fused lasso, have been used to identify molecular pathways. Extending the previous similar approaches, the Logit-Lapnet method incorporates a priori functional information contained in biological networks. We can consider Logit-Lapnet in a simple way, i.e., as a logistic regression method regularized by lasso and network two items. Its model estimation is formulated as a convex optimization problem, guaranteeing the identifiability of an optimal solution (Fig. 5.1). The optimization criteria, $L(\lambda, \alpha, \beta)$, contains the generalized L_2-norm penalty term using the Laplacian graphical matrix, which encourages smoothness on contribution coefficients (see [21] for a quantitative description of the grouping effect on Logit-Lapnet concerning the structure of network).

$$L(\lambda,\alpha,\beta) = \sum_{i=1}^{n}[-y_i X_i\beta + ln(1+e^{X_i\beta})] + \lambda\alpha \mid \beta \mid_1 + \lambda(1-\alpha)\beta^T L\beta$$
$$\hat{\beta} = \arg\min_{\beta} \mathrm{L}(\lambda,\alpha,\beta)$$

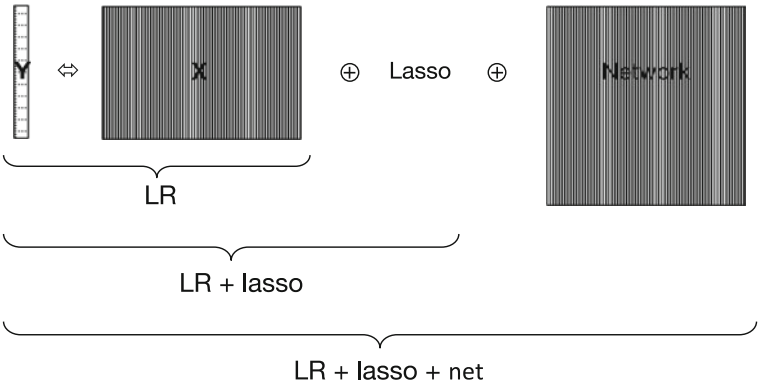


Fig. 5.1 Logit-Lapnet optimization criteria

5.2.3 *Network-Regularized Cox Model and Its Application*

For survival analysis of gene expression data, a Cox proportional hazard model with network regularization was used to select connected network modules predictive of survival of breast cancer patients [29]. Its optimization criterion to estimate gene contribution is a modified likelihood function of the Cox model: $h(t, x_j) = h_0(t)e^{x_j^T\beta}$, in which $h_0(t)$ is the baseline hazard function at time t, x_j the vector of biomarkers for genes, and β the gene coefficient vector. The estimation is defined as $\hat{\beta} = \arg\min_\beta \mathbb{C}(\lambda, \alpha, \beta)$, in which $\mathbb{C}(\lambda, \alpha, \beta)$ contains the negative log likelihood function with $L_1 + L_2$ norm and network regularizations on the coefficient vector. The new Cox model showed better performance in simulation than conventional Cox models and was much more sensitive to cancer-related genes and network modules. Genes identified by the new Cox model have clear biological functions involving cancer cell apoptosis and cell cycle.

5.2.4 *Application Results*

Performance assessment by simulation demonstrated that Logit-Lapnet outperforms elastic net and lasso, two alternative methods (Fig. 5.2) [21]. Application of network-regularized linear regression methods to glioblastoma gene expression data

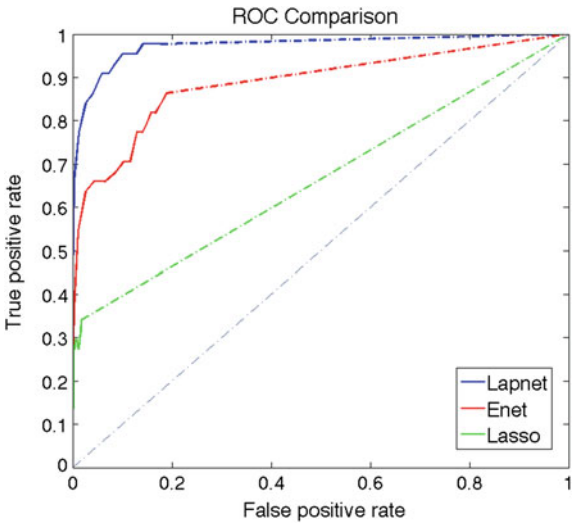


Fig. 5.2 Performance assessment by simulations

identified pathways that might be related to cancer survival time [20]. In a study of biomarkers for breast cancer, Logit-Lapnet selected 262 genes, 166 (∼63 %) of which interact with one another (Fig. 5.3). By comparison, lasso selected only 24 genes, 20 of which are isolated, while elastic net selected 393 genes, 232 (∼59 %) of which are interconnected [21]. The advantage of network-regularized Cox model was demonstrated by its application to breast cancer gene ascertainment [29], in which it selected more known mutated cancer biomarkers than the conventional means.

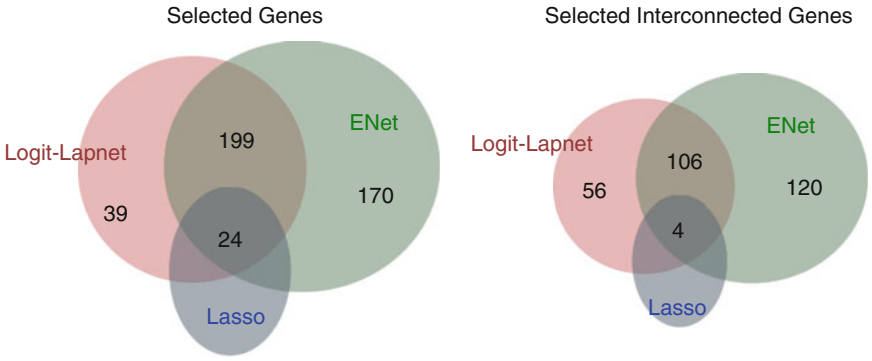


Fig. 5.3 Gene numbers selected by Logit-Lapnet, lasso, and elastic net

5.3 Analyzing Expression of mRNAs and miRNAs to Understand Disease Regulatory Mechanisms

Microarray- and sequencing-based gene expression profiling has been widely used to investigate complex diseases including cancer. Recent studies have discovered gene signatures of numerous diseases and biomarkers for prognosis prediction and disease sub-type classification. For example, Wang et al. [30] and van't Veer et al. [31], respectively, identified ~70 genes that predict breast cancer metastasis risk. Parker et al. [32] proposed a 50-gene PAM50 model, commonly used for breast cancer classification. These markers include genes that control cell cycle, proliferation, DNA replication, and repair, many of which are differentially expressed due to genomic mutations affecting transcriptional regulation.

Testing for differentially expressed genes can yield up to thousands of candidate genes, and one common way to study their functions is to analyze their enrichment in biological pathways. Because the experimentally validated canonical pathways (such as KEGG pathways) are largely incomplete [33], functional interpretation of the candidate genes based on them can be misleading. A less biased approach is based on biological networks, especially those derived from high-throughput data. It can reveal interactions among genes or gene products beyond pathways and has been shown to outperform methods for breast cancer metastasis prediction based on differential expression analysis only [34]. Co-expression networks and GRNs are two representative biological networks widely used to interpret mRNA expression data in disease phenotypes (Fig. 5.4). They are often constructed or inferred for each individual experiment and hence reveal cell type or conditional specific knowledge. In addition, many tools for network-based analysis and visualization have been developed, including GeneMANIA [35] and Cytoscape [36].

Among gene regulatory mechanisms, miRNAs have recently been revealed as one of the most important factors. miRNAs are small noncoding RNA molecules whose main function is to silence gene expression, mainly through transcription repression or mRNA degradation. They are known to be key regulators in important cellular processes such as development [37] and cycle progression [38]. In recent years, they have gained importance in different aspects of human disease research: as targets of miR mimics [39] or antagomirs [40] to reverse disease progression, as biomarkers to detect diseases [41, 42], and as drugs to improve the effect of already developed treatments [43]. Hence, mRNAs and miRNAs regulatory networks analyses are complementary, and both have become indispensable in the study of complex human diseases.

5.3.1 Co-expression Network

Co-expression networks aim at finding genes sharing similar expression patterns across diverse conditions by measuring the correlation of expression between each

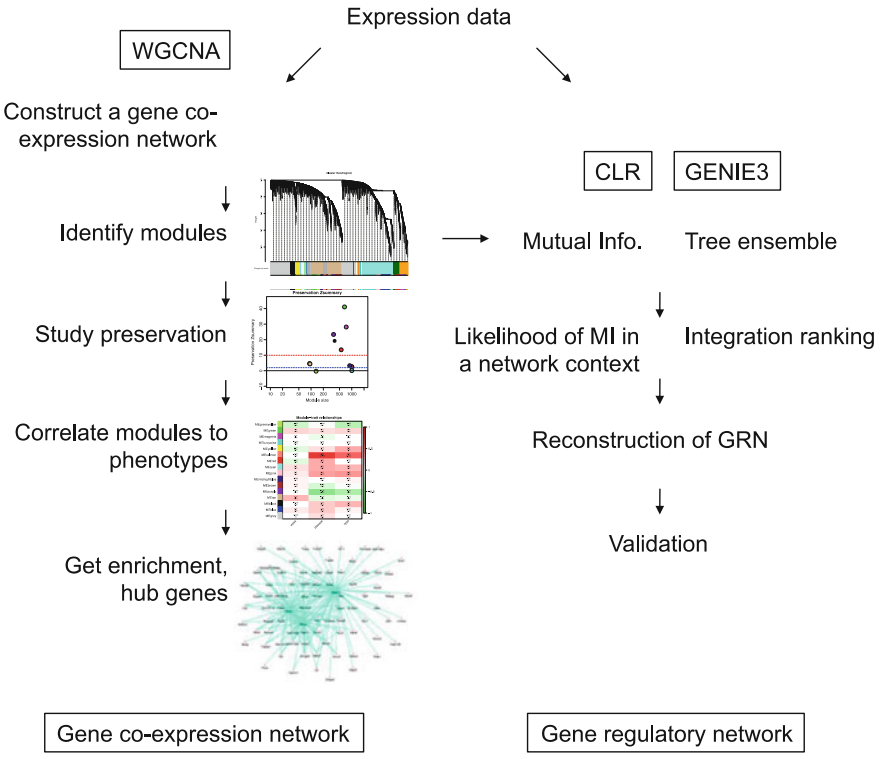


Fig. 5.4 Gene expression data analysis with gene networks

pair of genes, under the assumption that they function together in tightly connected biology processes. The weighted gene co-expression network analysis (WGCNA) [44] is now a popular way to find modules—i.e., groups of genes—as higher-order expression patterns and disease signatures. Gene–gene correlations are first quantified by Pearson's correlation coefficient, and modules are then identified using a topological overlap measure algorithm. A composite Z summary statistic indicates module preservation: whether the modules are robust in different conditions and independent datasets. One can then find contribution made by highly preserved modules to certain trait by measuring correlation coefficient between module eigengene value (the first principal component) and quantitative phenotypes. Hub genes (i.e., genes with many connections) in such modules are important. The WGCNA has been mostly used in developmental studies, where there are no controls and samples are usually arranged in a time course, such as hematopoietic stem cell ontogeny [45] and brain neuron formation [46]. Databases such as GeneMANIA [35] and COXPRESdb [47], which compile assorted datasets, are good co-expression data sources for query genes of interest.

5.3.2 Genetic Regulatory Network

Reconstruction of GRNs is an age-old challenge. Various algorithms can achieve this, but no single method shows the optimal performance across all datasets [48]. One of the well-established methods is context likelihood of relatedness (CLR), an extension of the relevance network technique based on mutual information (MI) [49]. The approach first scores the MI between each pair of a transcriptional regulator (TR) and its potential target gene, and then scores the likelihood of the regulation within its network context; those with high values are likely to form a regulatory relationship. Because a TR may regulates its targets in a nonlinear way, mutual information is a better choice than correlation for not requiring linearity or continuity of the dependence. In addition, the CLR method can be combined with WGCNA to find TRs in modules [45]. Recently, the DREAM4 in Silico network challenge [48] compared over 30 GRN-inference methods for high-throughput data. GENIE3 [50], a random forest-based method, is one of the top-performing methods. It treats GRN inference as a feature selection problem and predicts the expression of a target gene from the expression of all other genes (input genes) using random forests or extra-trees machine learning approaches. The contribution of an input gene on target gene expression is used to build the putative regulatory links. After aggregating links from all genes, the whole GRN is reconstructed from ranked interactions. Databases such as RegulonDB [51] provide experimentally confirmed regulatory interactions that can also validate the accuracy of the GRN inference methods.

5.3.3 miRNAs Regulation in Human Disease

Studies have implicated miRNAs in many diverse illnesses such as hepatitis B and C [52, 53], cardiac and heart diseases [54, 55], and even behavior and neuronal system diseases such as Tourette's syndrome [56]. In particular, important is the study of miRNAs in cancer research, as they are known to regulate important processes in cancer biology such as angiogenesis [57], apoptosis [58], and cell differentiation [59]. Here, we describe the common principle of these analyses—the integration of miRNAs and mRNAs expression, sequence pairwise information, and functional information.

miRNA regulation analysis. miRNAs regulate gene transcriptional activity by total or partial matching of nucleotide sequences with targeted mRNAs. Many computational algorithms are available to predict miRNA targets based on different criteria such as base pairing and target accessibility [60–62]. In general, their predictions are considered to be complementary and are usually combined to increase the overall sensitivity of the prediction [63, 64]. Each method, however, suffers from high false-positive and false-negative rates [65]. This happens even

with the inclusion of experimental validated interactions from databases such as miRWalk [66] or miRecords [67]. Thus, the predicted mRNA–miRNA interactions should be considered as working hypotheses, since they do not necessarily fit with the disease phenotypes. In the study of disease gene regulation, it is advisable to integrate these predictions not only with differential expression values of mRNAs from case and control individuals, but also with miRNAs expression values.

Identification of miRNA regulatory mechanisms. Regulatory mechanisms of biological processes generally involve more than one miRNA and mRNA functioning together. Many computational approaches have been proposed to identify such regulatory mechanisms. They differ from one another in their methodological approaches and their usage of mRNA/miRNA expression values and external information such as potentially involved pathways. Methods used in different contexts include Bayesian networks [68], probabilistic methods [69], LASSO regression [70], or rule-based methods [71]. Despite their differences, the overall analytical flow of these methods is similar (Fig. 5.5).

Functional analysis of miRNA regulation. It is common to infer the function of a miRNA from its gene targets (for possible bias in such an approach, see [72]).

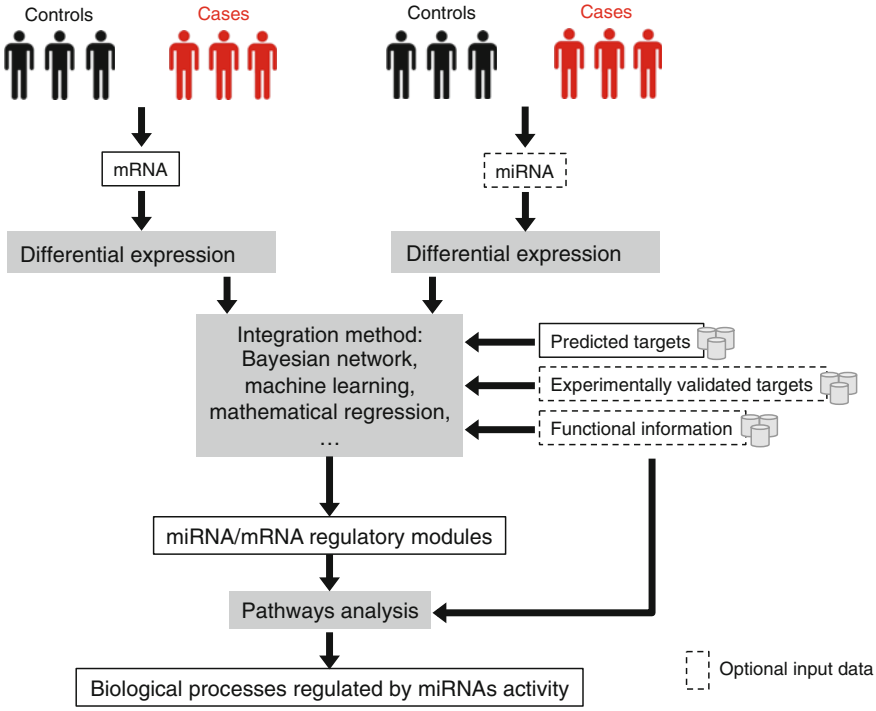


Fig. 5.5 miRNA analysis pipeline

The incorporation of external information, such as functional terms related to mRNA targets, makes it possible to deduce the involvement of miRNAs regulation in biological pathways [73]. This strategy can be used to interpret functional enrichment results and to find regulatory modules of miRNAs–mRNAs participating in the same processes [74]. Several resources provide direct functional annotation of miRNAs (Table 5.1).

Table 5.1 Useful resources of miRNA regulation for human disease studies

Resource	Description	Ref.
MiRNA database		
miRBase	Database of miRNA sequences and annotations for 206 different organisms	[99]
miRNA-target interaction		
microRNA.org	Database of predicted microRNA targets & target down-regulation scores. It includes experimentally observed expression patterns	[100]
miRWalk	Database that provides information on miRNA from human, mouse and rat on their predicted as well as validated binding sites on their target genes. It includes information on experimentally validated miRNA interaction information associated with genes, pathways, diseases, organs, OMIM disorders, cell lines, and literature on miRNAs	[66]
multiMiR	R package and database for miRNA-target interaction which includes information based on disease annotation and drug microRNA response, in addition to many experimental and computational databases	[101]
CancerMiner	Database including recurring microRNA-mRNA associations across cancer type	[102]
Functional information		
mir2Disease	A manually curated database providing a comprehensive resource of miRNA deregulation in various human diseases	[103]
mirFocus	Database providing leads for in-depth analysis of miRNA-target gene pathways and the related miRNA annotations	www.mirfocus.org
HmDD	Database with curated experiment-supported evidence for human microRNA (miRNA) and disease associations	[104]
miRCancer	Database providing a collection of miRNA expression profiles in various human cancers, automatically extracted from the published literatures in PubMed	[105]
Variant information		
PolymiRTS	Database of naturally occurring DNA variations in microRNA (miRNA) seed regions and miRNA-target sites underlying in gene expression and disease phenotypes	[106]
miRdSNP	Data source of dSNPs and robust tools to capture their spacial relationship with miRNA-target sites on the 3'UTRs of human genes	[107]

5.4 Predicting Disease Genes by Incorporating Knowledge of Gene Regulation

The identification of disease genes is a fundamental objective in medical research. With the advent of high-throughput genotyping technologies, a large number of disease-associated variants have been identified by genome-wide association studies (GWASs) [75]. Such disease variants provide valuable signals for uncovering underlying disease genes and unraveling disease mechanisms, which can be improved by leveraging the knowledge of gene regulation.

5.4.1 Importance of Knowledge of Gene Regulation in Complex Disease Prediction

Both genetic predisposition and environmental factors may contribute to the pathogenesis of complex diseases. The origins of genetic predisposition are genetic variants that affect gene functions and thus contribute to disease susceptibility. Some of these variants are located in coding regions and affect gene functions by altering the corresponding protein sequences. The others, located in noncoding regions, may affect (TREs), such as transcription factor binding sites, resulting in dysregulation of gene expression.

Uncovering disease causal genes that underlie the association signals discovered in GWAS is challenging. The simplest method is to select genes closest to disease-associated variants as the causal genes. However, because single nucleotide polymorphisms (SNPs) used in GWAS are tagging SNPs, representing linkage disequilibrium (LD) blocks, disease-associated SNPs discovered in GWAS are most likely not causal SNPs but mere their proxies. Another more sophisticated method is to first define the LD regions tagged by GWAS SNPs and then identify genes overlapping LD regions as candidate causal genes [76]. Causal genes near GWAS SNPs are likely to be included in this way. However, causal genes whose expression is affected by causal SNPs through modifying their TREs will almost certainly be missed, as they fall outside LD regions. To include these "distal" causal genes, it requires knowledge of gene regulation and, more specifically, knowledge of regulatory relationship between loci and genes.

5.4.2 Gene Regulation Data Resources and Complex Disease Risk Loci

Studies have shown that disease-associated SNPs are overrepresented in loci implicated in gene regulations [77–79] (Fig. 5.6). There are several important resources for the knowledge of aforementioned gene-locus regulation linkage.

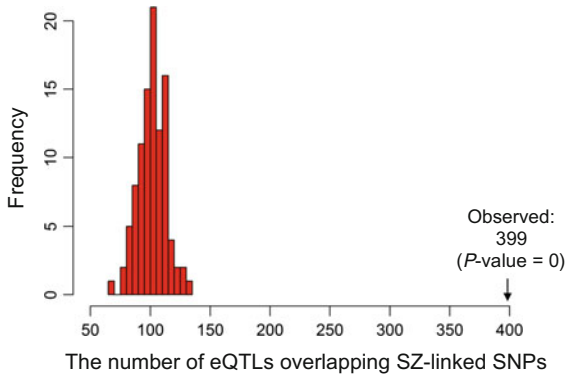


Fig. 5.6 Enrichment of schizophrenia-associated SNPs at eQTLs. We compiled 125,568 eQTLs from GTEx studies and identified 15,027 SNPs in high linkage disequilibrium with 261 schizophrenia-associated SNPs that we collected from the GWAS catalog [111] and a meta-analysis of schizophrenia [76]. 399 eQTLs are SZ-linked SNPs (P = 0, permutation test with 100 repetitions)

Expression quantitative trait loci (eQTL) are genomic loci whose genotypes are associated with transcript levels. eQTL data provide valuable information of gene-locus regulatory relationship and are useful in prioritizing GWAS signals [80]. In addition, the ENCODE Project inferred regulatory relationship from correlation between DNase I hypersensitivity of loci and promoters in different cell and tissue types [81]. Furthermore, FANTOM5 generated regulation information between enhancers and target genes by comparing their transcriptional activities across different cell types [78]. These regulatory data repositories serve as important information resources for not only prioritizing but also exploring new disease causal factors, on both SNP and gene levels.

5.4.3 Linking Distal Candidate Causal Genes by Incorporating the Knowledge of Gene Regulation

As mentioned earlier, causal genes may not always fall in the same haplotype block carrying GWAS SNPs, and thus, it requires other information in addition to LD to identify them. Figure 5.7 shows an example of successfully uncovering a promising causal gene underlying a GWAS SNP by using the gene regulatory information. SNP rs2159767 is a GWAS SNP associated with schizophrenia [82]. The LD region indexed by rs2159767 is in a gene desert and thus devoid of any genes. In it, however, we found two TREs that are likely to regulate two distal genes, fragile X mental retardation 1 (FMR1), and fragile X mental retardation 1 neighbor (FMR1NB), respectively. Notably, FMR1 is a literature-supported SZ gene [83],

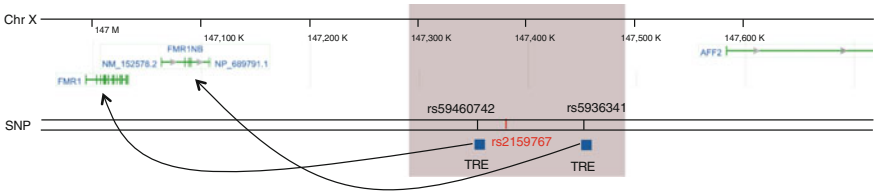


Fig. 5.7 Distal disease causal gene candidates. Gene regulatory information can link genes far away from the disease-associated GWAS SNP (schizophrenia-associated rs2159767 in this case) to the disease risk region (the *red block*)

and we found that a SNP (rs59460742) within the TRE associated with FMR1 is in strong LD (r^2 = 0.587) with rs2159767. Those evidences imply that the causal factor of the GWAS signal could be the SNP within the TRE that results in the dysregulation of FMR1.

5.4.4 Distal and Proximal Candidate Causal Genes

In general, incorporating LD information can improve the detection of causal genes in the proximity of GWAS signals, but finding distal causal genes relies on the knowledge of gene regulation. Using LD and gene regulation information, we identified three overlapping sets of candidate causal genes for schizophrenia (Fig. 5.8). There are 485 proximal and 158 distal candidate causal genes. Together, these two numbers indicate that incorporating gene regulatory information can substantially expand the set of candidate causal genes (about one-third in the aforementioned schizophrenia case). Although irrelevant distal genes could be

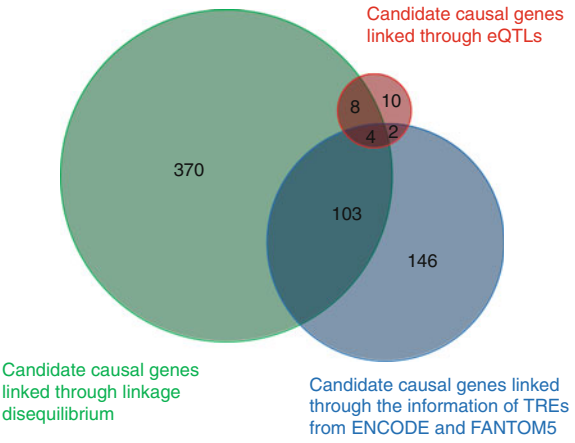


Fig. 5.8 Schizophrenia causal gene candidates. Candidates genes are linked to 261 schizophrenia-associated SNPs through different gene regulatory information

introduced due to false regulatory linkage, incorporating the knowledge of gene regulation can cover potential risk genes in a more comprehensive manner, which will also facilitate the downstream analysis.

5.5 Characterizing the Network and Association Properties of Disease Genes

Since last decade, a large number of causal or closely related genes have been reported for many diseases by experimental or computational methods [84, 85]. However, a complex disease usually reflects the perturbation to the complex intracellular network, rather than a consequence of an abnormality within a single gene [86]. By studying disease genes in the context of biological networks, we consider the disease genes as a whole instead of studying them individually. Such studies may not only provide clues to uncover the molecular mechanisms of diseases, but also reveal distinguishing properties of disease genes, which can be used to predict unknown disease genes.

5.5.1 Network Characteristics Analysis of Disease Genes

Interactions among disease genes in biological networks. Disease genes can be mapped into the network (Fig. 5.9a), and a sub-network around them can be extracted to obtain a view of the local interactions among them [27]. It is well-known that the protein products of different genes harboring causal mutations for the same Mendelian disease often physically interact. A recent study suggested that in many complex diseases, proteins encoded by genes from disease-associated regions also tend to physically interact [87]. This characteristic is the foundation of "guilty-by association" policy to predict unknown disease genes.

Distinct network properties of disease genes. Studies have found that some network properties can distinguish a group of disease genes from background genes or another set of genes, and thus are particularly informative for the relevant disease (Fig. 5.9b, c). In yeast, it was found that disease genes in general tend to have higher degrees, cluster together, and locate at the central network locations [88], but another study on human did not find higher degrees for disease genes [89]. In humans, it was reported that cancer proteins tend to have higher degrees and locate at central part of the network [90]. Moreover, it was found that cancer proteins tend to have higher betweenness (which measures the importance of a gene in communication between other gene pairs) and shorter shortest-paths than both the essential and the background proteins [91]. The specificity of network characteristics of disease genes can provide us clues to specific mechanisms behind the diseases.

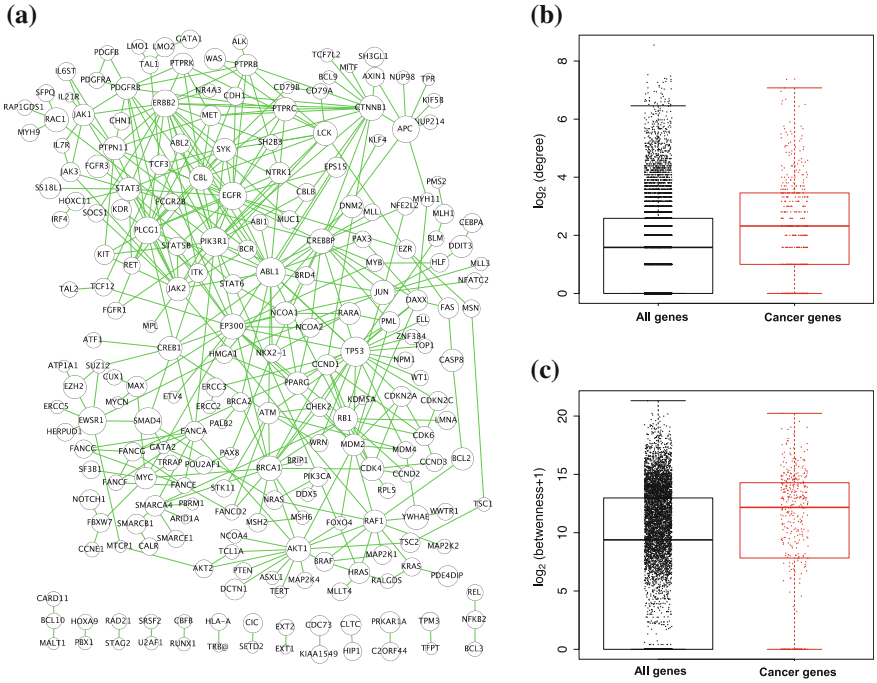


Fig. 5.9 Network characteristics of cancer genes. Among 547 cancer genes from COSMIC (Version 70; Aug 2014) [112], 386 of them were analyzed in the background network HINT [113]. **a** 394 directly physical interactions between cancer genes products. **b** Cancer genes tend to have higher degrees than background genes in HINT ($P = 5.136 \times 10^{-22}$, Wilcoxon rank-sum test). **c** Cancer genes tend to have higher betweenness than background genes in HINT ($P = 3.509 \times 10^{-18}$, Wilcoxon rank-sum test)

Network characteristics of disease genes in different biological networks and species. A recent cancer study found that prognostic genes are less likely to be hub genes in co-expression networks, and this pattern is unique to the corresponding cancer-type-specific network. Enriched in modules, prognostic genes are especially likely to be module genes conserved across different cancer co-expression networks [92]. In addition to co-expression network, researchers also integrated tissue-specific gene expression with protein interaction to derive tissue-specific PPI networks [93]. This provides an opportunity to study network characteristics of disease genes in tissue-specific PPI networks.

5.5.2 *Software Tools for Network Characteristics Analysis*

Many software tools have been developed for network characteristics analysis (Table 5.2). Some allow users to upload their own gene list for targeted analysis.

Table 5.2 Tools for network characteristics analysis

Tools	Description and access	Ref.
Targeted analysis		
TopoGSA	Generate 2D or 3D plots of network characteristics to visualize the network characteristic for each uploaded gene. Comparison with known gene sets based on 2D or 3D plots to visually identify similar pathways to the uploaded dataset The Web server can be accessed at http://www.topogsa.org	[94]
SNOW	Compute several network characteristics and estimate the statistical significance by comparing the network characteristics of the uploaded genes to those of the background genes or those in random networks The Web server can be accessed at http://snow.bioinfo.cipf.es	[95]
NetworkAnalyzer	Compute and display a comprehensive set of topological parameters. It can analyze the whole network or subset of nodes from the network It is a plug-in of Cytoscape	[96]
General analysis		
CentiScaPe	Compute 9 kinds of centralities of genes (proteins) in biological networks. It can highlight the genes whose centralities are higher or (lower) than the user-defined thresholds. It can generate "plot by node," which shows the centralities of one gene with background information about the centralities (e.g., min, mean). It can also generate "plot by centrality" to identify group of genes clustered together according to combinations of centralities. Attributes from experiments can be also uploaded to analyze relationship between experimental data and gene centralities It is a plug-in of Cytoscape	[108]
CenTiBiN	Compute and explore 17 kinds of centralities of genes (proteins) in biological networks The Web server can be accessed at http://centibin.ipk-gatersleben.de, and there is also instable Windows application.	[109]
CentiLib	CentiLib is a Java-based library and user-friendly plug-in for the analysis and visual exploration of centralities in networks. CentiLib can achieve similar functions as CenTiBiN, but it is easier to use and it can deal with weighted networks The software and manual can be downloaded at http://centilib.ipk-gatersleben.de/	[110]

For example, TopoGSA can generate 2D or 3D plots for submitted genes, which show difference network characteristics simultaneously [94]. When microarray data are uploaded, differentially expressed genes can be automatically identified and used as targeted genes for the analysis. TopoGSA can also compare the network characteristics of targeted genes with those of known gene sets (e.g., pathways). SNOW [95], a similar tool, can calculate the network characteristics and estimate their statistical significance. NetworkAnalyzer can also carry out a similar analysis when genes from the network are selected [96]. In addition to these methods, several tools for general network analysis can also be helpful (Table 5.2).

5.5.3 *Association Between Disease Genes and Other Gene Sets*

Another important utility of networks is to find the association between disease genes and other functional groups of genes. For example, recent studies suggested that the it is important to consider the relationship between genetic diseases and the aging process for understanding the molecular mechanisms of complex diseases. To better understand such association, one study investigated the relationship among aging genes and disease genes in a human disease-aging network [97]. The study found that (1) human disease genes are much closer to aging genes than expected by chance; (2) aging genes contribute significantly to association among diseases compared with nonaging genes with similar degrees.

It is important to assess functional association between a group of genes (e.g., candidate disease genes) and predefined gene sets. Overrepresentation-based enrichment analysis is commonly used for this task. This method, however, has several shortcomings. First, only shared genes between the input gene list and the known gene sets are considered, but current data of gene sets are not complete. Second, genes in the gene sets are treated equally, disregarding the network structure of physical or functional interactions between genes. To address these limitations, it is applicable to combine information of protein–protein interaction network with known get sets. To tackle these problems, several such tools have been developed. Glaab et al. [98] combined information from pathways databases and interaction networks and obtained more robust pathways and process representations. Their method first maps the genes in pathways into a protein–protein interaction network and then extends the pathways by including densely interacting partners. Later, Glaab et al. [24] proposed another tool for network-based gene set enrichment analysis. This approach first maps the target genes and reference gene sets into the network. It then scores the distance between the mapped target genes and reference dataset using a random walk with restart algorithm and compares the score against a background model. This method can use the network distance to differentiate gene sets with similar enrichment levels assessed by overrepresentation analysis. More importantly, it can identify novel functional associations (with no or few shared genes) and can evaluate tissue-specific association.

5.6 Conclusions

Gene expression is under tight regulation at all levels in normal cells. The characteristic forms and behaviors of different cell types are the result of their varying patterns of expression of the same set of genes. The dysregulation of gene expression can cause abnormal cell behaviors and result in diseases, and thus, gene expression profiling could provide the first clue about the molecular mechanisms of a disease. Two recent developments are spearheading the advancement of disease

research in this field: First, next-generation sequencing technologies have increased the throughput and the resolution of gene expression studies to an unprecedented level; second, new computational methods with sophisticated data integration, especially network integration, have been developed for gene expression data analysis. Biological networks can provide important a priori functional information in data analysis, and since last decade, many different types of them have been constructed: Not only the number has increased but also the coverage of them has increased dramatically. With such recent resource and technology development, biology has entered a new data-driven phase in the twenty-first century. Now is a particularly challenging and exciting time for disease research with gene expression assay, as more and more gene expression data are being generated at an ever-accelerating speed.

References

1. Manel E, Paul C, Stephen B, James H. A gene hypermethylation profile of human cancer. Cancer Res. 2001;61:3225–9.
2. Jones PA, Baylin SB. The fundamental role of epigenetic events in cancer. Nat Rev Genet. 2002;3:415–28.
3. Darnell JE Jr. Transcription factors as targets for cancer therapy. Nat Rev Cancer. 2002;2:740–9.
4. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144:646–74.
5. Herz HM, Hu D, Shilatifard A. Enhancer malfunction in cancer. Mol Cell. 2014;53:859–66.
6. Croce CM. Causes and consequences of microRNA dysregulation in cancer. Nat Rev Genet. 2009;10:704–14.
7. Akavia UD, Litvin O, Kim J, Sanchez-Garcia F, Kotliar D, et al. An integrated approach to uncover drivers of cancer. Cell. 2010;143:1005–17.
8. Kim YA, Wuchty S, Przytycka TM. Identifying causal genes and dysregulated pathways in complex diseases. PLoS Comput Biol. 2011;7:e1001095.
9. Herranz H, Cohen SM. MicroRNAs and gene regulatory networks: managing the impact of noise in biological systems. Genes Dev. 2010;24:1339–44.
10. Tsang J, Zhu J, van Oudenaarden A. MicroRNA-mediated feedback and feedforward loops are recurrent network motifs in mammals. Mol Cell. 2007;26:753–67.
11. Stark A, Brennecke J, Bushati N, Russell RB, Cohen SM. Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution. Cell. 2005;123:1133–46.
12. Lu M, Zhang Q, Deng M, Miao J, Guo Y, et al. An analysis of human microRNA and disease associations. PLoS ONE. 2008;3:e3420.
13. Ramos RG, Olden K. Gene-environment interactions in the development of complex disease phenotypes. Int J Environ Res Public Health. 2008;5:4–11.
14. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 2005;33:D514–7.
15. Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science. 1996;273:1516–7.
16. Mayeux R. Mapping the new frontier: complex genetic disorders. J Clin Invest. 2005;115:1404–7.

17. Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nat Genet. 2003;33 (Suppl):228–37.
18. Ritchie GR, Dunham I, Zeggini E, Flicek P. Functional annotation of noncoding sequence variants. Nat Methods. 2014;11:294–6.
19. Raney BJ, Cline MS, Rosenbloom KR, Dreszer TR, Learned K, et al. ENCODE whole-genome data in the UCSC genome browser (2011 update). Nucleic Acids Res. 2011;39:D871–5.
20. Li C, Li H. Network-constrained regularization and variable selection for analysis of genomic data. Bioinformatics. 2008;24:1175–82.
21. Zhang W, Wan YW, Allen GI, Pang K, Anderson ML, et al. Molecular pathway identification using biological network-regularized logistic models. BMC Genom. 2013;14 (Suppl 8):S7.
22. Wu C, Zhu J, Zhang X. Network-based differential gene expression analysis suggests cell cycle related genes regulated by E2F1 underlie the molecular difference between smoker and non-smoker lung adenocarcinoma. BMC Bioinform. 2013;14:365.
23. Ruan J, Dean AK, Zhang W. A general co-expression network-based approach to gene expression analysis: comparison and applications. BMC Syst Biol. 2010;4:8.
24. Glaab E, Baudot A, Krasnogor N, Schneider R, Valencia A. EnrichNet: network-based gene set enrichment analysis. Bioinformatics. 2012;28:i451–7.
25. Xia J, Gill EE, Hancock RE. NetworkAnalyst for statistical, visual and network-based meta-analysis of gene expression data. Nat Protoc. 2015;10:823–44.
26. Fryer RM, Randall J, Yoshida T, Hsiao LL, Blumenstock J, et al. Global analysis of gene expression: methods, interpretation, and pitfalls. Exp Nephrol. 2002;10:64–74.
27. Lemetre C, Zhang Q, Zhang ZD. SubNet: a Java application for subnetwork extraction. Bioinformatics. 2013;29:2509–11.
28. Marko NF, Weil RJ. Mathematical modeling of molecular data in translational medicine: theoretical considerations. Sci Transl Med. 2010;2:56rv54.
29. Wan YW, Nagorski J, Allen GI, Li ZH, Liu ZD. Identifying cancer biomarkers through a network regularized Cox model. In: Genomic Signal Processing and Statistics (GENSIPS), 2013 IEEE international workshop on IEEE. Houston, TX, 2013; pp. 36–39.
30. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet. 2005;365:671–9.
31. van't Veer LJ, Dai HY, van de Vijver MJ, He YDD, Hart AAM, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002;415:530–6.
32. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol. 2009;27:1160–7.
33. Atias N, Istrail S, Sharan R. Pathway-based analysis of genomic variation data. Curr Opin Genet Dev. 2013;23:622–6.
34. Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Network-based classification of breast cancer metastasis. Mol Syst Biol. 2007;3:140.
35. Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res. 2010;38:W214–20.
36. Saito R, Smoot ME, Ono K, Ruscheinski J, Wang PL, et al. A travel guide to Cytoscape plugins. Nat Methods. 2012;9:1069–76.
37. Lee RC, Feinbaum RL, Ambros V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell. 1993;75:843–54.
38. Zhao Y, Ransom JF, Li A, Vedantham V, von Drehle M, et al. Dysregulation of cardiogenesis, cardiac conduction, and cell cycle in mice lacking miRNA-1-2. Cell. 2007;129:303–17.

39. Trang P, Wiggins JF, Daige CL, Cho C, Omotola M, et al. Systemic delivery of tumor suppressor microRNA mimics using a neutral lipid emulsion inhibits lung tumors in mice. Mol Ther. 2011;19:1116–22.
40. Wahlquist C, Jeong D, Rojas-Munoz A, Kho C, Lee A, et al. Inhibition of miR-25 improves cardiac contractility in the failing heart. Nature. 2014;508:531–5.
41. Ludwig N, Nourkami-Tutdibi N, Backes C, Lenhof HP, Graf N, et al. Circulating serum miRNAs as potential biomarkers for nephroblastoma. Pediatr Blood Cancer. 2015;62:1360–1367.
42. van Schooneveld E, Wildiers H, Vergote I, Vermeulen PB, Dirix LY, et al. Dysregulation of microRNAs in breast cancer and their potential role as prognostic and predictive biomarkers in patient management. Breast Cancer Res. 2015;17:526.
43. Knezevic J, Pfefferle AD, Petrovic I, Greene SB, Perou CM, et al. Expression of miR-200c in claudin-low breast cancer alters stem cell functionality, enhances chemosensitivity and reduces metastatic potential. Oncogene. 2015; doi:10.1038/onc.2015.48.
44. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. Bmc Bioinform. 2008;9:559.
45. McKinney-Freeman S, Cahan P, Li H, Lacadie SA, Huang HT, et al. The transcriptional landscape of hematopoietic stem cell ontogeny. Cell Stem Cell. 2012;11:701–14.
46. Miller JA, Ding SL, Sunkin SM, Smith KA, Ng L, et al. Transcriptional landscape of the prenatal human brain. Nature. 2014;508:199.
47. Okamura Y, Aoki Y, Obayashi T, Tadaka S, Ito S, et al. COXPRESdb in 2015: coexpression database for animal species by DNA-microarray and RNAseq-based expression data with multiple quality assessment systems. Nucleic Acids Res. 2015;43:D82–6.
48. Marbach D, Costello JC, Kuffner R, Vega NM, Prill RJ, et al. Wisdom of crowds for robust gene network inference. Nat Methods. 2012;9:796.
49. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, et al. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol. 2007;5:54–66.
50. Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. Inferring regulatory networks from expression data using tree-based methods. PLoS One. 2010;5:e12776.
51. Salgado H, Peralta-Gil M, Gama-Castro S, Santos-Zavaleta A, Muniz-Rascado L, et al. RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. Nucleic Acids Res. 2013;41:D203–13.
52. Jiang J, Gusev Y, Aderca I, Mettler TA, Nagorney DM, et al. Association of MicroRNA expression in hepatocellular carcinomas with hepatitis infection, cirrhosis, and patient survival. Clin Cancer Res. 2008;14:419–27.
53. Jopling CL, Yi M, Lancaster AM, Lemon SM, Sarnow P. Modulation of hepatitis C virus RNA abundance by a liver-specific MicroRNA. Science. 2005;309:1577–81.
54. Wang X, Zhang X, Ren XP, Chen J, Liu H, et al. MicroRNA-494 targeting both proapoptotic and antiapoptotic proteins protects against ischemia/reperfusion-induced cardiac injury. Circulation. 2010;122:1308–18.
55. Xu J, Hu Z, Xu Z, Gu H, Yi L, et al. Functional variant in microRNA-196a2 contributes to the susceptibility of congenital heart disease in a Chinese population. Hum Mutat. 2009;30:1231–6.
56. Abelson JF, Kwan KY, O'Roak BJ, Baek DY, Stillman AA, et al. Sequence variants in SLITRK1 are associated with Tourette's syndrome. Science. 2005;310:317–20.
57. Yang F, Wang W, Zhou C, Xi W, Yuan L, et al. MiR-221/222 promote human glioma cell invasion and angiogenesis by targeting TIMP2. Tumour Biol. 2015;36:3763.
58. Zhao S, Yao D, Chen J, Ding N, Ren F. MiR-20a promotes cervical cancer proliferation and metastasis in vitro and in vivo. PLoS ONE. 2015;10:e0120905.
59. Houbaviy HB, Murray MF, Sharp PA. Embryonic stem cell-specific MicroRNAs. Dev Cell. 2003;5:351–8.
60. Enright AJ, John B, Gaul U, Tuschl T, Sander C, et al. MicroRNA targets in Drosophila. Genome Biol. 2003;5:R1.

61. Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E. The role of site accessibility in microRNA target recognition. Nat Genet. 2007;39:1278–84.
62. Thadani R, Tammi MT. MicroTar: predicting microRNA targets from RNA duplexes. BMC Bioinform. 2006;7(Suppl 5):S20.
63. Stingo FC, Chen YA, Vannucci M, Barrier M, Mirkes PE. A Bayesian graphical modeling approach to microRNA regulatory network inference. Ann Appl Stat. 2010;4:2024–48.
64. Tabas-Madrid D, Muniategui A, Sanchez-Caballero I, Martinez-Herrera DJ, Sorzano CO, et al. Improving miRNA-mRNA interaction predictions. BMC Genom. 2014;15(Suppl 10):S2.
65. Ritchie W, Flamant S, Rasko JE. Predicting microRNA targets and functions: traps for the unwary. Nat Methods. 2009;6:397–8.
66. Dweep H, Sticht C, Pandey P, Gretz N. miRWalk–database: prediction of possible miRNA binding sites by "walking" the genes of three genomes. J Biomed Inform. 2011;44:839–47.
67. Xiao F, Zuo Z, Cai G, Kang S, Gao X, et al. miRecords: an integrated resource for microRNA-target interactions. Nucleic Acids Res. 2009;37:D105–10.
68. Huang JC, Babak T, Corson TW, Chua G, Khan S, et al. Using expression profiling data to identify human microRNA targets. Nat Methods. 2007;4:1045–9.
69. Joung JG, Hwang KB, Nam JW, Kim SJ, Zhang BT. Discovery of microRNA-mRNA modules via population-based probabilistic learning. Bioinformatics. 2007;23:1141–7.
70. Muniategui A, Nogales-Cadenas R, Vazquez M, Aranguren XL, Agirre X, et al. Quantification of miRNA-mRNA interactions. PLoS ONE. 2012;7:e30766.
71. Tran DH, Satou K, Ho TB. Finding microRNA regulatory modules in human genome using rule induction. BMC Bioinform. 2008;9(Suppl 12):S5.
72. Bleazard T, Lamb JA, Griffiths-Jones S Bias in microRNA functional enrichment analysis. Bioinformatics. 2015;31:1592–1598.
73. Gusev Y, Schmittgen TD, Lerner M, Postier R, Brackett D. Computational analysis of biological functions and pathways collectively targeted by co-expressed microRNAs in cancer. BMC Bioinform. 2007;8(Suppl 7):S16.
74. Liu B, Li J, Cairns MJ. Identifying miRNAs, targets and functions. Brief Bioinform. 2014;15:1–19.
75. Welter D, MacArthur J, Morales J, Burdett T, Hall P, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014;42:D1001–6.
76. Schizophrenia Working Group of the Psychiatric Genomics C. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014;511:421–7.
77. Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, et al. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet. 2010;6:e1000888.
78. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, et al. An atlas of active enhancers across human cell types and tissues. Nature. 2014;507:455–61.
79. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012;337:1190–5.
80. Hou L, Zhao H. A review of post-GWAS prioritization approaches. Front Genet. 2013;4:280.
81. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, et al. The accessible chromatin landscape of the human genome. Nature. 2012;489:75–82.
82. Sullivan PF, Lin D, Tzeng JY, van den Oord E, Perkins D, et al. Genomewide association for schizophrenia in the CATIE study: results of stage 1. Mol Psychiatry. 2008;13:570–84.
83. Kelemen O, Kovacs T, Keri S. Contrast, motion, perceptual integration, and neurocognition in schizophrenia: the role of fragile-X related mechanisms. Prog Neuropsychopharmacol Biol Psychiatry. 2013;46:92–7.
84. Bauer-Mehren A, Bundschus M, Rautschka M, Mayer MA, Sanz F, et al. Gene-disease network analysis reveals functional modules in mendelian, complex and environmental diseases. Plos One. 2011;6:e20284.
85. Rappaport N, Nativ N, Stelzer G, Twik M, Guan-Golan Y, et al. MalaCards: an integrated compendium for diseases and their annotation. Database J Biol Databases Curation. 2013; bat018.

86. Barabasi AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet. 2011;12:56–68.
87. Rossin EJ, Lage K, Raychaudhuri S, Xavier RJ, Tatar D, et al. Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. Plos Genet. 2011;7:e1001273.
88. Ideker T, Sharan R. Protein networks in disease. Genome Res. 2008;18:644–52.
89. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, et al. The human disease network. Proc Natl Acad Sci USA. 2007;104:8685–90.
90. Jonsson PF, Bates PA. Global topological features of cancer proteins in the human interactome. Bioinformatics. 2006;22:2291–7.
91. Sun JC, Zhao ZM. A comparative study of cancer proteins in the human protein-protein interaction network. Bmc Genomics 2010;11.
92. Yang Y, Han L, Yuan Y, Li J, Hei NN, et al. Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. Nat Commun 2014;5:3231.
93. Magger O, Waldman YY, Ruppin E, Sharan R. Enhancing the prioritization of disease-causing genes through tissue specific protein interaction networks. Plos Comput Biol 2012;8:e1002690.
94. Glaab E, Baudot A, Krasnogor N, Valencia A. TopoGSA: network topological gene set analysis. Bioinformatics. 2010;26:1271–2.
95. Minguez P, Gotz S, Montaner D, Al-Shahrour F, Dopazo J. SNOW, a web-based tool for the statistical analysis of protein-protein interaction networks. Nucleic Acids Res. 2009;37: W109–14.
96. Assenov Y, Ramirez F, Schelhorn SE, Lengauer T, Albrecht M. Computing topological parameters of biological networks. Bioinformatics. 2008;24:282–4.
97. Wang JG, Zhang SH, Wang Y, Chen LN, Zhang XS. Disease-aging network reveals significant roles of aging genes in connecting genetic diseases. Plos Comput Biology 2009;5: e1000521.
98. Glaab E, Baudot A, Krasnogor N, Valencia A. Extending pathways and processes using molecular interaction networks to analyse cancer genome data. Bmc Bioinform. 2010;11:597.
99. Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res. 2014;42:D68–73.
100. Betel D, Wilson M, Gabow A, Marks DS, Sander C. The microRNA.org resource: targets and expression. Nucleic Acids Res. 2008;36:D149–53.
101. Ru Y, Kechris KJ, Tabakoff B, Hoffman P, Radcliffe RA, et al. The multiMiR R package and database: integration of microRNA-target interactions along with their disease and drug associations. Nucleic Acids Res. 2014;42:e133.
102. Jacobsen A, Silber J, Harinath G, Huse JT, Schultz N, et al. Analysis of microRNA-target interactions across diverse cancer types. Nat Struct Mol Biol. 2013;20:1325–32.
103. Jiang Q, Wang Y, Hao Y, Juan L, Teng M, et al. miR2Disease: a manually curated database for microRNA deregulation in human disease. Nucleic Acids Res. 2009;37:D98–104.
104. Purcell SM, Moran JL, Fromer M, Ruderfer D, Solovieff N, et al. A polygenic burden of rare disruptive mutations in schizophrenia. Nature. 2014;506:185–90.
105. Xie B, Ding Q, Han H, Wu D. miRCancer: a microRNA-cancer association database constructed by text mining on literature. Bioinformatics. 2013;29:638–44.
106. Bhattacharya A, Ziebarth JD, Cui Y. PolymiRTS Database 3.0: linking polymorphisms in microRNAs and their target sites with human diseases and biological pathways. Nucleic Acids Res. 2014;42:D86–91.
107. Bruno AE, Li L, Kalabus JL, Pan Y, Yu A, et al. miRdSNP: a database of disease-associated SNPs and microRNA target sites on 3'UTRs of human genes. BMC Genom. 2012;13:44.
108. Scardoni G, Petterlini M, Laudanna C. Analyzing biological network parameters with CentiScaPe. Bioinformatics. 2009;25:2857–9.

109. Junker BH, Koschutzki D, Schreiber F. Exploration of biological network centralities with CentiBiN. Bmc Bioinform. 2006;7:219.
110. Grassler J, Koschutzki D, Schreiber F. CentiLib: comprehensive analysis and exploration of network centralities. Bioinformatics. 2012;28:1178–9.
111. Hindorff LA MJEBI, Morales J (European Bioinformatics Institute), Junkins HA, Hall PN, Klemm AK, Manolio TA. (Available at: http://www.genome.gov/gwastudies). A catalog of published genome-wide association studies. Accessed 31 Mar 2015.
112. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res. 2014;43: D805–11.
113. Das J, Yu HY. HINT: High-quality protein interactomes and their applications in understanding human disease. BMC Syst Biol. 2012;6:92.

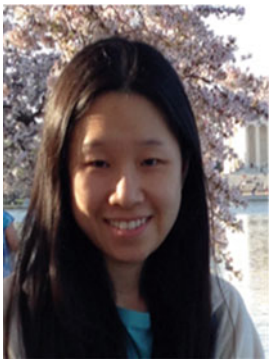
Author Biographies

Dr. Quanwei Zhang received his PhD degree at Xi'an Jiaotong University in China in 2010, after training in machine learning and bioinformatics. He joined the bioinformatics core in Northwestern University as a postdoc from 2011 to 2013, where his research focused on statistical and computational analysis of next-generation sequencing data, including nucleosome-positioning, histone methylation, and protein-binding sites analysis. He joined the Division of Computational Genetics at Albert Einstein College of Medicine in 2013 as a research fellow. Since then, he has been working on human aging. By integrating biological networks and the human genetic variation together, he is looking for clues to the basic mechanisms of aging and the evolution of sub-network of aging genes.

Dr. Wen Zhang obtained his PhD in computational science from Shanghai University in 2009. Afterward, he studied at Michigan Technological University where he got his second MS degree, in bioinformatics, in 2012. After working at Baylor College of Medicine for one year, he joined Prof. Zhengdong Zhang's laboratory at Albert Einstein College of Medicine in 2013, as a postdoctoral fellow working on bioinformatics and computational genetics. He is currently analyzing high-throughput data from human genome sequencing and genome-wide association studies.

Dr. Rubén Nogales-Cadenas is a scientific researcher at the Albert Einstein College of Medicine in New York. He has more than eight years of research experience in Bioinformatics and Computational Biology. His scientific activity is mainly focused on the integration of biological data and computational resources for genomics research of the immune system, cancer, and drug discovery. He is interested in a wide range of research topics, including the development and implementation of new data analysis methodologies and their application to real biological problems, always taking advantage of many different types of biological information from experimental, functional, structural, regulatory, and pharmacological data. Dr. Nogales-Cadenas has published 18 articles in international well-recognized journals.

Dr. Jhih-Rong Lin received his PhD degree in computer science at University of South Carolina in 2013 with the thesis on computational analysis of protein sorting signals and localization. In 2014, he joined the Genetics Department of Albert Einstein College of Medicine as a research fellow. Currently, he is working on prediction of causal genes and causal variants of complex diseases. His main research interest focuses on the development of computational methods for human genetic analysis. As the first author of three journal articles in the field of bioinformatics, he has developed free software tools implementing his computational methods for public access.



Ying Cai obtained her MS degree in Medical Genome Sciences from the University of Tokyo, where she studied miRNA expression in adult T-cell leukemia. As a PhD candidate in the Department of Genetics at Albert Einstein College of Medicine, she is currently working in the field of bioinformatics, focusing on the analysis of mRNA expression profile of breast cancer.

Prof. Zhengdong D. Zhang is an assistant professor in the Department of Genetics at Albert Einstein College of Medicine. His research interests are computational genomics and systems biology of complex human diseases, focusing on algorithm development, data integration, and software implementation (visit www.zdzlab.org for more information). He participated and played an active role in some of the most notable international genome projects. He has investigated human functional genomics on different levels—from single genes, to gene families, and to the whole genomes—with an integrative approach drawing form molecular biology, statistics, and computational biology. At the Baylor College of Medicine Human Genome Sequencing Center, as part of the Rat Genome Project, he performed a comparative analysis of the nuclear receptor family in the human, mouse, and rat. At Yale University, as part of the ENCODE and the 1000 Genomes Projects, he developed software pipelines to process microarray and high-throughput sequencing data and carried out a detailed statistical analysis of the genomic distribution and correlation of transcriptional regulatory elements in the ENCODE regions. At Albert Einstein College of Medicine, his research team recently developed two computational frameworks of data integration to identify risk genes of complex diseases. Using a novel framework of integrated post-GWAS analysis, they identified distal causal genes of complex human diseases linked to GWAS signals through nearby regulatory elements such as enhancers. In another study, they developed a network-regularized logistic regression method. Using it to analyze case–control sequencing data, they identified and prioritized risk genes of complex human diseases. He received the NIH Career Development Award from the National Library of Medicine and the New Scholar Award from the Ellison Medical Foundation for his scientific and medical research.