Received July 12, 2020, accepted July 19, 2020, date of publication August 3, 2020, date of current version August 14, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3013666

Disease Cluster Detection and Functional Characterization

WEI GUO[1,2], TAO ZENG[3], TAO HUANG[4], AND YU-DONG CAI[1]

[1]School of Life Sciences, Shanghai University, Shanghai 200444, China
[2]Key Laboratory of Stem Cell Biology, Shanghai Jiao Tong University School of Medicine (SJTUSM) & Shanghai Institutes for Biological Sciences (SIBS), Chinese Academy of Sciences (CAS), Shanghai 200025, China
[3]Shanghai Research Center for Brain Science and Brain-Inspired Intelligence, Shanghai 201210, China
[4]Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, Shanghai 200031, China

Corresponding authors: Tao Huang (tohuangtao@126.com) and Yu-Dong Cai (cai_yud@126.com)

This work was supported in part by the Shanghai Municipal Science and Technology Major Project under Grant 2017SHZDZX01; in part by the National Key Research and Development Program of China under Grant 2018YFC0910403; in part by the National Natural Science Foundation of China under Grant 31701151; in part by the Shanghai Sailing Program under Grant 16YF1413800; and in part by the Youth Innovation Promotion Association of Chinese Academy of Sciences (CAS) under Grant 2016245.

ABSTRACT Mechanisms underlying human diseases have been revealed with the development of molecular biology. The underlying molecular basis of disorders is valuable in prevention, diagnosis, and treatment. Decade-long efforts have been devoted to investigating disease–gene association through positional cloning of disease genes and genome-wide association studies. In particular, correlations among different diseases have been discovered by many clinical cases. The shared disease-associated genes may help reveal the intrinsic relationship in the genetic level, provide an access to evaluate disease similarity, and establish a human disease network. Although many methods have been proposed to measure disease similarity, they only consider the genes or functions directly annotated to diseases but ignore the interactions among genes or functions. These interactions cause deficiency in disease classification. Basing on network-based disease module, we presented a systematic research to further investigate the relationship among different human diseases and explore whether this correlation depends on the functions of corresponding disease genes. On the one hand, a disease clustering based on the separation score between diseases is applied to divide 299 diseases into 15 relatively separated disease clusters. On the other hand, an optimal clustering scheme discriminating 15 disease clusters was learned based on disease-associated genes, their GO terms, and KEGG pathways annotations. The detected key signatures showed the highest relevance to distinguishing distinct disease clusters and represented the essential functions in corresponding pathogenesis. This study provides a novel approach to predict the network and function characteristics and reveals the functional essence of diseases.

INDEX TERMS Disease cluster, feature selection, network embedding, K-means.

I. INTRODUCTION

By the early 1900s, the Mendelian law of inheritance described the pattern that some traits can be transmitted from one generation to another. This pattern contributes to building the linkage between phenotypes and genes. Gene mutations showed strong relevance to human diseases, particularly for genetically inherited disorders. With the development of molecular biology, mechanisms underlying human diseases have been revealed. Biological processes caused by gene dysfunction result in various diseases [1], [2]. Uncovering the underlying molecular basis is valuable in the prevention, diagnosis, and treatment of diseases. Decade-long efforts have been devoted to investigate disease–gene association through the positional cloning of disease genes and genome-wide association studies [3]. Since restriction fragment length polymorphisms (RFLPs) were initially proposed to construct the genetic map in human [4], positional cloning has become accurate, and many disorders have been traced to a specific region in genome [5], [6]. As summarized in the Online Mendelian Inheritance in Man (OMIM) database, more than 3000 human diseases with known related genes have been documented, and these diseases-gene associations greatly contribute to the understanding of disease pathogenesis [7].

The correlation among different diseases has been discovered by many clinical cases. As a complex and highly prevalent disorder, metabolic syndrome is involved in the

The associate editor coordinating the review of this manuscript and approving it for publication was Quan Zou.

pathogenesis and progression of prostatic diseases as indicated by epidemiological data [8]. Diabetes is viewed as a high risk for heart diseases because the high glucose level in the bloodstream can damage the arteries [9]. In addition, diabetes mellitus is the most common cause of neuropathy, and the combination of metabolic and ischemic mechanisms leads to severe neuropathy [10]. In addition to such connection at the phenotype level, genetic factors might partially explain the co-occurrence of particular diseases. Vascular endothelial growth factor (VEGF) plays crucial role in the pathogenesis of diabetes, and the polymorphism of its gene is substantially correlated to neuropathy, especially diabetic neuropathy; this finding suggests the potential pathological contribution of VEGF to disease complication [11]. These shared disease-associated genes may help reveal the intrinsic relationship in the genetic level, provide an access to evaluate the disease similarity, and establish the human disease network.

Many methods have been proposed to measure disease similarity. The most simple and direct approach is to cluster diseases based on their phenotypic similarity through clinical observation. Text mining-based method is popularly applied in the classification of numerous human phenotypes. Phenotypes refers to the deviation from normal morphology and reflects the biological representation correlated to diseases [12]. For instance, a text mining approach was developed based on over 5000 human phenotypes in the OMIM database, and the built phenomap may be used to predict candidate genes for diseases [13]. Masino *et al.* calculated the relationship between phenotypes depending on the information regarding their lowest common-ancestor in human phenotype ontology, in which human phenotypic abnormalities are described in structured and unified vocabularies [14]. An ontology-based technique also exhibits excellent performance in measuring disease similarity. Sachin *et al.* estimated disease similarity by using an ontological metric to measure semantic similarity between Gene Ontology (GO) processes associated with diseases [15]. However, these existing methods only consider the gene or function directly annotated to disease but ignore their interactions, which can cause a deficiency of disease classification.

Genes related to similar diseases display a high likelihood of physical interactions between their protein products and a high similarity in the expression pattern, thus supporting the existence of distinct disease-specific functional gene modules [16]. A blueprint of the human interactome was presented by compiling 141,296 physical interactions among 13,460 experimentally documented proteins [17]. The disease module was proposed to represent the network of disease-related genes and their interactions. The distance of two diseases can be calculated based on the separation degree of disease module in the network [17]. Basing on this network-based disease module, we conducted a systematic research to further investigate the relationship among different human diseases and explore its dependence on the functionals of corresponding disease genes. A total of 299 diseases and 2436 disease-associated genes were included in our analysis. First, disease clustering was performed based on the separation score between diseases. The 299 diseases were then divided into 15 relatively separated clusters as supported by literature to extend our understanding about disease mechanism and their potential relationship. Second, the 15 clusters were deemed as 15 class labels, which were assigned to each disease. Each disease was represented by features derived from a protein-protein interaction (PPI) network, gene ontology (GO) terms and KEGG pathways. Features and labels were analyzed by several machine algorithms to extract essential GO terms and KEGG pathways. These signatures with the most relevance to distinguish distinct disease clusters can represent the essential functions in corresponding pathogenesis. Our study provides a novel approach to predict the network and function characteristics and contributes to revealing the functional essence of diseases.

II. METHODS

A. DATA

A network can be applied to predict disease–disease relationships by investigating the protein–protein interactions involved in the disease module [17]. The network-based distance of two diseases was calculated depending on the separation degree of disease-related gene modules. Two diseases with the overlapping gene modules will have shared clinical characteristics due to the same pathological processes. OMIM and PheGenI databases are the two main sources of such gene-disease annotations, and the integrated information was compiled based on the MeSH vocabulary [17]–[19]. The documented diseases with less than 20 associated genes were filtered out to improve the accuracy of our disease cluster analysis. A total of 299 diseases involving 2436 associated genes were included in the computation and analysis.

B. DISEASE CLUSTERING

On the basis of the network-based disease module, an iterative clustering algorithm was adopted to divide the 299 diseases into different categories. Calculation was conducted using hierarchical clustering with a bottom-up strategy and the following steps:

(1) Find the pair of diseases (or disease clusters) with strongest association measured by [17], and put the two diseases into one cluster;

(2) Re-calculate the associations between diseases in one disease cluster and diseases outside this cluster, and use the average association as the new association between of diseases (or disease clusters);

(3) Repeat the above calculation, until the number of disease clusters achieve a given value;

In this work, we set the number of disease clusters from 2 to 20, and determine the best cluster number is 15, which is manually judged based on the biological significance of disease clusters and applied to give the class label of particular disease cluster.

C. DISEASE FEATURE EXTRACTION AND SELECTION

1) FEATURE REPRESENTATION

For each disease, its representation features were extracted for every disease-associated gene as two categories. One is using node2vec (https://snap.stanford.edu/node2vec/) [20] on the STRING PPI network [21] to get 500 network embedding features for every disease-associated gene, and the other is using functional enrichment on the direct PPI-neighboring genes to obtain 297 KEGG [22] features and 20618 GO [23] features for each disease-associated gene. Finally, each disease-associated gene has 25915 features in total. The average of all disease-associated genes' features was used as the disease features for a particular disease.

2) BORUTA FOR FEATURE FILTERING

Many features were used to represent each disease-associated gene, but not all were discriminable for different diseases or clusters. Thus, Boruta feature filtering [24] was applied to remove features non-relevant to target outputs. This method works in a wrapper manner and is based on random forest including mainly three calculation steps: the production of new shuffled data from original training data, the calculation of importance score for each feature on each shuffled data, and the selection of real features with remarkably high importance scores in the training data. Finally, the important features are filtered after a few iterations of such three computational steps. The present study used the Boruta codes downloaded from https://github.com/scikit-learn-contrib/boruta_py.

3) MINIMUM REDUNDANCY MAXIMUM RELEVANCE (mRMR) FOR FEATURE RANKING

mRMR [25]–[27] selects informative features on the basis of two assumptions: those with minimum redundancy among themselves and those with maximum relevance with class labels. The features satisfying the two assumptions simultaneously are also chosen using mutual information. The mRMR program used in this study was retrieved from http://home.penglab.com/proj/mRMR/index.htm.

4) INCREMENTAL FEATURE SELECTION (IFS)

IFS [28] can iteratively determine the best number of selected features in the order or ranked features. At first, IFS produces a series of feature subsets from the ranked features. The first feature subset includes the top-ranked one feature, and the second includes the top-ranked two features. A series of clusters on the data is produced with each feature subset, where the performance of each K-means clustering is evaluated by rand index. Finally, the feature subset with the highest clustering performance is selected as the optimum.

K-means and rand index evaluation. K-means is one unsupervised clustering method on data $X = [x_1, x_2, \ldots, x_N]$ that aims to divide N samples into k clusters $C = \{C_1, C_2, \ldots, C_k\}$ by minimizing the subsequent loss function:

$$E = \sum\nolimits_{i=1}^{k} \sum\nolimits_{x \in C_i} \|x - \mu_i\|^2, \tag{1}$$

where μ_i is the central point of cluster C_i such as:

$$\mu_i = \frac{1}{|C_i|} \sum\nolimits_{x \in C_i} x \tag{2}$$

Rand index is an evaluation measurement for clustering on N samples. For a clustering, let C is the true sample clusters (e.g., one cluster with one prior-known class label), K is the calculated clusters, x represents the number of pairs of samples where two samples have same cluster alignment in C and K, and y represents the number of pairs of samples where two samples have different cluster alignment in C and K. Rand index was then employed to evaluate the consistency between C and K by using the following calculation:

$$RI(C, K) = \frac{x + y}{\binom{N}{2}} \tag{3}$$

III. RESULTS AND DISCUSSION

In this study, we used several computational methods to uncover the functional differences between various diseases. The entire procedures are illustrated in **Figure 1**.

A. DISEASE CLUSTER IDENTIFICATION AND CHARACTERIZATION

An iterative algorithm depending on the association data of 299 diseases was employed to cluster them into 15 disease clusters (**Supplementary file 1**). Each disease cluster was assigned a class label to ensure that the machine learning approaches can be used to further extract the essential biological features and distinguish each disease cluster. The following two kinds of features were first produced on the basis of network function: (1) 500 features extracted by node2vec from STRING PPI network and (2) 297 KEGG features and 20681 GO features extracted by enrichment on disease-associated genes. Boruta feature selection was used to filter non-informative features, thus leaving 1731 features. mRMR was applied to rank these 1731 relevant features, and IFS combined with K-means was used to determine the best number of features and corresponding clustering scheme based on the ranked feature list, where the performance was evaluated by Rand index. The Rand index was able to achieve the best score (0.7150) when the 65 top-ranked features were used for K-means (**Figure 2 and Supplementary file 2**), including 58 GO features, 3 KEGG features, and 4 network features (**Supplementary file 3**).

Abnormality of gene regulation is the cause of human diseases. Therefore, genes that play roles in certain disease must be identified to reveal the pathogenesis. Evidence has revealed that a disease is rarely caused by an aberration of single gene and mostly the consequence of interaction of multiple molecular activities [29], [30]. Therefore, a disorder is a complex and multi-step biological process involving various gene functions. Studies on human diseases have discovered many relevant genes by animal experiments or genome sequence analysis. Many databases were

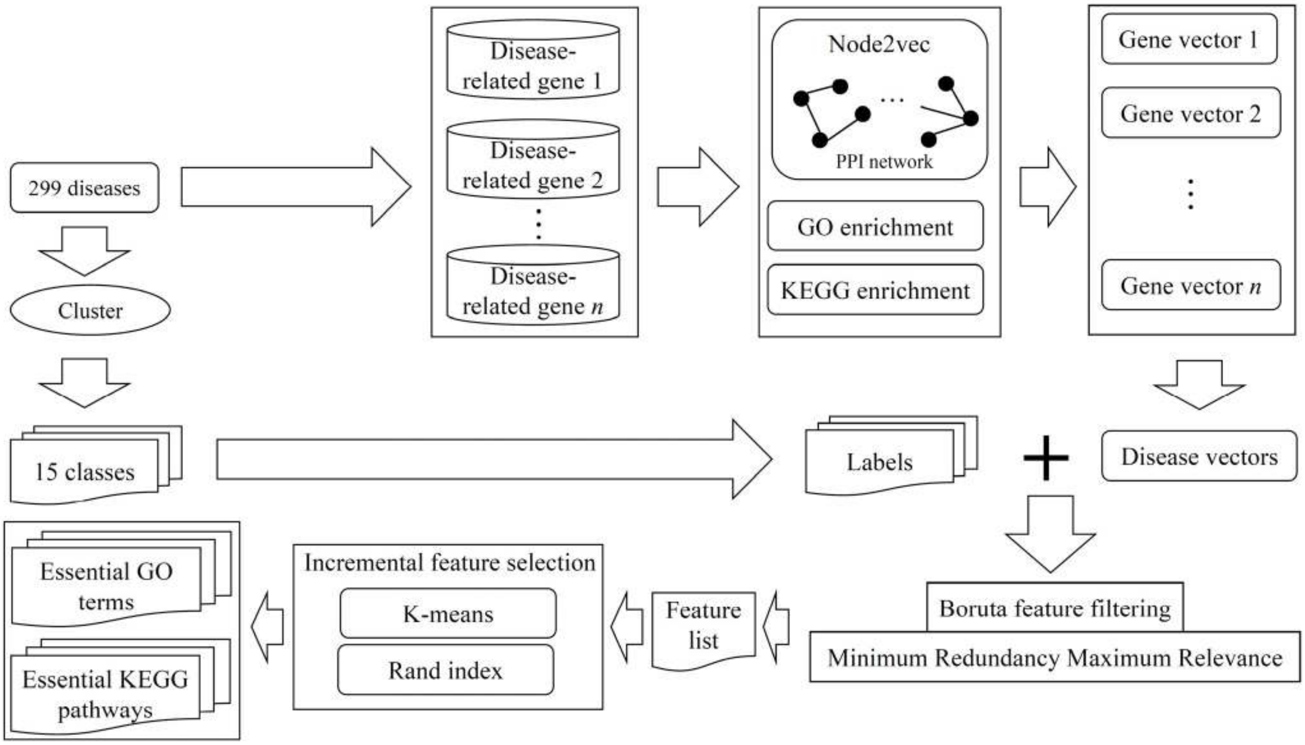

FIGURE 1. Entire analysis procedures. All 299 diseases are first clustered into 15 classes. Then, each disease is encoded into a vector, which is refined from the vectors of its related genes. The gene vector is derived from the PPI network, GO and KEGG enrichment. Finally, the disease vectors together with its labels are analyzed by two feature selection methods, yielding a feature list. The list is fed into the incremental feature selection method, incorporating K-means, to extract essential GO terms and KEGG pathways.

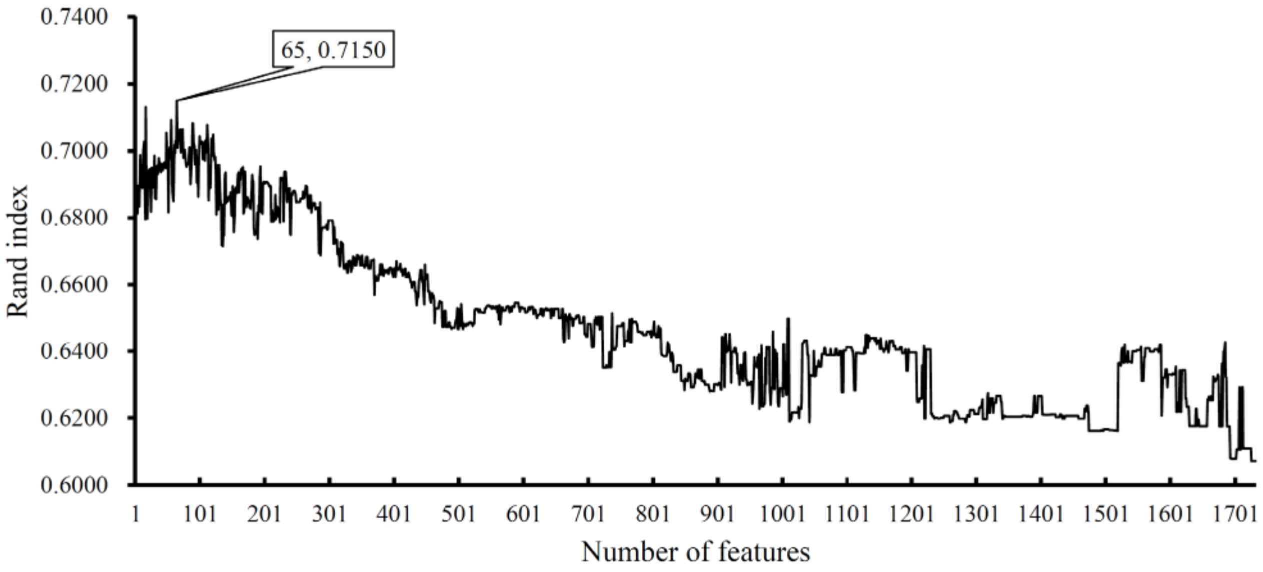

FIGURE 2. IFS curve with K-means on different number of features.

established with information on human genes and corresponding disorders, e.g., OMIM database. Gene–disease annotations greatly contribute to the understanding of disease etiology and the appropriate treatment design for rare clinical symptoms. In addition to revealing the molecular essence of disease, the relationships between genes and phenotypes can also aid in evaluating disease similarity on the basis of shared pathogenic genes. A disease correlation network that measures disease similarity extends the knowledge of disease mechanism, suggesting that diagnostic or

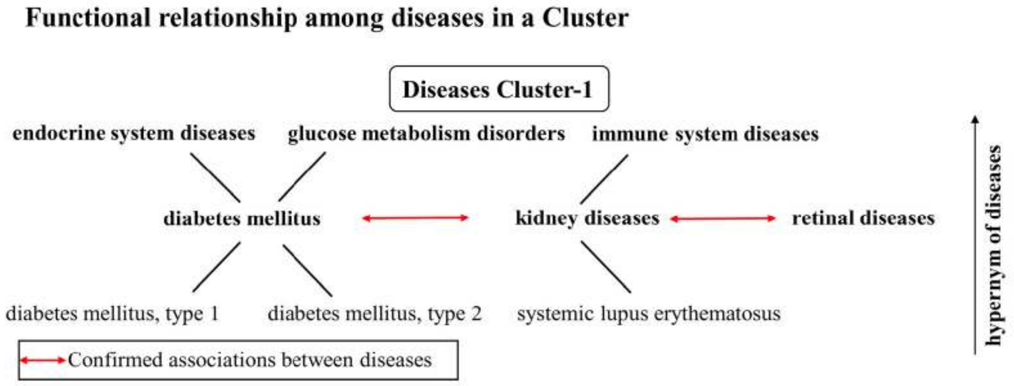
Functional relationship among diseases in a Cluster
Diseases Cluster-1
endocrine system diseases
glucose metabolism disorders
immune system diseases
diabetes mellitus
kidney diseases
retinal diseases
diabetes mellitus, type 1
diabetes mellitus, type 2
systemic lupus erythematosus
Confirmed associations between diseases
hypernym of diseases

FIGURE 3. Brief example of functional relationship among diseases in Cluster-1. By clustering analysis, diabetes mellitus, kidney diseases, retinal diseases and their hypernyms were assigned into cluster-1. The confirmed pathological associations between these diseases support the reasonability.

therapeutic approaches that can be appropriated from one disease to other similar types. In this study, a disease clustering analysis was performed based on the separation score between two diseases, and the 299 diseases were divided into 15 relatively separated clusters. This result suggests the potential relationship among the diseases in the same cluster and reveals the functional diversity among different disease clusters.

B. FUNCTIONAL RELATIONSHIP AMONG DISEASES SHARED IN A SAME CLUSTER

According to the results, disease clustering is reasonable. Diseases with similar text description exhibited close distance and mostly gathered in the same cluster in our analysis. For example, diabetes mellitus types 1 and 2 are naturally classified into the same cluster (cluster-1) because they are caused by the defects in insulin secretion, insulin action, or both [31]. Meanwhile, glucose metabolism disorders, which represents a class of diseases involved in the abnormal chemical reactions of glucose, were also assigned to cluster-1. Glucose metabolism is viewed as the hypernym of diabetes, and this finding is consistent with the classification into one disease cluster. In addition to the aggregation, an intrinsic connection was found among different types of diseases in the same cluster. Several autoimmune diseases, including rheumatoid arthritis, demyelinating autoimmune diseases, systemic lupus erythematosus, and their superior disease classes, were assigned to the cluster of diabetes. Insulin-dependent diabetes is also an autoimmune disorders that increasing inflammatory cells can be seen in the islets of Langerhans, and beta cell lesion is induced by the antibodies reaction and the following complement-dependent cytotoxicity [32], [33]. In clinical practice, systemic lupus erythematosus often coexist with diabetes, implying their potential relationship [34]. These findings supported the relevance of classification for the pathogenesis of aforementioned diseases and confirmed their assignment into one category (**Figure 3**).

Cluster-2 consisting of 18 different diseases also showed a degree of aggregation. In this disease cluster, 7 diseases clearly belong to the infection-related diseases including bacterial infections, mycoses and virus diseases. Strikingly, we noticed that the bile duct tract diseases and liver diseases were also assigned to this cluster. Liver and gallbladder disorders are closely associated that obstructed bile flows from liver into gallbladder will cause inflammation within liver [35]. Acute hepatitis or cirrhosis due to virus infection can result to cholestasis [36]. Thus, certain potential linkages among infection, inflammation, liver and gallbladder diseases was built, reflecting a specific pathological characteristic in cluster-2.

Cluster-3 is the largest disease cluster consisting of 169 various diseases. The relatively chaotic phenomenon in cluster-3 may be due to the limited maximum number of groups. Diseases with no association with other 14 clusters would be all assigned to this cluster. Although this big cluster cannot be briefly summarized, some interesting aggregations that many cancer-related diseases such as breast neoplasms, renal carcinoma and leukemia have gathered. This finding may imply the complex mechanism of cancer development, involvement of numerous biological processes, and potential association with various disorders.

Disorders of sex development and urogenital abnormalities were assigned to cluster-4, and both diseases showed association with the deletions of regions on 10q26, indicating their genetic relationship [37]. Cluster-5 showed a close association between amino acid metabolism and pigmentation disorders. L-tyrosine serves as the starting precursor in the synthesis of melanin, and the increased melanin level is the direct cause of hyperpigmentation [38]. The deficiency of sulfur amino acid also play roles in retinal pigmentary degeneration according to clinical studies [38]. These findings elucidate the pathological relationship of the two disease classes and confirm the reliability of our clustering results. Anemia, blood coagulation disorders, and blood platelet disorders were assigned to cluster-6 and belonged to hematologic disease class, reflecting an efficient clustering outcome. Another efficient clustering was found in cluster-9 consisting of seven diseases such as obesity, overweight, and overnutrition, all of which are related to nutrition disorders. In addition, brain disease and lipid metabolism disorders were assigned to the same cluster-10. Tissues in central nervous system have a high lipid concentration and deregulated lipid metabolism, which play a crucial role in brain injuries and disorders [39]. Lipid defect is relevant in Alzheimer disease, and membrane phospholipid degradation is a main pathogen mechanism of Alzheimer disease [40], [41]. Finally, in cluster-11, cardiomyopathy-related diseases and muscular disorders gathered in this same cluster. Given that cardiomyopathy is caused by the dysfunction of heart muscle, their close relationship is reasonable.

Overall, diseases in the same cluster showed some degree of relevance either directly or indirectly. This result validated the reliability of our clustering and confirmed the hypothesis that disease is the external representation of dysfunctional genes. In addition, this clustering analysis also indicated several potential correlations in diseases that have not been previously discovered. For example, a rare inherited disorder called lysosomal storage diseases was assigned to cluster-10 that contains brain diseases and lipid metabolism disorders. This finding suggested lysosomal function might participate in multiple processes of brain injuries. Our clustering analysis can help contribute to further understanding about the genetic essence of diseases and reveal the potential association among different types of diseases.

C. KEY FUNCTIONS ASSOCIATED WITH DISCRIMINATION OF DISEASE CLUSTERS

Considering that genes are the representors of biological functions, the functional essence of diseases was also explored. Each disease-associated gene was characterized by its enrichment values of GO terms and KEGG pathways to capture some concrete descriptions of functions corresponding to particular diseases. Protein encoded by each disease-associated gene was mapped into the protein–protein interaction network retrieved from STRING, thus providing a supplemental feature for the characterization of a given gene. According to such functional features from GO annotations and KEGG pathways, an optimal clustering scheme was built to assign any given disease into corresponding disease-cluster. This step provided an effective and novel approach to predict the characteristic of unknown diseases and aid in diagnosis and treatment. The most weighted features indicated the key functions in specific disease cluster / type and represented the functional essence of disorders. Existing evidence was evaluated to find the relevance between the disease-cluster and its related GO terms or KEGG pathways and validate the performance of the prediction model. The top-ranked features were selected as examples to extend the discussion because of their best ability for the discrimination of disease-clusters and close association with diseases.

The most discriminative feature by our computation is GO:0071976, which refers to the biological process of cell gliding. This GO term was substantially enriched in the disease cluster-2 and cluster-11. Diseases in cluster-2 are related to microbial infection, especially bacterial infection. Gliding is one of common motility types in prokaryotic cells and plays a role in adhesion and migration [42] and therefore may affect the contact between bacteria and host cell. GO:0071976 also influences the movement of host cells and often co-occurs with the biological process of T cell meandering migration and phagocytosis, resulting in the alteration of host immunity and consequent bacterial infection. An early study reported that Listeria-infected cells promote the formation of actin filaments, which may be important for the intracellular movement of Listeria bacteria [43]. Gene *MYO1G*, one of representative genes involved in cell gliding, encodes a plasma membrane-associated myosin and acts as a regulator of T-cell migration by generating membrane tension and enhancing pathogen detection and eradication [44]. These findings suggest that GO:0071976 plays crucial role in the motility of microbes and host cells and therefore has a close association with infection-related diseases. In addition, myosin is linked to hypertrophic cardiomyopathy and dilated cardiomyopathy by genotype-phenotype correlation study [45]. Elevated expression of *Myo1g* was detected in mouse phenotype of muscle atrophy, indicating the potential role of *MYO1G* in muscular disorders [46]. These findings built a close linkage between aberrance of myosin and cardiomyopathy-related diseases and confirmed that GO:0071976 is enriched in the disease cluster-11 related to cardiomyopathies and muscular disorders **(Figure 4)**.

GO:0060709 showed an enrichment tendency in cluster-4 and cluster-13, indicating that the related functions may play crucial role in these two types of diseases. GO:0060709 refers to the biological process of glycogen cell differentiation involved in embryonic placenta development. Given that this biological process is important in embryonic development, impaired function in GO:0060709 may result in development disorders, such as illnesses related to sex development and urogenital abnormalities observed in cluster-4. As a representative gene of GO:060709, gene *AKT1* displays strong relevance to various development impairments such as delayed bone development and severe growth deficiency [47], [48]. *AKT1* is also a critical mediator of growth factor-induced neuronal survival in the developing nervous system [49]. The only member in cluster-14 is epilepsy, a central nervous system disorder in which brain activity becomes abnormal. Enhanced phosphorylation of neuronal *AKT1* was observed in epilepsy, indicating that *AKT1* may be implicated in epilepsy pathogenesis [50]. Epilepsy or similar nervous-related diseases can be identified based on the enrichment of GO:060709.

A relatively high enrichment score of GO:0035722 was found in the disease cluster-9, suggesting the potential role of GO:0035722 in obesity-related diseases. GO:0035722 refers to the biological process of interleukin-12-mediated signaling pathway. A correlation was found between IL-12 and obesity. Serum IL-12 was remarkably higher in overweight and obese individuals than in normal weight controls as indicated by a clinical study [51]. Pro-inflammatory cytokines such as IL-12 may play role in mediating insulin response,

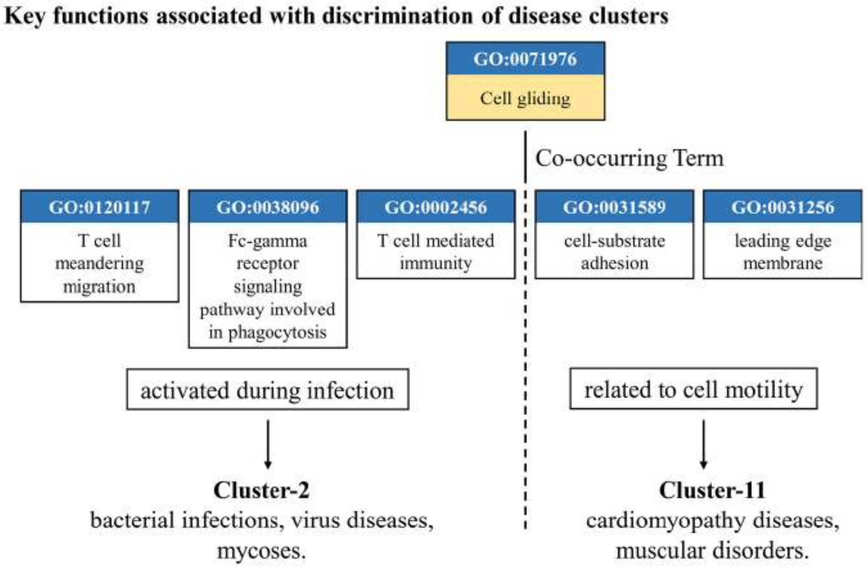
Key functions associated with discrimination of disease clusters
GO:0071976
Cell gliding
Co-occurring Term
GO:0120117
T cell meandering migration
GO:0038096
Fc-gamma receptor signaling pathway involved in phagocytosis
GO:0002456
T cell mediated immunity
GO:0031589
cell-substrate adhesion
GO:0031256
leading edge membrane
activated during infection
related to cell motility
Cluster-2
bacterial infections, virus diseases, mycoses.
Cluster-11
cardiomyopathy diseases, muscular disorders.

FIGURE 4. GO:0071976 is the key discriminative feature for classifying Cluster-2 and Cluster-11. The biological process of cell gliding is closely associated with processes of immune activation and cell motility, which are much crucial in infection diseases (Cluster-2) and muscular disorders (Cluster-11).

and elevated IL-12 is involved in the development of obesity-induced insulin resistance [52]. GO:0052331, which represents the biological process of hemolysis in other organism involved in symbiotic interaction, was highly enriched in cluster-6. Diseases in cluster-6 such as hemolytic anemia and blood coagulation disorders are closely related to the dysfunction of hemolysis [53]. Another gene ontology term GO:1990831 was found significantly enriched in cluster-14, which consists of neoplasms and sarcoma. GO:1990831 refers to the biological process of cellular response to carcinoembryonic antigen (CEA). CEA is a protein normally found in low levels in the blood, and its increased level are detected in cancer [54]. Thus, an activated response to CEA indicates tumorigenesis, thus confirming our prediction that the high enrichment of GO:1990831 serves as the indicator to sarcoma or neoplasms.

IV. CONCLUSION

An optimal clustering scheme was constructed based on the networks and functions of disease-related genes, which can provide a function classification for each disease, thus deepening our understanding on disease pathogenesis. The most relevant GO terms or KEGG pathways were identified and confirmed to play crucial roles in corresponding disease-cluster on the basis of wide literature review. These biological functions may partially represent the functional essence of diseases and contribute to further research on disease mechanism.

ACKNOWLEDGMENT

Support information: Supplementary file 1: Clusters of diseases; Supplementary file 2: Performance of IFS with K-means on different number of features; Supplementary file 3: Optimal features indicating essential GO and KEGG functions.

REFERENCES

[1] Y. Chen *et al.*, "Variations in DNA elucidate molecular networks that cause disease," *Nature*, vol. 452, no. 7186, pp. 429–435, Mar. 2008.

[2] D. Altshuler, M. J. Daly, and E. S. Lander, "Genetic mapping in human disease," *Science*, vol. 322, no. 5903, pp. 881–888, Nov. 2008.

[3] J. N. Hirschhorn and M. J. Daly, "Genome-wide association studies for common diseases and complex traits," *Nature Rev. Genet.*, vol. 6, no. 2, pp. 95–108, Feb. 2005.

[4] D. Botstein, R. L. White, M. Skolnick, and R. W. Davis, "Construction of a genetic linkage map in man using restriction fragment length polymorphisms," *Amer. J. Hum. Genet.*, vol. 32, p. 314, 1980.

[5] J. F. Gusella, N. S. Wexler, P. M. Conneally, S. L. Naylor, M. A. Anderson, R. E. Tanzi, P. C. Watkins, K. Ottina, M. R. Wallace, A. Y. Sakaguchi, A. B. Young, I. Shoulson, E. Bonilla, and J. B. Martin, "A polymorphic DNA marker genetically linked to Huntington's disease," *Nature*, vol. 306, no. 5940, pp. 234–238, Nov. 1983.

[6] L. K. Billings and J. C. Florez, "The genetics of type 2 diabetes: What have we learned from GWAS?" *Ann. New York Acad. Sci.*, vol. 1212, p. 59, Nov. 2010.

[7] J. Amberger, C. A. Bocchini, A. F. Scott, and A. Hamosh, "McKusick's online Mendelian inheritance in man (OMIM)," *Nucleic acids Res.*, vol. 37, no. 1, pp. D793–D796, 2009.

[8] C. De Nunzio, W. Aronson, S. J. Freedland, E. Giovannucci, and J. K. Parsons, "The correlation between metabolic syndrome and prostatic diseases," *Eur. Urol.*, vol. 61, no. 3, pp. 560–570, Mar. 2012.

[9] S. M. Haffner, *Coronary Heart Disease in Patients With Diabetes*. Waltham, MA, USA: Mass Medical Soc, 2000.

[10] G. Said, "Diabetic neuropathy—A review," *Nature Clin. Pract. Neurol.*, vol. 3, no. 6, pp. 331–340, 2007.

[11] J. Tavakkoly-Bazzaz, M. M. Amoli, V. Pravica, R. Chandrasecaran, A. J. M. Boulton, B. Larijani, and I. V. Hutchinson, "VEGF gene polymorphism association with diabetic neuropathy," *Mol. Biol. Rep.*, vol. 37, no. 7, pp. 3625–3630, Oct. 2010.

[12] P. N. Robinson, "Deep phenotyping for precision medicine," *Hum. Mutation*, vol. 33, no. 5, pp. 777–780, May 2012.

[13] M. A. van Driel, J. Bruggeman, G. Vriend, H. G. Brunner, and J. A. M. Leunissen, "A text-mining analysis of the human phenome," *Eur. J. Hum. Genet.*, vol. 14, no. 5, pp. 535–542, May 2006.

[14] A. J. Masino, E. T. Dechene, M. C. Dulik, A. Wilkens, N. B. Spinner, I. D. Krantz, J. W. Pennington, P. N. Robinson, and P. S. White, "Clinical phenotype-based gene prioritization: An initial study using semantic similarity and the human phenotype ontology," *BMC Bioinf.*, vol. 15, no. 1, p. 248, 2014.

[15] S. Mathur and D. Dinakarpandian, ''Finding disease similarity based on implicit semantic similarity,'' *J. Biomed. Informat.*, vol. 45, no. 2, pp. 363–371, Apr. 2012.

[16] K. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A. L. Barabási, ''The human disease network,'' *Proc. Nat. Acad. Sci. USA*, vol. 104, no. 21, pp. 8685–8690, 2007.

[17] J. Menche, A. Sharma, M. Kitsak, S. D. Ghiassian, M. Vidal, J. Loscalzo, and A.-L. Barabási, ''Uncovering disease-disease relationships through the incomplete interactome,'' *Science*, vol. 347, no. 6224, 2015, Art. no. 1257601.

[18] A. Hamosh, ''Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders,'' *Nucleic Acids Res.*, vol. 30, no. 1, pp. D514–D517, Jan. 2002.

[19] E. M. Ramos, D. Hoffman, H. A. Junkins, D. Maglott, L. Phan, S. T. Sherry, M. Feolo, and L. A. Hindorff, ''Phenotype–genotype integrator (PheGenI): Synthesizing genome-wide association study (GWAS) data with existing genomic resources,'' *Eur. J. Hum. Genet.*, vol. 22, no. 1, pp. 144–147, Jan. 2014.

[20] A. Grover and J. Leskovec, ''node2vec: Scalable feature learning for networks,'' presented at the 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, San Francisco, CA, USA, 2016.

[21] D. Szklarczyk, A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic, N. T. Doncheva, J. H. Morris, P. Bork, L. J. Jensen, and C. V. Mering, ''STRING v11: Protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets,'' *Nucleic Acids Res.*, vol. 47, no. D1, pp. D607–D613, Nov. 2018.

[22] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe, ''KEGG for integration and interpretation of large-scale molecular data sets,'' *Nucleic Acids Res.*, vol. 40, no. D1, pp. D109–D114, Jan. 2012.

[23] J. A. Blake *et al.*, ''Gene ontology consortium: Going forward,'' *Nucleic Acids Res.*, vol. 43, no. D1, pp. D1049–D1056, Jan. 2015.

[24] M. Kursa and W. Rudnicki, ''Feature selection with the Boruta package,'' *J. Stat. Softw.*, vol. 36, no. 11, pp. 1–13, 2010.

[25] H. Peng, F. Long, and C. Ding, ''Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[26] L. Chen, X. Pan, X. Hu, Y. Zhang, S. Wang, T. Huang, and Y. Cai, ''Gene expression differences among different MSI statuses in colorectal cancer,'' *Int. J. Cancer*, vol. 143, no. 7, pp. 1731–1740, Oct. 2018.

[27] X. Zhao, L. Chen, and J. Lu, ''A similarity-based method for prediction of drug side effects with heterogeneous information,'' *Math. Biosci.*, vol. 306, pp. 136–144, Dec. 2018.

[28] H. Liu and R. Setiono, ''Incremental feature selection,'' *Appl. Intell.*, vol. 9, no. 3, pp. 217–230, Nov./Dec. 1998.

[29] E. E. Schadt, ''Molecular networks as sensors and drivers of common human diseases,'' *Nature*, vol. 461, no. 7261, pp. 218–223, Sep. 2009.

[30] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, ''Network medicine: A network-based approach to human disease,'' *Nature Rev. Genet.*, vol. 12, no. 1, pp. 56–68, Jan. 2011.

[31] M. M. Funnell, T. L. Brown, B. P. Childs, L. B. Haas, G. M. Hosey, B. Jensen, M. Maryniuk, M. Peyrot, J. D. Piette, D. Reader, L. M. Siminerio, K. Weinger, and M. A. Weiss, ''Diagnosis and classification of diabetes mellitus,'' *Diabetes Care*, vol. 36, pp. S67–S74, Jan. 2013.

[32] J. Nerup and A. Lernmark, ''Autoimmunity in insulin-dependent diabetes mellitus,'' *Amer. J. Med.*, vol. 70, no. 1, pp. 135–141, Jan. 1981.

[33] J.-F. Bach, ''Insulin-dependent diabetes mellitus as an autoimmune disease,'' *Endocrine Rev.*, vol. 15, no. 4, pp. 516–542, Aug. 1994.

[34] S. Cortes, S. Chambers, A. Jerónimo, and D. Isenberg, ''Diabetes mellitus complicating systemic lupus erythematosus—Analysis of the UCL lupus cohort and review of the literature,'' *Lupus*, vol. 17, no. 11, pp. 977–980, Nov. 2008.

[35] C. G. Tag, S. Sauer-Lehnen, S. Weiskirchen, E. Borkham-Kamphorst, R. H. Tolba, F. Tacke, and R. Weiskirchen, ''Bile duct ligation in mice: Induction of inflammatory liver injury and fibrosis by obstructive cholestasis,'' *J. Vis. Exp.*, no. 96, Feb. 2015, Art. no. e52438.

[36] S. E. Davies, B. C. Portmann, J. G. O'grady, P. M. Aldis, K. Chaggar, G. J. M. Alexander, and R. Williams, ''Hepatic histological findings after transplantation for chronic hepatitis b virus infection, including a unique pattern of fibrosing cholestatic hepatitis,'' *Hepatology*, vol. 13, no. 1, pp. 150–157, Jan. 1991.

[37] T. Ogata, K. Muroya, I. Sasagawa, T. Kosho, K. Wakui, S. Sakazume, K. Ito, N. Matsuo, H. Ohashi, and T. Nagai, ''Genetic evidence for a novel gene(s) involved in urogenital development on 10q26,'' *Kidney Int.*, vol. 58, no. 6, pp. 2281–2290, Dec. 2000.

[38] M. Kosmadaki, A. Naif, and P. Hee-Young, ''Recent progresses in understanding pigmentation,'' *Giornale Italiano Dermatologia Venereologia, Organo Ufficiale, Societ Italiana Dermatologia Sifilografia*, vol. 145, no. 1, pp. 47–55, 2010.

[39] R. M. Adibhatla and J. Hatcher, ''Altered lipid metabolism in brain injury and disorders,'' in *Lipids in Health and Disease*. Dordrecht, The Netherlands: Springer, 2008, pp. 241–268.

[40] S. M. de la Monte and M. Tong, ''Brain metabolic dysfunction at the core of Alzheimer's disease,'' *Biochem. Pharmacol.*, vol. 88, no. 4, pp. 548–559, Apr. 2014.

[41] R. M. Nitsch, J. K. Blusztajn, A. G. Pittas, B. E. Slack, J. H. Growdon, and R. J. Wurtman, ''Evidence for a membrane defect in Alzheimer disease brain,'' *Proc. Nat. Acad. Sci. USA*, vol. 89, no. 5, pp. 1671–1675, 1992.

[42] K. F. Jarrell and M. J. McBride, ''The surprisingly diverse ways that prokaryotes move,'' *Nature Rev. Microbiol.*, vol. 6, no. 6, pp. 466–476, Jun. 2008.

[43] V. Zhukarev, F. Ashton, J. M. Sanger, J. W. Sanger, and H. Shuman, ''Organization and structure of actin filament bundles inListeria-infected cells,'' *Cell Motility Cytoskeleton*, vol. 30, no. 3, pp. 229–246, 1995.

[44] J. den Haan, N. Sherman, E. Blokland, E. Huczko, F. Koning, J. Drijfhout, J. Skipper, J. Shabanowitz, D. Hunt, V. Engelhard, and A. Et, ''Identification of a graft versus host disease-associated human minor histocompatibility antigen,'' *Science*, vol. 268, no. 5216, pp. 1476–1480, Jun. 1995.

[45] A. J. Marian and R. Roberts, ''The molecular genetic basis for hypertrophic cardiomyopathy,'' *J. Mol. Cellular Cardiol.*, vol. 33, no. 4, pp. 655–670, Apr. 2001.

[46] D. Bosnakovski, S. S. K. Chan, O. O. Recht, L. M. Hartweck, C. J. Gustafson, L. L. Athman, D. A. Lowe, and M. Kyba, ''Muscle pathology from stochastic low level DUX4 expression in an FSHD mouse model,'' *Nature Commun.*, vol. 8, no. 1, pp. 1–9, Dec. 2017.

[47] X.-D. Peng, ''Dwarfism, impaired skin development, skeletal muscle atrophy, delayed bone development, and impeded adipogenesis in mice lacking Akt1 and Akt2,'' *Genes Develop.*, vol. 17, no. >11, pp. 1352–1365, Jun. 2003.

[48] Z.-Z. Yang, O. Tschopp, M. Hemmings-Mieszczak, J. Feng, D. Brodbeck, E. Perentes, and B. A. Hemmings, ''Protein kinase Bα/Akt1 regulates placental development and fetal growth,'' *J. Biol. Chem.*, vol. 278, no. 34, pp. 32124–32131, Aug. 2003.

[49] A. Saito, ''Neuroprotective role of a proline-rich AKT substrate in apoptotic neuronal cell death after stroke: Relationships with nerve growth factor,'' *J. Neurosci.*, vol. 24, no. 7, pp. 1584–1593, Feb. 2004.

[50] L. Miles, H. M. Greiner, M. V. Miles, F. T. Mangano, P. S. Horn, J. L. Leach, J. H. Seo, and K. H. Lee, ''Interaction between Akt1-positive neurons and age at surgery is associated with surgical outcome in children with isolated focal cortical dysplasia,'' *J. Neuropathol. Experim. Neurol.*, vol. 72, no. 9, pp. 884–891, Sep. 2013.

[51] K. Suárez-Álvarez, L. Solís-Lozano, S. Leon-Cabrera, A. González-Chávez, G. Gómez-Hernández, M. S. Quiñones-Álvarez, A. E. Serralde-Zúñiga, J. Hernández-Ruiz, J. Ramírez-Velásquez, F. J. Galindo-González, J. C. Zavala-Castillo, M. A. De León-Nava, G. Robles-Díaz, and G. Escobedo, ''Serum IL-12 is increased in Mexican obese subjects and associated with low-grade inflammation and obesity-related parameters,'' *Mediators Inflammation*, vol. 2013, Feb. 2013, Art. no. 967067.

[52] H. Nam, B. S. Ferguson, J. M. Stephens, and R. F. Morrison, ''Impact of obesity on IL-12 family gene expression in insulin responsive tissues,'' *Biochimica Biophysica Acta-Mol. Basis Disease*, vol. 1832, no. 1, pp. 11–19, Jan. 2013.

[53] R. Znazen, H. Kaabi, S. Hmida, H. B. Abid, S. B. Tahar, I. Zammit, A. Hafsia, and K. Boukef, ''Detection of serum hemolysins in autoimmune hemolytic anemia,'' *Transfusion Clinique Biologique*, vol. 13, no. 6, pp. 341–345, Dec. 2006.

[54] R. H. Fletcher, ''Carcinoembryonic antigen,'' *Ann. internal Med.*, vol. 104, pp. 66–73, Jan. 1986.

WEI GUO received the B.S. degree in life science and technology from Wuhan University, in 2015. Since September 2015, he has been with the Laboratory of Molecular Genetics, Shanghai Jiao Tong University School of Medicine, and the Institute of Health Sciences, beginning his master's and Ph.D. education. His research interests include bioinformatics, microbiome, and cancers.

TAO ZENG received the B.S., M.S., and Ph.D. degrees from Wuhan University, Wuhan, China, in 2003, 2006, and 2010, respectively. Since 2013, he has been an Associate Professor with the Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences. He is currently with the Shanghai Research Center for Brain Science and Brain-Inspired Intelligence, Shanghai, China. His research interests include bioinformatics, network biology, computational biology, machine learning, and graph theory.

TAO HUANG received the B.S. degree in bioinformatics from the Huazhong University of Science and Technology, Wuhan, China, in 2007, and the Ph.D. degree in bioinformatics from the Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China, in 2012.

Since 2014, he has been an Associate Professor and the Director of the Bioinformatics Core Facility at the Institute of Health Sciences and the Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences. From 2012 to 2014, he was a Postdoctoral Fellow with the Department of Genetics and Genomics Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA. His research interests include bioinformatics, computational biology, systems genetics, and big data research. He has published over 100 articles. His works have been cited for over 3000 times with an H-index of 26 and an i10-index of 64. He has been a Reviewer of over 20 journals and an Editor/Guest Editor of 7 journals and books.

YU-DONG CAI has been a Professor of bioinformatics with the School of Life Science, Shanghai University, since 2015. His main research interests include systems biology and bioinformatics, such as protein–protein interaction, disease biomarkers prediction, drug-target interaction, and protein functional sites prediction. He has published over 200 peer-reviewed scientific articles, including invited reviews. His works have been cited for more than 7500 times, with H-index of 51. He is an Editorial Board Member of *Biochimica et Biophysica Acta-Proteins and Proteomics* (BBA) and *Biochemistry Research International*. He has been a Guest Editor of *Computational Proteomics, Systems Biology and Clinical Implications* and *Biochimica et Biophysica Acta–Proteins and Proteomics* (BBA).