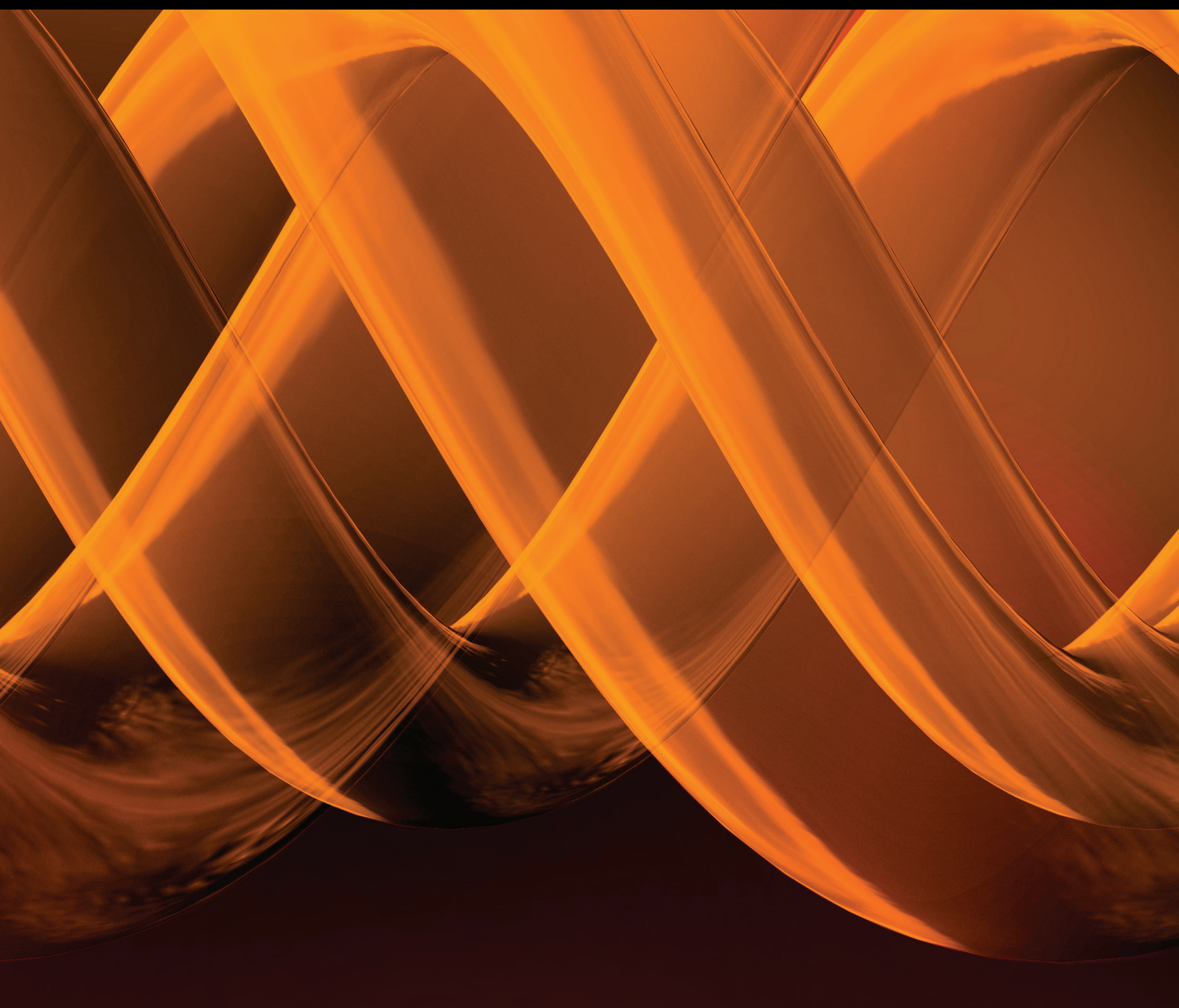
International Journal of Genomics

# Advancing Genomics for Drug Development and Safety Evaluation

Lead Guest Editor: Zhichao Liu
Guest Editors: Joshua Xu and Zhining Wen



# Advancing Genomics for Drug Development and Safety Evaluation

International Journal of Genomics

# Advancing Genomics for Drug Development and Safety Evaluation

Lead Guest Editor: Zhichao Liu
Guest Editors: Joshua Xu and Zhining Wen

Copyright © 2018 Hindawi. All rights reserved.

This is a special issue published in "International Journal of Genomics." All articles are open access articles distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

# Editorial Board


Jacques Camonis, France
Luigi Ceci, Italy
Maria Luisa Chiusano, Italy
Prabhakara V. Choudary, USA
Martine A. Collart, Switzerland
Giandomenico Corrado, Italy
Monika Dmitrzak-Weglarz, Poland
Antonio Ferrante, Italy
Marco Gerdol, Italy
João Paulo Gomes, Portugal
Soraya E. Gutierrez, Chile
M. Hadzopoulou-Cladaras, Greece
Sylvia Hagemann, Austria
Henry Heng, USA
Eivind Hovig, Norway
Hieronim Jakubowski, USA
B.-H. Jeong, Republic of Korea
Atsushi Kurabayashi, Japan
Sang Hong Lee, Australia
Julio Martin-Garcia, USA
Giuliana Napolitano, Italy
Corey Nislow, Canada
Michael Nonnemacher, USA
Ferenc Olasz, Hungary
Elena Pasyukova, Russia
Graziano Pesole, Italy
Giulia Piaggio, Italy
Ernesto Picardi, Italy
Mohamed Salem, USA
Wilfred van IJcken, Netherlands
Brian Wigdahl, USA
Jinfa Zhang, USA

# Contents

Hindawi
International Journal of Genomics
Volume 2018, Article ID 3126820, 2 pages
https://doi.org/10.1155/2018/3126820

*Editorial*

# Advancing Genomics for Drug Development and Safety Evaluation

**Zhichao Liu**,[1] **Joshua Xu**,[1] **and Zhining Wen**[2]

[1]*Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, U.S. Food and Drug Administration, Jeffersonville, IN, USA*
[2]*Sichuan University, Chengdu, China*

Correspondence should be addressed to Zhichao Liu; zhichao.liu@fda.hhs.gov

Received 4 February 2018; Accepted 4 February 2018; Published 23 May 2018

Copyright © 2018 Zhichao Liu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This special issue is focused on advancing genomics for drug development and safety evaluation. Gaining wide adoption in biomedical fields, genomic technologies have tremendously improved our molecular understanding of disease etiology and pathogenesis. Furthermore, the accumulated genomic datasets allow us to articulate new hypothesis, revisit and rethink the conventional paradigm for drug development and safety evaluation, and generate more efficient tools to promote biomedical researches.

Cancer genomics has made a great progress to uncover the underlying mechanism of tumorigenesis. Consequently, a lot of cancer genomic biomarkers have been developed for improving cancer diagnosis and prognosis. Questions arise on how the biomarkers developed from one study could be extrapolated to another. The advancement of machine learning technologies provides us a great opportunity to address these crucial questions such as model transfer and data integrity issues in the cancer field. R. Jing et al. proposed an ensemble method with voting protocols to integrate multiple machine learning algorithms to predict cancer outcome. It was indicated the proposed ensemble approaches could greatly improve the robustness and stability of cancer prediction modeling.

Genomic technologies such as next-generation sequencing provided an unprecedented resolution to better understand complex regulatory relationship among different genetic elements. The consortium efforts such as The Cancer Genomics Atlas (TCGA) not only covers diverse cancer subtypes but also includes a lot of genomic elements and events (i.e., miRNA, copy number variation, and DNA methylation). Accordingly, the approaches to integrate these genomic elements to decipher the complex regulatory relationship tailored to different cancer mechanisms are urgently needed. J. Xue et al. applied graphical lasso models (GLMs) to predict miRNA-mRNA relationship for three cancer types including acute myeloid leukemia (AML), breast invasive carcinoma (BRCA), and kidney renal clear cell carcinoma (KIRC) in TCGA. The results suggested that the proposed network approaches could improve the prediction performance to enrich more tumorigenesis-related miRNA-mRNA relationship.

Shift work is a common social issue due to the rapid pace of modern lifestyle. The disruption in circadian clock system leads to a lot of health concerns and increases the risk to develop serious diseases including sleep disorders, metabolic disorders, psychiatric disorders, and even cancers. S. Khan et al. summarized the shift work-related disorders and elaborated on the potential risk factors and mechanisms with a comprehensive literature survey. It is very interesting that some gene expressions and genetic variants were identified for playing an important role in shift work-related health disorders, which paves a way to further uncover the genetic contribution to shift work-related diseases and develops therapy to control and relieve the syndromes.

Protein classification based on organisms is a hot topic in microbiology. Considering huge amount of protein sequencing data that were generated in the past two decades, supplicated model development strategies and novel approaches are needed to categorize the proteins from different organisms. H.-B. Guo et al. developed novel fingerprints based on protein distribution densities in the LD space and implemented a machine learning framework to improve the

accuracy of protein organism classification. The proposed approach could be potentially applied to microbiome field and related disciplines.

A lot of drug candidates in the clinical trial failed due to unexpected adverse drug reactions (ADRs). ADR especially idiosyncratic adverse drug reaction (IADR) is difficult to study. Drug-induced myopathy as an IADR is unpredictable and dose independent. To better understand the causes for drug-induced myopathy, D. Li et al. developed a systematic approach to integrating different data profiles including chemical structure information, drug-protein relationship, side effects, and transcriptomic data profiles to identify the risk factors for drug-induced myopathy. This study sets a great example for fusing the genomic data with other types of data profiles to elucidate the hidden genotype-phenotype relationship. Furthermore, the key factors including structure alerts could be applied to develop predictive models for early detection and prevention of drug-induced myopathy.

Many aspects of genomics were not covered in this special issue. For example, the reproducibility of genomic studies, inconsistent results from different data analysis pipelines, and data storage are also of great importance for better utilization of genomic technology to promote and improve public health. We hope this special issue could serve as a trigger to stimulate the common interest in the community for advancing genomic technologies.

*Zhichao Liu*
*Joshua Xu*
*Zhining Wen*

Hindawi
International Journal of Genomics
Volume 2018, Article ID 9784161, 12 pages
https://doi.org/10.1155/2018/9784161

*Research Article*

# Classification of Complete Proteomes of Different Organisms and Protein Sets Based on Their Protein Distributions in Terms of Some Key Attributes of Proteins

**Hao-Bo Guo**,[1] **Yue Ma**,[1] **Gerald A. Tuskan**,[2] **Xiaohan Yang**,[2] **and Hong Guo**[1]

[1]*Department of Biochemistry and Cellular and Molecular Biology, University of Tennessee, Knoxville, TN 37996, USA*
[2]*Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 3783, USA*

Correspondence should be addressed to Xiaohan Yang; yangx@ornl.gov and Hong Guo; hguo1@utk.edu

Received 27 July 2017; Revised 12 November 2017; Accepted 20 November 2017; Published 4 March 2018

Academic Editor: Joshua Xu

Copyright © 2018 Hao-Bo Guo et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The existence of complete genome sequences makes it important to develop different approaches for classification of large-scale data sets and to make extraction of biological insights easier. Here, we propose an approach for classification of complete proteomes/protein sets based on protein distributions on some basic attributes. We demonstrate the usefulness of this approach by determining protein distributions in terms of two attributes: protein lengths and protein intrinsic disorder contents (ID). The protein distributions based on $L$ and ID are surveyed for representative proteome organisms and protein sets from the three domains of life. The two-dimensional maps (designated as fingerprints here) from the protein distribution densities in the LD space defined by $\ln(L)$ and ID are then constructed. The fingerprints for different organisms and protein sets are found to be distinct with each other, and they can therefore be used for comparative studies. As a test case, phylogenetic trees have been constructed based on the protein distribution densities in the fingerprints of proteomes of organisms without performing any protein sequence comparison and alignments. The phylogenetic trees generated are biologically meaningful, demonstrating that the protein distributions in the LD space may serve as unique phylogenetic signals of the organisms at the proteome level.

## 1. Introduction

Determination of complete genome sequences for a number of organisms has offered an unprecedented opportunity for biological community and transformed biology into a discipline that depends significantly on how to classify and interpret large-scale data sets and to extract biological insights from these data sets. The traditional ways of thinking and approaches from the pregenomic era (e.g., the sequence comparison/alignment and homology identification) are of fundamental importance in the postgenomic era. Nevertheless, new approaches based on some global features of omics data sets need to be explored in order to make classification and comparison of large-scale data sets easier. For proteomes, this may be achieved, for instance, through identification of key parameters or attributes of proteins and comparison of protein distributions within complete proteomes of different organisms or protein sets in terms of such parameters or attributes.

In this paper, we adapt this approach and use two parameters of proteins for the purpose of classifying complete proteomes of different organisms (for simplicity, proteomes) and protein sets: the length of protein amino acid (aa) sequence (protein length $L$ hereafter) and intrinsic disorder content (protein disorder ID hereafter). It had been proposed that the protein sizes, folding rates, and many other physical properties could be associated or even determined by $L$ [1, 2]. At the level of proteomes, previous studies have suggested that the eukaryotic proteomes may exhibit averagely longer $L$ compared to the prokaryotic

proteomes [3, 4], even though further analysis may still be necessary. The importance of intrinsically disordered proteins (IDPs) and protein regions (IDPRs) has been recognized [5–13], and it has been observed that relatively high contents of intrinsic disorders may exist for eukaryotic proteins than for prokaryotic proteins [14]. Moreover, proteins expressed in two eukaryotic organelles, chloroplasts and mitochondria, which evolved from cyanobacteria and alphaproteobacteria, respectively, seem to have a lower disorder content, on average, compared to nuclear-encoded proteins in their host eukaryotes [15]. Interestingly, it has been demonstrated that intrinsically disordered proteins are associated with a variety of human diseases [16, 17], including cancers [18, 19]. As a result, intrinsically disordered proteins have become important targets for drug design [20–25]. Thus, understanding intrinsically disordered proteins at the proteomic levels would be of considerable interest. The observations that the distributions of proteins in terms of ID and $L$ may be different for proteomes and for different protein sets suggest that such distributions may be used to classify proteomes of different organisms or protein sets. They may also be used in the future to help understand the properties of proteomes in different disease states, as there seems to be a wide variability of predicted disorder among different diseases [26]. It is interesting to see that a recent study revealed that the overall disorder fractions are positively correlated to the size of the proteomes (estimated by the total aa numbers) and that the disorder fractions of the proteomes of large bacteria (more than 2.5 M aa) are comparable to those of eukaryotes [27].

Here we analyze the protein distributions in terms of $L$ and ID from proteomes of different organisms across the three domains of life, collective data sets of organelles (plasmids, chloroplasts, and mitochondria), and the proteome data of two giant DNA viruses (termed giruses in literature). We noticed that the eukaryotic proteomes do not always exhibit averagely longer proteins than the prokaryotic proteomes. Our observation on protein disorder agrees well with the previous finding, that is, the average disorder contents in eukaryotic proteins are indeed higher than those in prokaryotic proteins. The two-dimensional maps (designated as fingerprints here) based on the protein distribution densities in the LD space defined by $\ln(L)$ and ID for the representative proteomes of different organisms and protein sets were constructed, and these fingerprints show distinct patterns for different organisms and protein sets. The features and relationships among the fingerprints are analyzed and compared. To test if our classification of proteomes of different organisms and protein sets proposed here is meaningful, we generated phylogenetic trees based on the protein distribution densities in the fingerprints of proteomes of different organisms without performing any protein sequence comparison and alignments. The phylogenetic trees generated in this way were found to be meaningful, as they contain important information of evolution. Thus, the proposed approach may represent a useful and simple way for proteome classification and comparison. In present study, for each protein-encoding gene locus only the prime protein has been used, therefore, the protein densities (Figure 1 and Figure S1) could be regarded as the gene densities. Moreover, using the poplar proteome as an example, it was found that the phylogenies show little difference with or without using alternative splicing proteins (Figure S3). Discussions are made concerning the possibility for extending this approach through introduction of additional attributes.

## 2. Results

*2.1. Protein Distributions in Terms of L and ID.* Here, we discuss the proteins (811,600 entries in total) from the proteomes of different organisms and protein sets listed in Table 1, with the protein lengths varying over three degrees of magnitude from 5 (*Os06g47230* of rice) to 34,350 aa (*titin* of human). For the protein length comparison, as pointed out previously [4], the median length is a better measure than the average length to avoid biases from extremely long proteins. Table 1 lists both the median and average lengths of all the proteomes and proteins from gene sets. It should be pointed out that in the present analysis, only the primary protein at each gene locus is selected. This allows a significant simplification of proteome classification. This approximation seems to be reasonable for the main purpose of this work, as there is little difference in the results for the test cases with or without using alternative splicing proteins. Table 1 shows that the eukaryotic proteomes do not *always* have averagely longer proteins than those in the prokaryotic proteomes, as previously suggested [3, 4]. For instance, the basal flowering plant *Amborella trichopoda* has a median protein length of 218, shorter than all prokaryotes (Archaea and bacteria) surveyed here. In addition, *Giardia intestinalis* in the Eukaryota domain has an even shorter median protein length of 147. The average $L$s show the same trend as the median values (Table 1).

Nevertheless, the proteins in a eukaryotic proteome do have a significantly higher intrinsic disorder in average ($41.1 \pm 6.4\%$) compared to those in a prokaryotic proteome ($15.6 \pm 6.5\%$), consistent with previous studies [14, 28]. This trend stands for the average disorder contents of all residues from the proteomes ($47.5 \pm 6.4\%$ for eukaryotes compared to $32.9 \pm 1.4\%$ for prokaryotes). Proteomes from the archaeon *N. equitans* and bacterium *Rickettsiales* have the lowest disorder content at the protein level (7.0% for *N. equitans* and 7.7% for *Rickettsiales*) for the systems examined. As the smallest known archaeon, *N. equitans* is an obligate symbiont on the other archaeon *I. hospitalis*, which is the smallest known free-living archaeon [29]. The free-living alphaproteobacterium *Rickettsiales*, on the other hand, was suggested to be a living candidate that is close to the ancient endosymbiotic alphaproteobacteria that were merged into an archaeon and eventually transferred into the mitochondria of the first eukaryotic cell [30]. These two symbiotic or presymbiotic organisms have retained more ordered proteins compared to other free-living bacteria and Archaea surveyed here.

Consistent with previous studies [15], the proteins from the mitochondrion (88,405 proteins from 6119 species) and chloroplast (80,807 proteins form 935 species) sets have relatively low disorder contents compared to the proteins

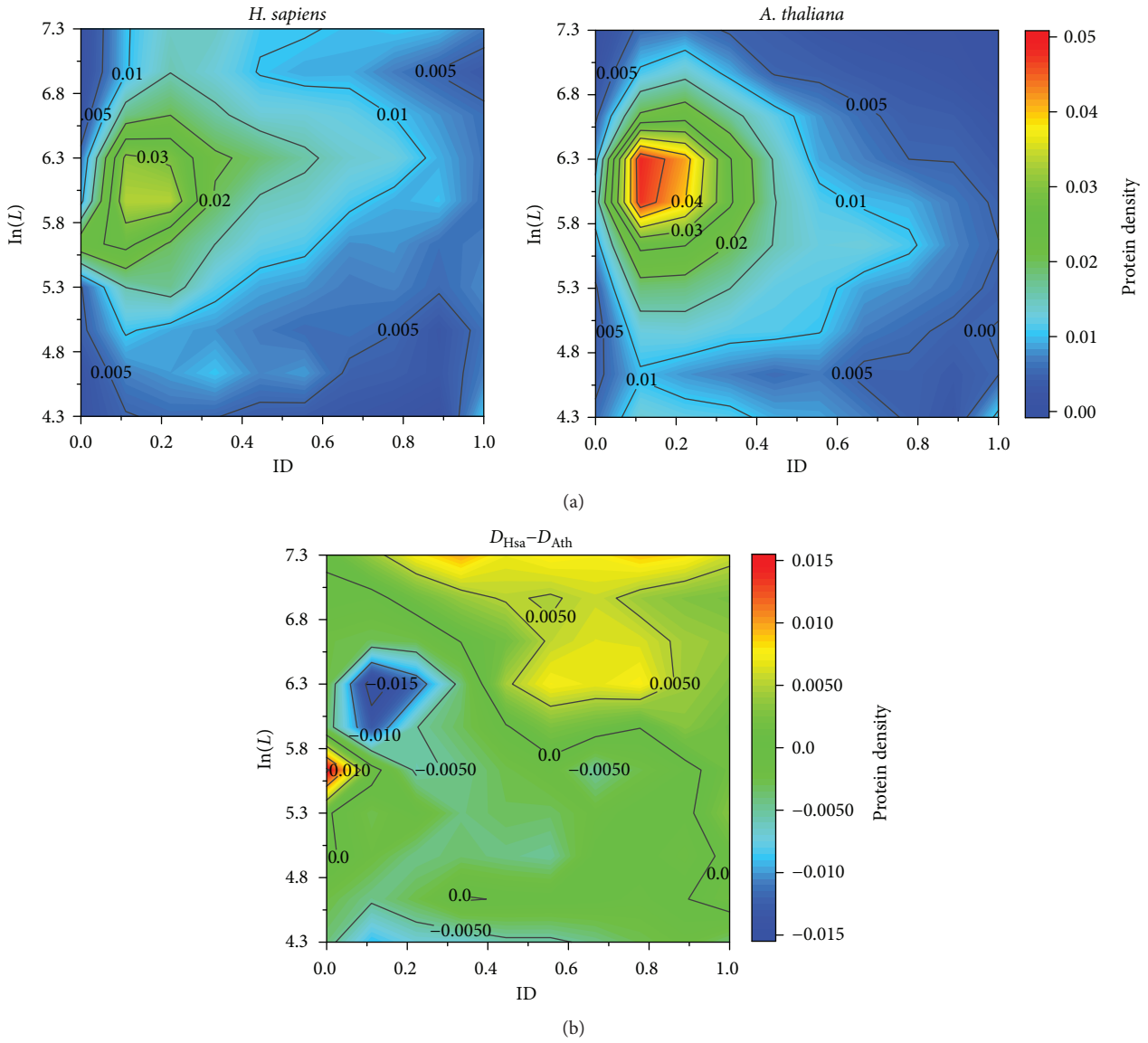



Figure 1: (a) Representative protein-density contour maps of (left) an animal (*H. sapiens*) and (right) a plant (*A. thaliana*) proteome. Short proteins ($\ln(L) < 4.3$ or $L < 74$) and long proteins ($\ln(L) > 7.3$ or $L > 1480$) are treated as $\ln(L) = 4.3$ and $\ln(L) = 7.3$, respectively, for statistics. (b) Differential protein density contour map between *H. sapiens* ($D_{Hsa}$) and *A. thaliana* ($D_{Ath}$) indicates that short disordered proteins are enriched in the plant proteome; and the animal proteome has more long disordered proteins.

encoded in nuclear genes of eukaryotic organisms, for example, the mitochondrial protein set has a considerably lower disorder content of 8.6% at protein level. The mitochondria have lost most of their ancestral genes either by transferring to the nucleus or by being discarded [31]. Here, we show that the mitochondrial proteins have relatively low disorder contents (i.e., highly ordered) at both the protein and the residue levels (Table 1). The genes retained in the mitochondrial genomes have been proposed preferentially to encode core proteins involved in electron transfers [32], and a colocalization of the redox regulation (CoRR) mechanism was proposed to explain why the mitochondrial and chloroplastic organelles retain their own genes, or proteins [33, 34]. Our analysis indicates that the chloroplast genes have their proteins with disorder contents close to the free-living prokaryotes, but higher than those from the symbiotic Archaeon *N. equitans* and alphaproteobacterium *Rickettsiales*, as well as the mitochondrial set (Table 1).

The proteomes of two giant DNA viruses (giruses), the Mimivirus and Pandoravirus, were also analyzed. The numbers of proteins encoded in these two giruses are comparable to the prokaryotic proteomes. The disorder content of the proteome of the Mimivirus is larger than that of the prokaryotes, but smaller than that of the eukaryotes surveyed here. However, the Pandoravirus has a proteome with disorder content close to that of the eukaryotes.

Finally, the viral and plasmid gene sets were analyzed. The viral gene set contains 237,463 genes collected from 4942 strains and the plasmid set contains 95,214 genes cultivated from 985 bacteria. Interestingly, the proteins from

Table 1: A summary of the proteomes and gene sets.

| Domain[a] | Species | Gene number[b] | Ave[c] | Med[c] | Max[c] | Min[c] | $ID_{pep}$%[b,d] | $ID_{res}$%[b,e] |
|---|---|---|---|---|---|---|---|---|
| Eukaryota | *H. sapiens* | 20,193 | 561.0 | 417 | 34,350 | 16 | 45.2 | 49.3 |
| | *D. melanogaster* | 13,700 | 537.2 | 396 | 22,949 | 11 | 44.3 | 49.0 |
| | *S. cerevisiae* | 5917 | 494.1 | 405 | 4910 | 16 | 38.1 | 44.6 |
| | *A. thaliana* | 27,407 | 405.2 | 348 | 5393 | 7 | 36.8 | 43.6 |
| | *P. trichocarpa* | 41,434 | 385.0 | 317 | 5410 | 29 | 35.5 | 42.6 |
| | *A. comosus* | 29,772 | 372.6 | 288 | 5407 | 31 | 39.5 | 45.4 |
| | *O. sativa* | 48,788 | 376.1 | 290 | 4957 | 5 | 38.0 | 44.5 |
| | *A. trichopoda* | 26,460 | 317.0 | 218 | 4990 | 29 | 37.5 | 43.9 |
| | *C. reinhardtii* | 17,819 | 732.9 | 498 | 23,859 | 31 | 54.8 | 61.9 |
| | *P. patens* | 32,400 | 351.9 | 250 | 5199 | 13 | 40.2 | 45.5 |
| | *G. intestinalis* | 9667 | 353.8 | 147 | 8161 | 33 | 35.1 | 41.7 |
| | *Monocercomonoides* | 16,780 | 784.6 | 393 | 14,902 | 49 | 52.7 | 60.1 |
| Archaea | *Lokiarchaeum* | 5348 | 268.4 | 224 | 3592 | 20 | 20.0 | 33.0 |
| | *I. hospitalis* | 1434 | 278.3 | 240 | 1392 | 33 | 20.4 | 34.3 |
| | *N. equitans* | 540 | 280.2 | 228 | 2197 | 45 | 7.0 | 30.6 |
| Bacteria | *E. coli* | 4140 | 316.9 | 282 | 2358 | 14 | 17.5 | 32.2 |
| | *S. elongatus* | 2612 | 305.3 | 258 | 1807 | 29 | 20.8 | 34.3 |
| | *Rickettsiales* | 1780 | 365.2 | 251 | 2243 | 31 | 7.7 | 32.8 |
| Giruses | *Mimivirus* | 979 | 356.7 | 289 | 2959 | 25 | 25.0 | 36.6 |
| | *Pandoravirus* | 2541 | 259.2 | 178 | 2321 | 26 | 36.4 | 43.5 |
| Gene sets | *Viruses* | 237,463 | 251.8 | 154 | 8573 | 9 | 28.0 | 38.8 |
| | *Plasmids* | 95,214 | 258.9 | 206 | 16,990 | 9 | 27.2 | 38.1 |
| | *Mitochondria* | 88,405 | 286.1 | 261 | 2640 | 13 | 8.6 | 20.0 |
| | *Plastids* | 80,807 | 280.0 | 156 | 5242 | 12 | 20.5 | 32.0 |
| All proteins[f] | | 811,600 | 325.7 | 225 | 34,350 | 5 | 32.2 | 39.8 |

[a]Proteomes in the three domains of life; the giant DNA viruses (giruses) and collective protein sets are listed after the cellular species; [b]Total gene numbers; [c]Protein length statistics: Ave: average; Med: median; Max: maximal; Min: minimal protein lengths; [d]Percentage of the intrinsically disordered proteins in the proteome or gene set; [e]Average intrinsic disorder contents of all residues carried by the proteome or gene set; [f]All proteins studied in the present work. The protein length statistics covers all proteins in a proteome or gene set; however, the proteins with unknown sequence(s) (*X* residues) are excluded in the intrinsic disorder calculations.

these two sets yield similar trends in both length and disorder distributions.

*2.2. Definition of the LD Space.* Consistent with a previous report [3], exponential distributions of the protein lengths ($L$) in all proteomes and protein data sets have been observed. In this analysis, all proteins of a proteome or protein set have been ranked hierarchically from the shortest to the longest, and the proteins then distribute linearly on $\ln(L)$ (the natural base was used for the logarithm function in this study). Similar linear distribution trend is observed for the percentage of residues located in the IDPR, ID (Figure 2). Therefore, a two-dimensional LD space could be defined with one phase for the content of the protein intrinsic disorder, ID, and the other phase for the logarithm of the protein length, $\ln(L)$. Figure 2 exemplifies the protein distribution in the LD space of the human proteome.

*2.3. Dependency of the Two Attributes for the LD Space.* We defined a two-dimensional LD space with the two attributes, $\ln(L)$ and ID, and these two attributes need to be independent of each other. Therefore, we calculated the correlation coefficients (CCs) between $\ln(L)$ and ID of proteins in all proteomes and protein sets (Figure 3). Pearson's and Spearman's CCs for all proteins (811,600 entries, Table 1 and S1) are −0.101 and −0.129, respectively. The overall slight negative CC (anticorrelation) indicates that there may be a trend that shorter proteins have averagely higher disorder contents than the longer proteins. However, the anticorrelational trend does not hold for all species surveyed in this study and positive CC values were found, too, such as in the animals (human and fruit fly) and green algae *C. reinhardtii* (Table S1). The variations in the correlational trends between $\ln(L)$ and ID, therefore, may have been driven by the evolutionary processes rather than a cause-and-effect relationship. As such, the validity of the protein LD space and the related architecture of protein distributions in the LD space (i.e., the "fingerprint") should be discussed in an evolutionary framework (see below).

*2.4. Architecture of Protein Distribution (Fingerprint) in the LD Space.* The most thoroughly annotated animal and plant

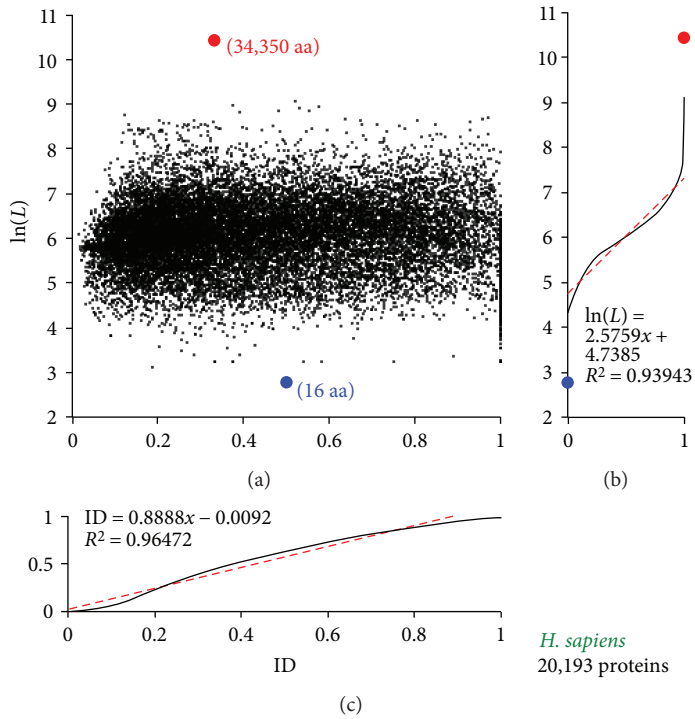



Figure 2: Protein distributions for the human (*H. sapiens*) proteome in the LD space defined by ln($L$) (the protein length in a logarithm scale) and ID (protein intrinsic disorder contents with 1.0 corresponding to proteins with 100% residues disordered and 0.0 corresponding to proteins with 0% residues disordered). The distributions in the hierarchical scale are shown in (b) and (c), respectively (see text). Linear fittings of ln($L$) and ID are shown in red dashed lines with satisfactory $R^2$ and hence support the linear participations shown in Table 2. The blue and red dots indicate the shortest (16 aa) and longest (34,350 aa) proteins, respectively.

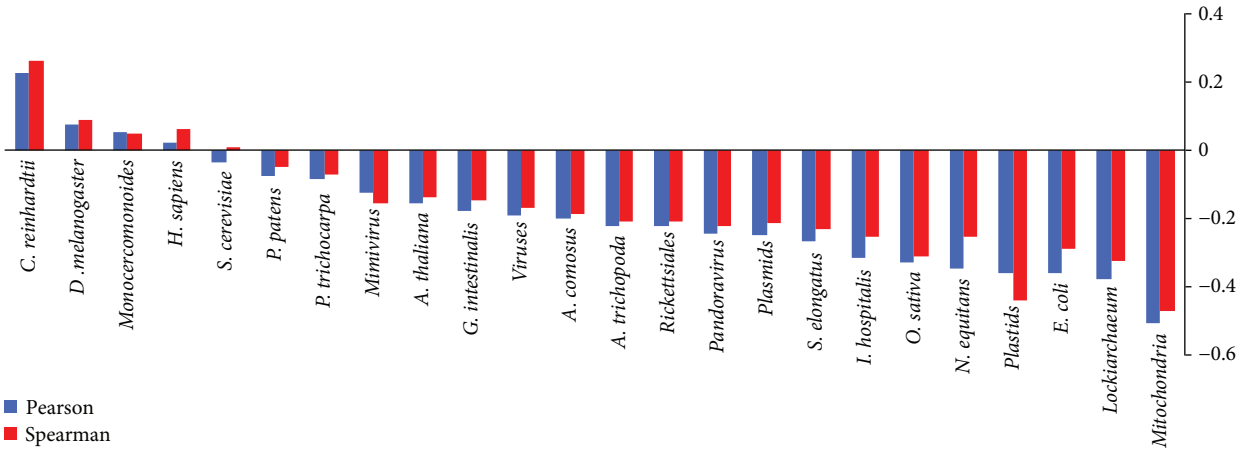



Figure 3: Pearson's (blue) and Spearman's (red) correlation coefficients (CCs) between ln($L$) and ID of the proteins in proteomes and gene sets surveyed in the present work. All species were ranked by the Pearson's CCs from the highest positive (*C. reinhardtii*) to highest negative (mitochondrial gene set).

genomes may be those of human (*H. sapiens*) and *Arabidopsis thaliana*, respectively. Using proteomes form the two representative animal and plant, the protein distributions of proteomes in the LD space were converted to the protein-density contour maps in Figure 1(a) (see Materials and Methods). As we will show below, this approach may be useful in comparative proteomes/genomics.

At a first glance, the plant proteome has more proteins of medium lengths ($\sim 5.7 < \ln(L) < 6.4$ or $\sim 300 < L < 600$) and relatively low disorder contents (ID < 0.3) whereas the animal proteome contains more long and disordered proteins (e.g., $L > 600$ and ID > 0.5). This may partly explain the slightly positive correlations between ln($L$) and ID in the animal proteomes but negative correlations in the plant proteomes. The protein distribution contour maps of other proteomes and gene sets can be found in Figure S1 in Supplementary Materials and have been trimmed in the phylogenetic tree in Figure 4 (see below).

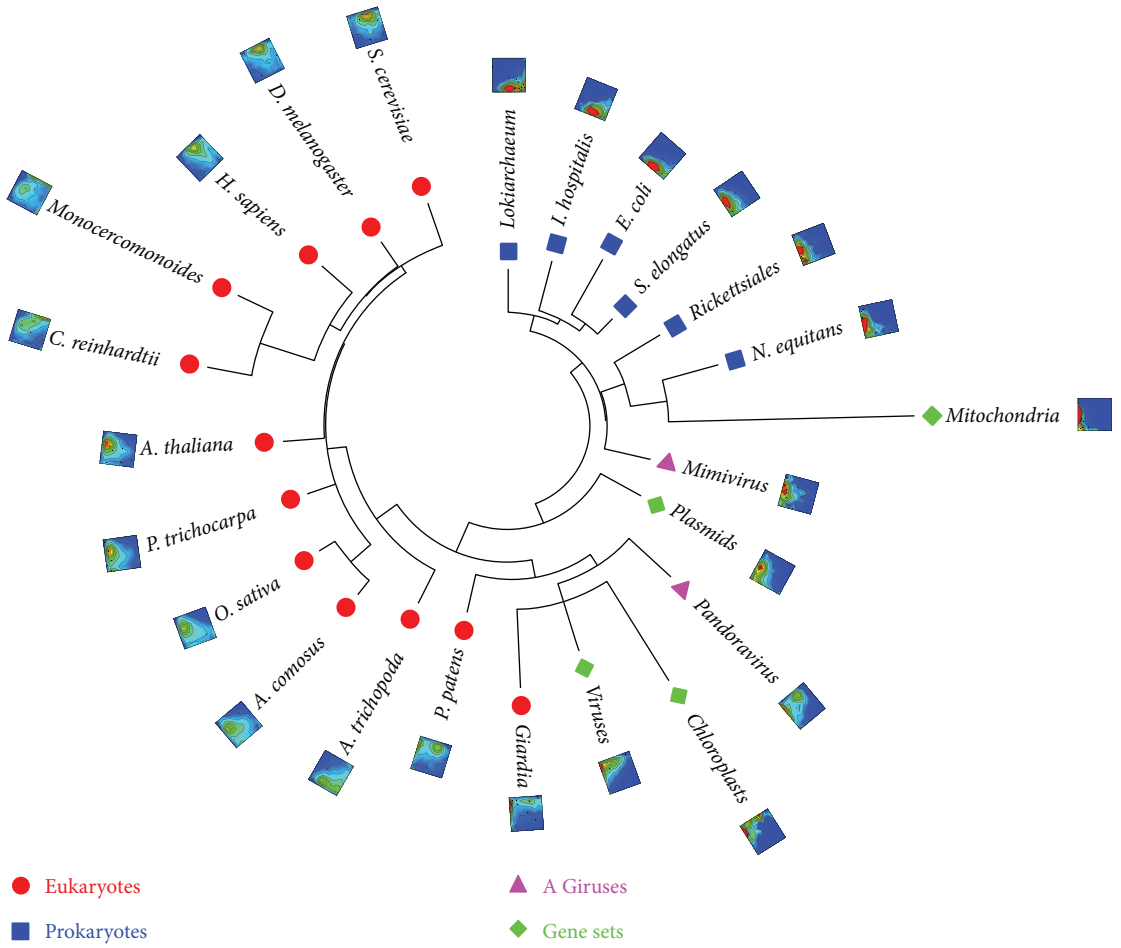



Figure 4: The phylogenetic tree reconstructed using the protein distribution densities on the LD space. The protein density distributions in the LD space for each species or gene set are also shown (Figure S1 in Supplementary Materials shows higher resolution figures). MEGA5 [69] was used to plot the tree.

It is straightforward to visualize the differences of these two proteomes using the differential contour in Figure 1(b). The *H. sapiens* proteome has 657 short proteins (i.e., $L < 100$ or $\ln(L) < 4.6$), among which 294 (1.5% of all proteins) are considered disordered (ID > 0.5); in the *A. thaliana* proteome, 888 (3.2%) out of 2292 short proteins are disordered. On the other hand, in the *H. sapiens* proteome, 1135 (5.6%) out of 2384 long proteins (i.e., $L \geq 1000$ or $\ln(L) \geq 6.9$) are disordered; whereas, in the *A. thaliana* proteome, 306 (1.1%) out of 1157 long proteins are disordered. Therefore, a significant difference between the animal (*H. sapiens*) and the plant (*A. thaliana*) could be recognized as that the former has more long disordered proteins, whereas the latter has more short disordered proteins. This difference shown in Figure 1(b) allows us to narrow down the protein/gene distributions related to the architectural differences between the two organisms.

A recent report also indicates that the overall disorder contents of the *A. thaliana* proteome are lower than those of the *H. sapiens* proteome [35], which was attributed that more IDP genes functioning in environmental adaptations may have been enriched in plants [35]. Based on our analysis and the apparent abundance of the short disordered proteins in *A. thaliana* compared to *H. sapiens* (Figure 1(b)), we focus on the 888 short (<100 aa, see above) IDP (sIDP) genes of *A. thaliana*. Among these genes, the GO annotations of 203 sIDPs could not be identified, that is, they may be considered among the "dark matter" of the *A. thaliana* proteome [36]. However, among the 685 annotated sIDPs (occupying 545 GO terms), only 20 (~0.2% of all sIDPs) with 32 GO annotations were included in the previous analysis showing "enrichment" of 74 GO annotations related to the environmental adaptations in *A. thaliana* compared to *H. sapiens* [35]. Based on our analysis, this enrichment might not be significant for the sIDPs. We suggest that it might be possible that in animals and other organism (e.g., the green algae *C. reinhardtii*), some of the sIDPs had been lost whereas long IDPs were enriched. Here, GO annotations of the plant genes were adopted from the plant comparative genomics database PLAZA 3.0 [37].

*2.5. Phylogeny Reconstructed Based on Protein Distribution Densities in the LD Space.* As the first test concerning whether our classification of proteomes and protein sets is biologically reasonable, we generated phylogenetic trees based on the protein distribution densities in the fingerprints of proteomes

Table 2: Intervals that partition the LD spaces into $M \times N$ blocks with $M = N = 10$.

| | Number | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | ln($L$) | (0,4.6) | (4.6,4.9) | (4.9,5.2) | (5.2,5.5) | (5.5,5.8) | (5.8,6.1) | (6.1,6.4) | (6.4,6.7) | (6.7,7.0) | (7.0,∞) |
| 10 × 10 | $L$ | (1100) | (101,135) | (135,182) | (182,245) | (245,331) | (331,446) | (446,602) | (602,813) | (813,1097) | (1097,∞) |
| | ID% | (0,0.1) | (0.1,0.2) | (0.2,0.3) | (0.3,0.4) | (0.4,0.5) | (0.5,0.6) | (0.6,0.7) | (0.7,0.8) | (0.8,0.9) | (0.9,1.0) |

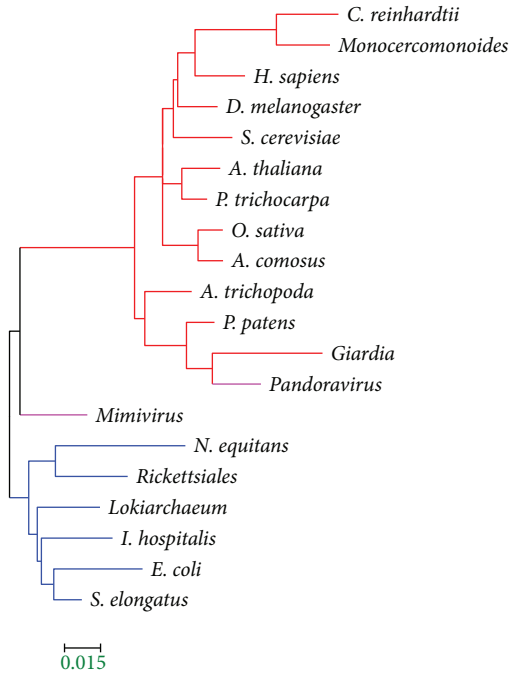



Figure 5: The phylogenetic tree reconstructed from the protein distributions in the LD space using $M = N = 10$ in (3) in Materials and Methods. Eukaryotes are in red, prokaryotes (bacteria and Archaea) in blue and giruses in pink branches. MEGA5 [69] was used to plot the trees. Because this tree is based on normalized protein densities (of 100 blocks in the $M = N = 10$ tree here), the branch length of the tree is relatively small with a scale bar of 0.015.

without performing any protein sequence comparison and alignments. Here, aiming to quantify the *architectural* differences among proteomes, the LD space was divided into $M \times N$ blocks and then, the distance between two species A and B was calculated using a Euclidian-type formula based on the protein distributions in all blocks (see (3) in Materials and Methods). In this *architectural*-distance calculation, no rigorous biological function annotations and/or genomic comparisons using BLAST or other protocols are required.

By dividing the LD space with $M = N = 10$ (Table 2), the distance matrices for all proteomes including those from giruses (Table 1) were calculated and converted to phylogenetic trees as shown in Figure 5. We also tested the $5 \times 5$ or $2 \times 2$ partitioning; the $10 \times 10$ partitioning of the LD space seems to yield relatively high accuracy (Table S2 and Figure S2 in Supplementary Materials). Nevertheless, some of the key properties are not very sensitive to the $M$ and $N$ values. Several interesting features have been found in the trees that we reconstructed: (1) the eukaryotes are clearly separated from the prokaryotes and (2) plants and animals are grouped together, even the eudicot plants (*A. thaliana* and *P. trichocarpa*) and monocot plants (*O. sativa* and *A. comosus*) are separated. The tree in Figure 5 correctly puts *A. trichopoda* before the other plant species and after *P. patents*. Interestingly, it is consistent with our understanding of the plants-fungi-animals phylogenetic relationships [38] and stays in the framework of the natural classification of three domains of life [39]. Based on the phylogenetic tree, the definition of the protein LD space might be considered meaningful to the proteomes, at least to those chosen in present work.

## 3. Discussion

To the best of our knowledge, this is the first time to classify proteomes and protein sets based on the protein distribution densities in the LD space (fingerprints), and a detailed comparison with the previous work is therefore not straightforward. Nevertheless, the survey of protein distributions in terms of each of the two attributes is consistent with the work published previously. We noticed that the eukaryotic proteomes do not always exhibit averagely longer proteins than the prokaryotic proteomes. Our observation on protein disorder agrees well with the previous finding, that is, the average disorder contents in eukaryotic proteins are indeed higher than those in prokaryotic proteins. We have also generated phylogenetic trees based on the protein distribution densities in the fingerprints of proteomes, and this allows us to make some comparisons of the results that we obtained here with the knowledge in the field and to examine the consistency and differences with earlier investigations. Such comparison may also provide certain alternative views that were generated through this unique approach.

*3.1. Giant DNA Viruses and the Tree of Life.* It has been in the debate over the years concerning if viruses should be included in the tree of life [40, 41] or if they are alive at all [42, 43]. The discovery of Mimivirus [44] that belongs to the nucleocytoplasmic large DNA viruses (NCLDV) and the following discoveries of other giant DNA viruses (giruses) [45], for example, the Pandoravirus with a genome size exceeding some of the cellular organisms [46], invoked questions on if a "fourth domain" should be added to the tree of life [46, 47] and potentially important roles that viruses played in eukaryogenesis [48]. Interestingly, we found that Mimivirus is located in between the Eukaryota and prokaryote (Archaea + Eubacteria) branches, that is, at the prokaryote-to-eukaryote transition zone. This is consistent with the original phylogenetic analysis inferred based on seven universally conserved protein sequences [44]. The Pandoravirus, on the other hand, is located within the

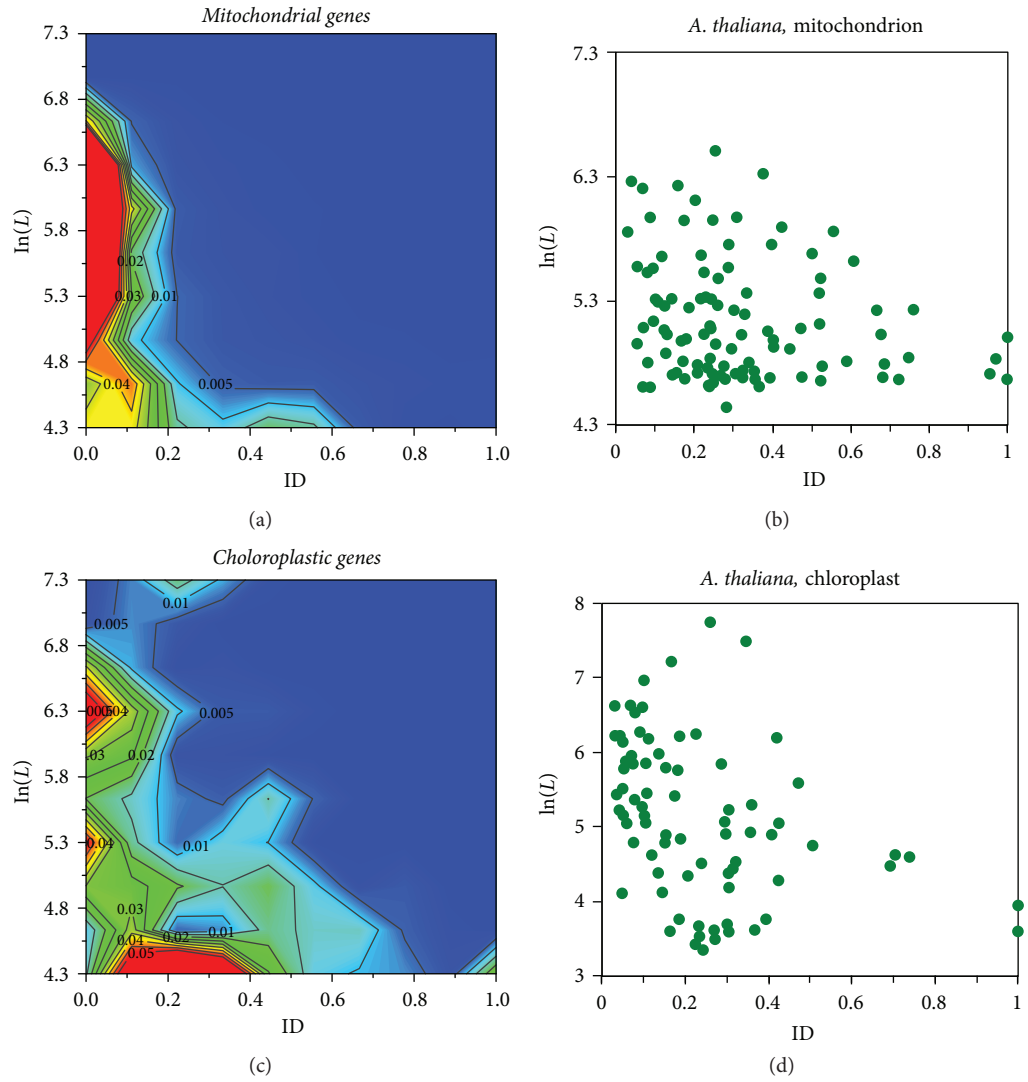



FIGURE 6: Protein distributions in the LD space for (a) the mitochondrial gene set, (b) the mitochondrial genes in *A. thaliana*, (c) the chloroplast gene set, and (d) the chloroplastic genes in *A. thaliana*.

Eukaryota branch. The vast majority (>93%) of the Pandoravirus genes exhibit no homology to anything known [46]; however, our approach puts it in the same branch of the parasite Giardia (Figure 5(c)), owing to the abundance of short proteins (both in ordered and disordered states) in these two organisms (Figure S1).

*3.2. Organelles.* The phylogenetic tree with the viral and organelle (mitochondria, chloroplasts, and plasmids) gene sets is shown in Figure 4 along with the fingerprints in the LD space. In this tree, the viral gene set is located in the same branch as the Pandoravirus. The plasmid gene set is located in between prokaryotic and eukaryotic branches, or more accurately, between Mimivirus and Pandoravirus. These results suggest the importance of horizontal gene transfers in eukaryogenesis carried by the viral and plasmid genes.

In Figure 4, the mitochondrial gene set sits in the same branch as the symbiont *N. equitans* and alphaproteobacterium *Rickettsiales*, owing to that majorities of the proteins in these proteomes and protein set are highly ordered (Table 1). The chloroplast set is located at the same branch as the viral gene set and *Giardia* (Figure 4). Using the full set of annotated mitochondrial genomes for 2015 species, a recent report [32] revealed that the proteins retained in the eukaryotic mitochondria are preferentially the structural cores in the electron transportation chains. Our survey with the mitochondrial proteins obtained from the NCBI database indicates that the mitochondrial proteins are mainly structurally ordered (Figure 6(a)), thereby possibly structurally and functionally conserved, too. However, using the model plant species *A. thaliana* as an example, the mitochondrial protein distribution in the LD space (Figure 6(b)) does not match that from the mitochondrial gene set (Figure 6(a)). This inconsistency may originate from a considerable amount of highly disordered proteins retained in the mitochondria. For instance, *A. thaliana* has 115 mitochondrial genes, 23 of which are IDPs (i.e., ID $\geq$ 0.5; here, ID refers to the ratio of residues). However, we found that 19

(out of 23) mitochondrial IDPs have unknown functions involved in unknown biological processes (Table S3 in Supplementary Materials), immediately raising a question on the validity of the results obtained from annotated mitochondrial genomes (Figure 6(a) in the present study and [32]). The protein distribution profile of *A. thaliana* chloroplast (Figure 6(d)) resembles that of the collective chloroplast gene set (Figure 6(c)). Only 6 out of 85 *A. thaliana* proteins are IDPs, all of which have been annotated as ribosomal proteins (Table S3).

## 4. Conclusion

Our two-dimensional contour maps (or proteome fingerprints) based on the protein distribution densities in the LD space show distinct patterns for different organisms and protein sets and may therefore be used for classification of proteomes and protein sets. The phylogenetic trees generated based on the protein distribution densities from the fingerprints were found to be meaningful, as they seem to contain important information of evolution. Thus, the proposed approach and its further extension may represent a useful and alternative way for proteome classification and comparison. It should be pointed out that although in the present work we used protein lengths ($L$) and protein intrinsic disorder contents (D) as the basic attributes, other attributes (not limited to those from proteins) may be introduced as well. One can imagine that one of the properties for the attributes would be that protein distributions in terms of the new attributes would be different for different proteomes (protein sets) so that the purpose of classification of proteomes (protein sets) can be achieved.

## 5. Materials and Methods

*5.1. Proteomes and Gene Set.* The plant proteomes in this study were downloaded from Phytozome, and the proteomes of bacteria, Archaea, and animals were downloaded from UniProt; the organelle protein sets were obtained from NCBI, at or before December 2016.

Here, we surveyed 12 eukaryotic proteomes from two animal species *Homo sapiens* [49, 50] and *Drosophila melanogaster* [51], two monocot plant species *Oryza sativa L. ssp. indica* [52] and *Ananas comosus* [53], two dicot plant species *Arabidopsis thaliana* [54] and *Populus trichocarpa* [55], the basal angiosperm *Amborella trichopoda* [56], the moss *Physcomitrella patens* [57], the fungus *Saccharomyces cerevisiae strain S288C* [58], the green algae *Chlamydomonas reinhardtii* [59], the metamonada Giardia (previously known as an Archezoa that lacks conventional mitochondrion) [60], and *Monocercomonoides sp. PA203* that completely lacks the mitochondrial or mitochondrial-derived genes [61]. We also analyzed three bacterial species *Escherichia coli K12 MG1655* [62], the cyanobacterium *Synechococcus elongatus PCC 7942* [63], and the alphaproteobacterium *Rickettsiales bacterium Ac37b* [64] and three Archaea species *Ignicoccus hospitalis kin4/i*, *Nanoarchaeum equitans* [29], and *Lokiarchaeum sp. GC14_75* [65]. Two giant DNA-viruses (giruses) were also analyzed, including the *Mimivirus* [44] and *Pandoravirus salinus* [46]. In addition, we downloaded several gene collections from the NCBI gene libraries containing the viral set (237,463 genes), plasmid set (95,214 genes), mitochondrial set (88,405 genes), and chloroplast set (80,807 genes). Table 1 gives a summary of the proteomes and gene sets.

The proteomes and gene sets listed above comprise 811,600 proteins, among which 2401 proteins (~0.3%) contain unknown "*X*" residues and were excluded for analysis in this work.

It should be pointed out that in the present analysis, only the primary protein at each gene locus is selected. The poplar (*P. trichocarpa*) proteome [55] was selected to test the potential influence of the versions of the proteomes and splicing alternatives. From the *P. trichocarpa* genome, there are three versions (v01, v02, and v03) of the proteomes, of which the v03 proteome has 41,434 primary proteins and 31,579 splicing alternatives (73,013 proteins in total). Using the primary proteins of all three versions and the full proteome of the v03 version as separated entries, a phylogenetic tree was constructed (Figure S3 in Supplementary Materials) and there is little difference with or without using alternative splicing proteins or by using different proteome versions.

*5.2. Intrinsic Disorder (ID) Prediction.* The PONDR-VSL2 algorithm [66] was applied to predict the ID content of all residues in a protein. This program had achieved ~81% accuracy for both short and long proteins. By default, a residue is in an ordered state if its PONDR score is less than 0.5, but in a disordered state when the PONDR score is larger than or equal to 0.5. PONDR scores of 0 and 1 corresponding to the fully ordered and fully disordered states, respectively. Here, this criterion was adopted and extended to calculate the ID content of a protein:

$$\mathrm{ID}_{\mathrm{pep}} = \frac{N_{\mathrm{D}}}{L}, \tag{1}$$

where $N_{\mathrm{D}}$ is the number of disordered residues and $L$ is the total number of residues of the protein (i.e., protein length). $\mathrm{ID}_{\mathrm{pep}}$ is also termed as the "*rough definition*" of the disorder contents in [27] and ranges from 0 to 1, with 0 and 1 corresponding to the fully ordered and fully disordered proteins, respectively.

It had been suggested that the total proteome information content (PIC) could be defined as the total number of amino acids of the primary proteins (longest isoform at each gene locus) that the proteome carries [67]. In accordance with this definition, we also calculated the average intrinsic disorder content per residue as

$$\mathrm{ID}_{\mathrm{res}} = \sum_{i=1}^{X} \frac{D_i}{X}, \tag{2}$$

where $X$ is the total number of amino acids and $D_i$ is the PONDR score of the $i$th residue of the proteome or protein set. $\mathrm{ID}_{\mathrm{res}}$ corresponds to the definition adapted in [27]. Both $\mathrm{ID}_{\mathrm{pep}}$ and $\mathrm{ID}_{\mathrm{res}}$ are listed in Table 1. Because in present work distributions of genes (or proteins) are used to discuss the evolutionary dynamics, $\mathrm{ID}_{\mathrm{pep}}$ (simplified as ID in the main

text) had been chosen to act as one of the attributes of the LD space.

*5.3. Generation of the Fingerprints and Phylogenetic Analysis.* To generate the fingerprints, the LD space of species $X$ was first divided into $M \times N$ blocks (e.g., Table 2), $M$ for $\ln(L)$ and $N$ for ID. This separation is reasonable because both $\ln(L)$ and ID exhibit linearity (Figure 2). Then, the protein density in the $ij$th block ($i$ in $\ln(L)$ and $j$ in ID%) is calculated as $X_{ij} = n_{ij}/n_{\text{tot}}$, where $n_{ij}$ is the number of proteins in the $ij$th block and $n_{\text{tot}} = \sum_{l=1}^{M}\sum_{d=1}^{N} n_{ld}$ is the total number of proteins in the proteome of species $X$. Normalization of the protein density is realized by default since $\sum X_{ij} = 1$.

Using the protein densities, the distance between two organisms A and B can be calculated using the Euclidean equation:

$$r_{AB} = \sqrt{\sum_{l=1}^{M}\sum_{d=1}^{N} (A_{ld} - B_{ld})^2}, \quad (3)$$

where $r_{AB}$ is the distance between A and B and $X_{ij}$ ($X = A$ or $B$) is the protein density in the $ij$th block. The calculated distance matrix is converted to the phylogenetic tree using the neighbor-joining method by the T-REX web server [68]. $M$ and $N$ and detailed block separations may serve as variables to fine tune the final phylogenetic tree. As a proof of concept, the reconstructed phylogenetic tree using $M = N = 10$ is shown in Figure 5.

The overall working flow of phylogenetic tree reconstruction is as follows: selection of the proteomes and protein sets → calculations and statistics of the intrinsic disorder contents (ID) and protein length of primary proteins (logarithm, $\ln(L)$) → calculations of the protein densities in all blocks (Table 2) → calculations of the Euclidian distance between each pair of proteomes or protein sets (3) → reconstruction of the phylogenetic tree based on the distance matrix.

## Disclosure

The College of Engineering & Computer Science, SimCenter, University of Tennessee Chattanooga, 701 East M. L. King Blvd., Chattanooga, TN 37403, USA, is the current address of Hao-Bo Guo. Oak Ridge National Laboratory is managed by UT-Battelle LLC for the US DOE under Contract no. DE-AC05-00OR22725. A presentation for a part of this work has been given at the Quantitative Biology 2017 Meeting in Beijing.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Acknowledgments

This work is supported by the U.S. Department of Energy (DOE), Office of Science, Genomic Science Program, under Award no. DE-SC0008834.

## Supplementary Materials

Table S1. Correlation coefficients between $\ln(L)$ and ID%. Table S2. Intervals that partition the LD spaces into $M \times N$ blocks with $M = N = 2$ and 5. Table S3. IDPs in the mitochondrion and chloroplast of *A. thaliana*. Figure S1. Protein-density contour maps (see Figure 1(a) in main text for the scale bar). Figure S2. Phylogenetic trees reconstructed from the protein distributions in the LD space using A—($M = N = 2$) and B ($M = N = 5$). Eukaryotes are in red, prokaryotes (bacteria and Archaea) in blue, and giruses in pink branches. MEGA5 (1) was used to plot the trees. Compared to that of the $M = N = 10$ tree (Figure 5), the branch length of the $M = N = 10$ tree is larger. Figure S3. Phylogenetic tree reconstructed from gene densities on the LD space. Different versions (v01–v03) of the *P. trichocarpa* proteomes have been used. By default of the present work, only proteins from primary transcripts are chosen for all proteomes. Here, for *P. trichocarpa* proteome v03, we tested both the primary transcripts (41,434 proteins) and all transcripts (73,013 proteins). We show here that progressive improvements including the splicing variants did not make significant changes in the phylogeny. *(Supplementary Materials)*

## References

[1] D. Thirumalai, E. P. O'Brien, G. Morrison, and C. Hyeon, "Theoretical perspectives on protein folding," *Annual Review of Biophysics*, vol. 39, no. 1, pp. 159–183, 2010.

[2] K. A. Dill, K. Ghosh, and J. D. Schmit, "Physical limits of cells and proteomes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 44, pp. 17876–17882, 2011.

[3] J. Z. Zhang, "Protein-length distributions for the three domains of life," *Trends in Genetics*, vol. 16, no. 3, pp. 107–109, 2000.

[4] L. Brocchieri and S. Karlin, "Protein length in eukaryotic and prokaryotic proteomes," *Nucleic Acids Research*, vol. 33, no. 10, pp. 3390–3400, 2005.

[5] V. N. Uversky, "Natively unfolded proteins: a point where biology waits for physics," *Protein Science*, vol. 11, no. 4, pp. 739–756, 2002.

[6] V. N. Uversky and A. K. Dunker, "Understanding protein non-folding," *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, vol. 1804, no. 6, pp. 1231–1264, 2010.

[7] P. Tompa, "Intrinsically disordered proteins: a 10-year recap," *Trends in Biochemical Sciences*, vol. 37, no. 12, pp. 509–516, 2012.

[8] V. N. Uversky, "A decade and a half of protein intrinsic disorder: Biology still waits for physics," *Protein Science*, vol. 22, no. 6, pp. 693–724, 2013.

[9] R. B. Berlow, H. J. Dyson, and P. E. Wright, "Functional advantages of dynamic protein disorder," *FEBS Letters*, vol. 589, no. 19, Part A, pp. 2433–2440, 2015.

[10] A. K. Dunker, S. E. Bondos, F. Huang, and C. J. Oldfield, "Intrinsically disordered proteins and multicellular organisms," *Seminars in Cell & Developmental Biology*, vol. 37, pp. 44–55, 2015.

[11] V. N. Uversky, "The multifaceted roles of intrinsic disorder in protein complexes," *FEBS Letters*, vol. 589, no. 19, Part A, pp. 2498–2506, 2015.

[12] P. Tompa, E. Schad, A. Tantos, and L. Kalmar, "Intrinsically disordered proteins: emerging interaction specialists," *Current Opinion in Structural Biology*, vol. 35, pp. 49–59, 2015.

[13] P. E. Wright and H. J. Dyson, "Intrinsically disordered proteins in cellular signalling and regulation," *Nature Reviews Molecular Cell Biology*, vol. 16, no. 1, pp. 18–29, 2015.

[14] B. Xue, A. K. Dunker, and V. N. Uversky, "Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life," *Journal of Biomolecular Structure and Dynamics*, vol. 30, no. 2, pp. 137–149, 2012.

[15] I. Yruela and B. Contreras-Moreira, "Protein disorder in plants: a view from the chloroplast," *BMC Plant Biology*, vol. 12, no. 1, p. 165, 2012.

[16] V. N. Uversky, V. Dave, L. M. Iakoucheva et al., "Pathological unfoldomics of uncontrolled chaos: intrinsically disordered proteins and human diseases," *Chemical Reviews*, vol. 114, no. 13, pp. 6844–6879, 2014.

[17] L. M. Iakoucheva, C. J. Brown, J. D. Lawson, Z. Obradović, and A. K. Dunker, "Intrinsic disorder in cell-signaling and cancer-associated proteins," *Journal of Molecular Biology*, vol. 323, no. 3, pp. 573–584, 2002.

[18] A. C. Joerger and A. R. Fersht, "The p53 pathway: origins, inactivation in cancer, and emerging therapeutic approaches," *Annual Review of Biochemistry*, vol. 85, no. 1, pp. 375–404, 2016.

[19] V. N. Uversky, I. Na, K. S. Landau, and R. O. Schenck, "Highly disordered proteins in prostate cancer," *Current Protein & Peptide Science*, vol. 18, no. 5, pp. 453–481, 2017.

[20] V. N. Uversky, "Targeting intrinsically disordered proteins in neurodegenerative and protein dysfunction diseases: another illustration of the $D^2$ concept," *Expert Review of Proteomics*, vol. 7, no. 4, pp. 543–564, 2010.

[21] S. J. Metallo, "Intrinsically disordered proteins are potential drug targets," *Current Opinion in Chemical Biology*, vol. 14, no. 4, pp. 481–488, 2010.

[22] D. Marasco and P. L. Scognamiglio, "Identification of inhibitors of biological interactions involving intrinsically disordered proteins," *International Journal of Molecular Sciences*, vol. 16, no. 4, pp. 7394–7412, 2015.

[23] J. S. Lazo and E. R. Sharlow, "Drugging undruggable molecular cancer targets," *Annual Review of Pharmacology and Toxicology*, vol. 56, no. 1, pp. 23–40, 2016.

[24] D. Kumar, N. Sharma, and R. Giri, "Therapeutic interventions of cancers using intrinsically disordered proteins as drug targets: c-Myc as model system," *Cancer Informatics*, vol. 16, 2017.

[25] S. Ambadipudi and M. Zweckstetter, "Targeting intrinsically disordered proteins in rational drug discovery," *Expert Opinion on Drug Discovery*, vol. 11, no. 1, pp. 65–77, 2016.

[26] U. Midic, C. J. Oldfield, A. K. Dunker, Z. Obradovic, and V. N. Uversky, "Protein disorder in the human diseasome: unfoldomics of human genetic diseases," *BMC Genomics*, vol. 10, Supplement 1, p. S12, 2009.

[27] M. Y. Lobanov and O. V. Galzitskaya, "How common is disorder? Occurrence of disordered residues in four domains of life," *International Journal of Molecular Sciences*, vol. 16, no. 8, pp. 19490–19507, 2015.

[28] Z. Peng, J. Yan, X. Fan et al., "Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life," *Cellular and Molecular Life Sciences*, vol. 72, no. 1, pp. 137–151, 2015.

[29] M. Podar, I. Anderson, K. S. Makarova et al., "A genomic analysis of the archaeal system Ignicoccus hospitalis-Nanoarchaeum equitans," *Genome Biology*, vol. 9, no. 11, article R158, 2008.

[30] S. G. Ball, D. Bhattacharya, and A. P. Weber, "Pathogen to powerhouse," *Science*, vol. 351, no. 6274, pp. 659-660, 2016.

[31] W. Neupert, "Mitochondrial gene expression: a playground of evolutionary tinkering," *Annual Review of Biochemistry*, vol. 85, no. 1, pp. 65–76, 2016.

[32] I. G. Johnston and B. P. Williams, "Evolutionary inference across eukaryotes identifies specific pressures favoring mitochondrial gene retention," *Cell Systems*, vol. 2, no. 2, pp. 101–111, 2016.

[33] J. F. Allen, "Why chloroplasts and mitochondria retain their own genomes and genetic systems: colocation for redox regulation of gene expression," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 112, no. 33, pp. 10231–10238, 2015.

[34] J. F. Allen, W. B. M. de Paula, S. Puthiyaveetil, and J. Nield, "A structural phylogenetic map for chloroplast photosynthesis," *Trends in Plant Science*, vol. 16, no. 12, pp. 645–655, 2011.

[35] N. Pietrosemoli, J. A. Garcia-Martin, R. Solano, and F. Pazos, "Genome-wide analysis of protein disorder in *Arabidopsis thaliana*: implications for plant environmental adaptation," *PLoS One*, vol. 8, no. 2, article e55524, 2013.

[36] N. Perdigao, J. Heinrich, C. Stolte et al., "Unexpected features of the dark proteome," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 112, no. 52, pp. 15898–15903, 2015.

[37] S. Proost, M. Van Bel, D. Vaneechoutte et al., "PLAZA 3.0: an access point for plant comparative genomics," *Nucleic Acids Research*, vol. 43, no. D1, pp. D974–D981, 2015.

[38] P. O. Wainright, G. Hinkle, M. L. Sogin, and S. K. Stickel, "Monophyletic origins of the metazoa: an evolutionary link with fungi," *Science*, vol. 260, no. 5106, pp. 340–342, 1993.

[39] C. R. Woese, O. Kandler, and M. L. Wheelis, "Towards a natural system of organisms: proposal for the domains Archaea, bacteria, and Eucarya," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 87, no. 12, pp. 4576–4579, 1990.

[40] D. Moreira and P. Lopez-Garcia, "Ten reasons to exclude viruses from the tree of life," *Nature Reviews Microbiology*, vol. 7, no. 4, pp. 306–311, 2009.

[41] J.-M. Claverie and H. Ogata, "Ten good reasons not to exclude giruses from the evolutionary picture," *Nature Reviews Microbiology*, vol. 7, no. 8, p. 615, 2009.

[42] E. V. Koonin and P. Starokadomskyy, "Are viruses alive? The replicator paradigm sheds decisive light on an old but misguided question," *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, vol. 59, pp. 125–134, 2016.

[43] P. Forterre, "To be or not to be alive: how recent discoveries challenge the traditional definitions of viruses and life," *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, vol. 59, pp. 100–108, 2016.

[44] D. Raoult, S. Audic, C. Robert et al., "The 1.2-megabase genome sequence of Mimivirus," *Science*, vol. 306, no. 5700, pp. 1344–1350, 2004.

[45] M. G. Fischer, "Giant viruses come of age," *Current Opinion in Microbiology*, vol. 31, pp. 50–57, 2016.

[46] N. Philippe, M. Legendre, G. Doutre et al., "Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes," *Science*, vol. 341, no. 6143, pp. 281–286, 2013.

[47] D. Moreira and P. Lopez-Garcia, "Evolution of viruses and cells: do we need a fourth domain of life to explain the origin of eukaryotes?," *Philosophical Transactions of the Royal Society B-Biological Sciences*, vol. 370, no. 1678, article 20140327, 2015.

[48] P. Forterre and M. Gaia, "Giant viruses and the origin of modern eukaryotes," *Current Opinion in Microbiology*, vol. 31, pp. 44–49, 2016.

[49] J. C. Venter, M. D. Adams, E. W. Myers et al., "The sequence of the human genome," *Science*, vol. 291, no. 5507, pp. 1304–1351, 2001.

[50] M. Olivier, A. Aggarwal, J. Allen et al., "A high-resolution radiation hybrid map of the human genome draft sequence," *Science*, vol. 291, no. 5507, pp. 1298–1302, 2001.

[51] M. D. Adams, S. E. Celniker, R. A. Holt et al., "The genome sequence of *Drosophila melanogaster*," *Science*, vol. 287, no. 5461, pp. 2185–2195, 2000.

[52] J. Yu, S. Hu, J. Wang et al., "A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*)," *Science*, vol. 296, no. 5565, pp. 79–92, 2002.

[53] R. Ming, R. VanBuren, C. M. Wai et al., "The pineapple genome and the evolution of CAM photosynthesis," *Nature Genetics*, vol. 47, no. 12, pp. 1435–1442, 2015.

[54] I. Arabidopsis Genome, "Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*," *Nature*, vol. 408, no. 6814, pp. 796–815, 2000.

[55] G. A. Tuskan, S. DiFazio, S. Jansson et al., "The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray)," *Science*, vol. 313, no. 5793, pp. 1596–1604, 2006.

[56] P. Amborella Genome, "The *Amborella* genome and the evolution of flowering plants," *Science*, vol. 342, no. 6165, article 1241089, 2013.

[57] S. A. Rensing, D. Lang, A. D. Zimmer et al., "The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants," *Science*, vol. 319, no. 5859, pp. 64–69, 2008.

[58] J. M. Cherry, E. L. Hong, C. Amundsen et al., "Saccharomyces genome database: the genomics resource of budding yeast," *Nucleic Acids Research*, vol. 40, no. D1, pp. D700–D705, 2012.

[59] S. S. Merchant, S. E. Prochnik, O. Vallon et al., "The *Chlamydomonas* genome reveals the evolution of key animal and plant functions," *Science*, vol. 318, no. 5848, pp. 245–250, 2007.

[60] C. Aurrecoechea, J. Brestelli, B. P. Brunk et al., "GiardiaDB and TrichDB: integrated genomic resources for the eukaryotic protist pathogens *Giardia lamblia* and *Trichomonas vaginalis*," *Nucleic Acids Research*, vol. 37, pp. D526–D530, 2009.

[61] A. Karnkowska, V. Vacek, Z. Zubacova et al., "A eukaryote without a mitochondrial organelle," *Current Biology*, vol. 26, no. 10, pp. 1274–1284, 2016.

[62] F. R. Blattner, G. Plunkett 3rd, C. A. Bloch et al., "The complete genome sequence of *Escherichia coli* K-12," *Science*, vol. 277, no. 5331, pp. 1453–1462, 1997.

[63] B. E. Rubin, K. M. Wetmore, M. N. Price et al., "The essential gene set of a photosynthetic organism," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 112, no. 48, pp. E6634–E6643, 2015.

[64] Z. Wang and M. Wu, "An integrated phylogenomic approach toward pinpointing the origin of mitochondria," *Scientific Reports*, vol. 5, no. 1, article 7949, 2015.

[65] A. Spang, J. H. Saw, S. L. Jorgensen et al., "Complex Archaea that bridge the gap between prokaryotes and eukaryotes," *Nature*, vol. 521, no. 7551, pp. 173–179, 2015.

[66] K. Peng, P. Radivojac, S. Vucetic, A. K. Dunker, and Z. Obradovic, "Length-dependent prediction of protein intrinsic disorder," *BMC Bioinformatics*, vol. 7, no. 1, p. 208, 2006.

[67] E. Schad, P. Tompa, and H. Hegyi, "The relationship between proteome size, structural disorder and organism complexity," *Genome Biology*, vol. 12, no. 12, article R120, 2011.

[68] A. Boc, A. B. Diallo, and V. Makarenkov, "T-REX: a web server for inferring, validating and visualizing phylogenetic trees and networks," *Nucleic Acids Research*, vol. 40, no. W1, pp. W573–W579, 2012.

[69] K. Tamura, D. Peterson, N. Peterson, G. Stecher, M. Nei, and S. Kumar, "MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods," *Molecular Biology and Evolution*, vol. 28, no. 10, pp. 2731–2739, 2011.

Hindawi
International Journal of Genomics
Volume 2018, Article ID 8576890, 11 pages
https://doi.org/10.1155/2018/8576890

*Review Article*

# Shiftwork-Mediated Disruptions of Circadian Rhythms and Sleep Homeostasis Cause Serious Health Problems

**Suliman Khan**[1], **Pengfei Duan**[2], **Lunguang Yao**[2] and **Hongwei Hou**[1]

[1]*The Key Laboratory of Aquatic Biodiversity and Conservation of Chinese Academy of Sciences, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan, Hubei 430072, China*
[2]*Collaborative Innovation Center of Water and Security for Water Source Region of Mid-Line of South-to-North Diversion Project, College of Agricultural Engineering, Nanyang Normal University, Nanyang, Henan, China*

Correspondence should be addressed to Lunguang Yao; lunguangyao@163.com and Hongwei Hou; houhw@ihb.ac.cn

Received 19 September 2017; Accepted 12 December 2017; Published 21 January 2018

Academic Editor: Zhining Wen

Copyright © 2018 Suliman Khan et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Shiftwork became common during the last few decades with the growing demands of human life. Despite the social inactivity and irregularity in habits, working in continuous irregular shifts causes serious health issues including sleep disorders, psychiatric disorders, cancer, and metabolic disorders. These health problems arise due to the disruption in circadian clock system, which is associated with alterations in genetic expressions. Alteration in clock controlling genes further affects genes linked with disorders including major depression disorder, bipolar disorder, phase delay and phase advance sleep syndromes, breast cancer, and colon cancer. A diverse research work is needed focusing on broad spectrum changes caused by jet lag in brain and neuronal system. This review is an attempt to motivate the researchers to conduct advanced studies in this area to identify the risk factors and mechanisms. Its goal is extended to make the shift workers aware about the risks associated with shiftwork.

## 1. Introduction

Fast growing needs demand doing work in recurring periods other than the traditional diurnal periods. The rotations in working shifts disrupt natural sleep-wake cycle and eating patterns [1], which in turn cause serious health problems by affecting mental health and work effectiveness [2]. This disruption alters circadian rhythms [3] and neuronal functions to cause neuronal disorders [2]. Shiftwork for a long period increases the risk of fatigue, aggression, sleep disorders, metabolic disorders, mental abnormalities, and death [2, 4–9]. Shiftwork directly affects alertness [10, 11], causes depression, and promotes anxiety [12].

Working repeatedly during night shifts affects hormonal system and disrupts hormonal secretions and their control factors [13–16]. This altered hormonal profile increases the risk of breast cancer, prostate cancer, gastrointestinal abnormalities, cardiovascular diseases, and reproductive aberrations [17–19]. Alterations in physiological, behavioral, and psychological mechanisms [5] further develop abnormalities associated with peptic ulcer, diabetes type II, and rheumatoid arthritis [20–22]. The measurement of shiftwork is considered complex as it requires several parameters including sleep quality, fatigue level, types of sleep problems, use of stimulants, and sleepiness [2, 23].

Shiftwork has gained huge importance which is inevitable in modern world. This situation alerts researchers to focus on shiftwork and related abnormalities. Without knowing the genetic mechanisms and identifying factors related to shiftwork-promoted health problems, development of prevention and curing strategies may not be achievable. Suitable model organism, monitoring systems, and mimicking the shiftwork in lab conditions are major challenges in studying shiftwork to investigate related disorders. These challenges make the identification genes, majorly involved in shiftwork-related abnormalities, difficult. In this review, we have focused on major health problems that are linked with shiftwork directly or indirectly.

## 2. Shiftwork Dysregulates Circadian Rhythms by Affecting Clock Genes

Circadian rhythm mainly controls the daily wake and sleep cycle and regulates physiological processes including hormone secretion, body temperature, feeding behavior, cell cycle progression, and drug, glucose, and xenobiotic metabolism. Its disruption through environmental and genetic means causes aberrations in physiological processes. Clock genes with the effects of oscillators and endocrine and neural signals regulate circadian rhythm [24] which may be disrupted apparently by shiftwork. Circadian clock, through appropriate physiological activities in relevance to time, controls the circadian rhythms [4]. Shiftwork (chronic jet lag) suppresses the expression levels of core clock genes, including Per1 and Per2 in suprachiasmatic nuclei (SCN) and MT1 melatonin and glucocorticoid receptors in the liver. It further causes delay in acrophases of circadian expression of Per1, Per2, BMAL1, and Dbp in liver. Besides clock genes, expression levels of some cell cycle-related genes including c-Myc and p53 are also altered [25]. Vasopressin (V1a and V1b) receptors (expressed in SCN neurons) promote shiftwork effect in combination with core clock genes. Individuals lacking these receptors are normally resistant to jet lag/shiftwork effects [26].

Circadian oscillator period is determined around 24 hours genetically and adjusted by synchronizers such as LD cycle. Circadian rhythms synchronized to day time work and night time sleep require phase adjustment in altered routines by central and peripheral oscillators. This phase adjustment in certain cases disrupts the normal organization and sequence of the clock. Clock genes behave differently during phase shifts [1, 27]. Shift workers undergo circadian dysrhythmia that adversely effects mental health. This further leads to circadian rhythm disorder causing aberrations in neurogenesis and spatial cognition [28].

Many aspects of circadian rhythmicity can be modulated by serotonergic agents that indicate that serotonin is involved in the regulation of circadian rhythm. Serotonin transporter (5-HTT) control serotonin reuptake depending on serotonin transporter gene (*SLC6A4*) promoter region (5-HTTLPR) [29]. Significant associations between shiftwork and S variant of the SLC6A4 promoter and 5-HT and 5-HIAA contents of platelet can help in investigating the circadian rhythm-related mechanisms imposed by shiftwork [29]. The effect of shiftwork on circadian rhythm and sleep may cause the metabolic dysregulation and depressions by affecting the genetic pathways. The affect may lead to the disruption of other functioning systems and cause relative disorders either through direct or indirect genetic interaction in continuous/discontinuous forms.

## 3. Shiftwork Disrupts Normal Sleep and Behavior

Sleep consists of two repeated cyclic patterns "nonrapid eye movement and rapid eye movement" [30]. It is controlled genetically with the influence of environmental factors [31, 32]. Adenosine, a sleep-promoting molecule [33], mediates wake-promoting effects of caffeine by acting on adenosine receptors antagonistically [34]. GABA (gamma-aminobutyric acid) promotes sleep whereas dopamine, acetylcholine, norepinephrine, and histamine promote wakefulness [35]. The cyclic guanosine monophosphate (cGMP) kinase [36, 37], regulatory subunit of Shaker [38], Sleepless (sss) gene [39], and CLK and CYC proteins [40] are key players in sleep regulation. All these genes and regulators may get disrupted in shift workers which will lead to abnormal sleep and other serious health conditions.

Shiftwork due to the rotation of working schedules and light/dark disturbance directly affects sleep (as shown in Figure 1) and may cause health abnormalities including insomnia, sleep apnea, periodic leg movements, restless leg syndrome (RLS), and sleep-wake state dissociation disorders such as rapid eye movement (REM) and narcolepsy, sleep behavior disorder, and sleep walking [34]. Furthermore, the shiftwork-induced deficits in sleep homeostasis and circadian rhythms may lead to different psychiatric disorders and affect SUR2 gene, which was found involved in energy metabolism and aetiology of cardiomyopathies [41]. Shiftwork may disrupt daily patterns of human physiology controlled by circadian rhythms and sleep, including regulation of energy patterns expenditure [42, 43] and glucose metabolism [42].

Shiftwork may negatively affect genes and factors involved in sleep disorders such that a SNP marker [44] and chemokine (C-C motif) receptor 3 (*CCR3*) susceptibility gene [45] associated with narcolepsy, MEIS1 locus, and neuronal nitric oxide synthase (NOS1) and BTBD9 associated with RLS [46–48]. Serine to glycine mutation in *PER2* and mutation in casein kinase I$\delta$ gene (*CSNK1D)* develop familial advanced sleep phase syndrome (FASPS) [34, 35] whereas point mutation in *PRNP* causes *fatal familial insomnia (FFI)* [35]. The disrupted functions of the above genes through shiftwork may lead to related disorders. Shiftwork can alter the functionality of sleep-promoting immune genes [49], glucocorticoids and NSAIDs [50], protein NF-kB (upregulated during sleep deprivation) [34], TAK1 (TGF-b-activated kinase) and *Sik3* (control sleep behaviors), and *Nalcn* gene (involved in REMS) [51]. Widening its effectiveness, shiftwork promotes acute coronary syndrome [52], menstrual disturbances, and abnormal insulin levels [53] and increases sleepiness and alcohol consumption [54]. It also lifts sleep deprivation that ultimately leads to loss in cognitive, physical, and metabolic consequences. These alterations can possibly end up with developing cardiovascular morbidity and mortality [29].

## 4. Shiftwork Dysregulates Metabolic Process and Develops Related Disorders

Social jet lag is a key player in developing metabolic syndrome by inducing changes in cholesterol levels and disrupting normal food processes. It was observed in an experiment that social jet lag potentiated body weight gain by increasing overconsumption of cafeteria food. As a result, it promoted high insulin and dyslipidemia indicating the risk of metabolic syndrome [55]. A chronic shift in light/dark (LD) cycles induces obesity, increases body weight and

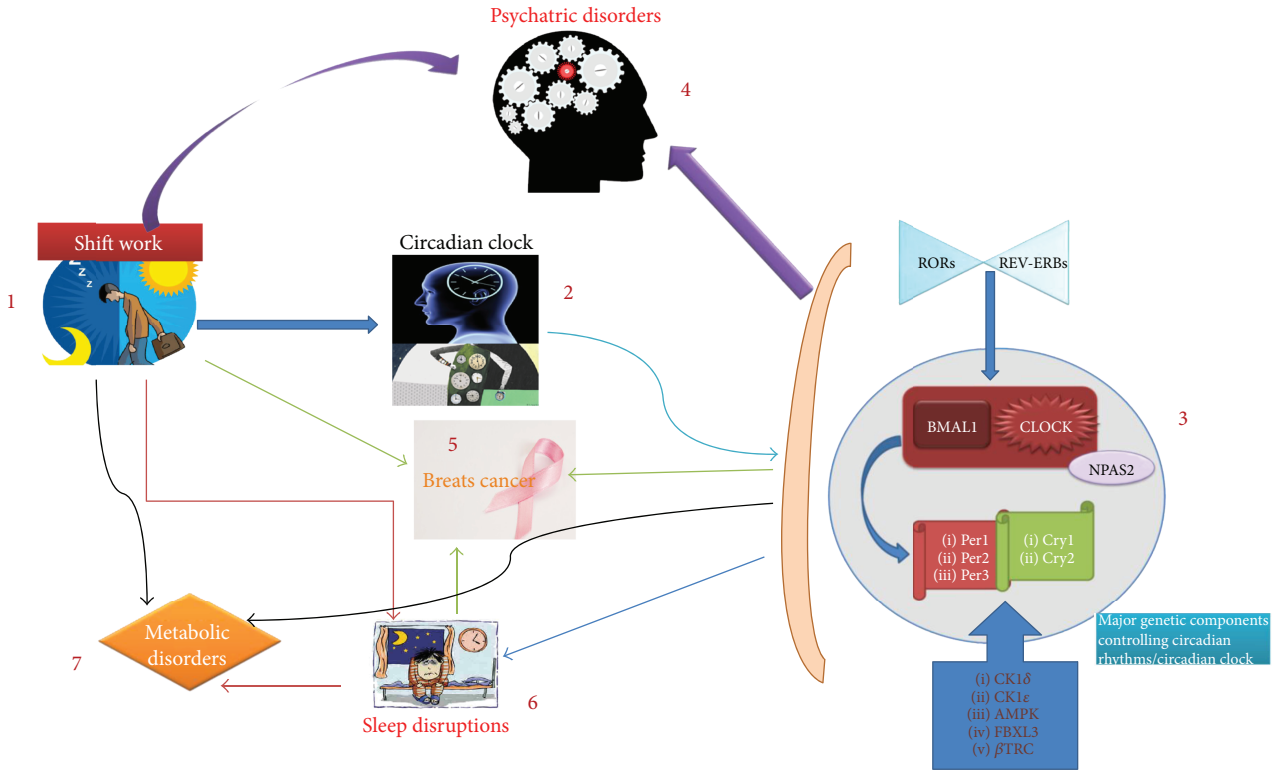



Figure 1: The abnormalities associated with shiftwork. (1) Shiftwork primarily affects the circadian clock and leads to several disruptions and disorders. (2) Disruption in circadian clock further affects the circadian system and alters important gene expressions which play an important role in maintaining normal body functions. (3) BMAL1 and CLOCK genes are the key factors that control circadian rhythms. Any alteration in these groups of genes further alters the genetic expression of genes (Per 1–3 and Cry 1/2) involved in clock maintenance. Other factors and proteins which play a major role in circadian rhythm are RORs, REV-ERBs, CK1$\delta$, CK1$\varepsilon$, AMPK, FBXL3 and $\beta$TRC. (4) Shiftwork either directly or through circadian system alterations may cause severe psychiatric disorders including major depression, anxiety, and mood disorders. (5) Due to exposure to light irregularly, shiftwork enhances the chances of breast cancer. Breast cancer development is promoted by sleep disruptions, circadian system imbalances, and dietary conditions. (6) Sleep disorders and disrupted sleep is another condition developed by shiftwork. (7) This disruption in sleep affects the metabolism and causes metabolic disorders including obesity.

glucose intolerance, and accumulates more fat in white adipose tissues. It changes the expression profiles of metabolic genes in liver [56].

According to timed sleep restriction (TSR) study, sleep timing directly affects clock-controlled genes including ClockD19, BMAL1, Per1, Rev-erb$\alpha$ and Dbp, and circadian machinery. This alteration affects metabolic processes including carbohydrate and lipid utilization in liver. BMAL1, Per1, and Dbp were upregulated during early light phase and Rev-erb$\alpha$ during mid-light phase. Carbohydrate regulators affected were insulin receptor substrate 2 (Irs2), glucose-responsive fork head box O1 (Foxo1), and glucose transporter 2 (Slc2a2). These regulators control the functions of pyruvate carboxylase, the pyruvate transporter Slc16a7, pyruvate dehydrogenase kinase 4, the glycerol transporter aquaporin, fructose-2,6-biphosphatase 1, and liver pyruvate kinase. These alterations in carbohydrate regulators further affect glycerol kinase and glycerol phosphate dehydrogenase 2, glutamic-pyruvate transaminase, and regulators of glycerol biosynthesis [4]. This dysregulation may induce critical alterations in genetic system to develop metabolic disorders.

The presence of metabolic process controlling genes in rhythmic transcriptome indicates that alteration in circadian rhythm causes disruption in metabolic process [4, 57–59]. CLOCK, BMAL, and PER2 were found associated with obesity by affecting the metabolic process [60, 61]. Shiftwork brings about changes in appetitive hormones through circadian misalignment which causes reduction in leptin level that ultimately leads to weight gain. Shift workers are considered at higher risk of type II diabetes [42, 62].

## 5. Shiftwork-Related Health Risks

*5.1. Shiftwork Is Associated with Mortality.* Quick return occurs due to rotations in between shifts, causing major accidents [22] that normally lead to death. Shiftwork increases the risk of several disorders including mental disorders and sleep-related disorders, which may lead to death. In an experiment related to jet lag effects on transgenic aged rats, it was found that mortality light scheduled rotations at 6 hours phase advance, each week. The survival rate of rats exposed to shifted light schedule was found lesser (47%) than the survival rate (83%) of rats exposed to normal

light (12L/12D) scheduled rotations. Although jet lag has no mortal effect on younger mice, it alters the behavior and circadian rhythms which further causes serious health issues by affecting brain and liver [7].

*5.2. Shiftwork Induces Mental and Related Disorders.* The role of SLC6A4 (serotonin transporter gene) in shift workers was confirmed to be related with time period. The proportion of short allele becomes higher than large allele if the period is more than 5 years. The effectiveness of shiftwork becomes higher in case of internal desynchronization of circadian rhythm [2, 63].

Endocrine imbalance in depression and psychosis is the hyperactivity of the HPA axis [64]. Glucocorticoid functions as an immunosuppressant and interacts with melatonin in a form of suppressive effect on the production of ACTH-induced glucocorticoid [65]. Its resistance results in hypercortisolemia in psychiatric patients and increased pituitary volume in depressed patients [66, 67]. Shiftwork causes desynchronization in circadian rhythm which in turn leads to reduced NK activity and weakening cellular immunity [68], disruptions in norepinephrine, melatonin, and serotonin production [69, 70] which may induce anxiety, depression, hypercortisolemia, and psychosis.

*5.3. Shiftwork Disorder and Associated Factors.* Shiftwork disorder is characterized by insomnia and excessive sleepiness which develops due to working schedules overlapping sleeping time [71]. According to a study, shiftwork disorder was found to occur frequently in males as compared to females, working during the night. Approximately, 9% of the night shift workers reported severe shiftwork disorder while one-third of the total had mild symptoms [71]. Shiftwork sleep disorder changes the behavior and sleeping periods permanently which could possibly be recovered through advanced therapies only. Treatment with modafinil, an effective compound against narcolepsy and obstructive sleep apnea, was found ineffective [6], suggesting that the mechanisms involved in shiftwork sleep disorder are not related to sleep apnea and narcolepsy. Molecular level studies with a broad range of patients, to investigate mechanisms, are needed to find out the main alterations in the system to prevent and cure the conditions properly.

The high number of nights and age were found associated with shiftwork disorder [22], whereas genotypes were found to be associated with morningness and eveningness. Allele 3111C is associated with extreme eveningness and shiftwork with semidominant influence on sleep phases without having an obvious influence on morning/evening preferences. 3111C/C homozygotes are associated with delayed shift of sleep [72] where melatonin can phase-shift the circadian clock being chronobiotic and a sleep promoting agent [73].

*5.4. Shiftwork Sleep Disorder.* Shiftwork sleep disorder is a condition defined by excessive sleepiness or insomnia accompanied by total sleep time reduction. It is 10–38% prevalent in shift workers [5, 43]. The disturbance in general is the sleep-wake cycle distortion of extrinsic origin. This disorder is related to the night and early morning timings. Reduction in alertness and performance along with the linkage to higher rates of comorbidity with GI disorders [67] make SWSD a severe and attention-requiring condition. Sleepiness during the night and insomnia during the day become more severe with continuous shiftwork for longer periods. Depression, ulcers, and sleepiness-related accidents are the ultimate risks associated with shiftwork sleep disorder. This sleepiness behavior during night is similar to the day time sleepiness in people with narcolepsy [6].

*5.5. Advanced and Delayed Sleep Phase Syndrome.* Advanced sleep phase syndrome (ASPS) is characterized by 3-4 hours advanced awakening times and sleep onsets relative to normal times. It is inherited in autosomal dominant mode, caused by circadian cycle irregularity as circadian clock genes are key players in its development. Missense mutations S662G (occurs in phosphorylation site within CSNK1-binding domain of PER2) and CSNK1D are involved in ASPS [58].

Delayed sleep phase syndrome (DSPS) is characterized by chronic inability to fall asleep and awaken at normal timings [58]. Circadian system genes are majorly involved in this dysregulated sleep behavior. Significant associations with T3111C polymorphism in the 3′UTR of CLOCK and the association of a SNP in the 5′UTR of PER2 with morning preference have been reported in DSPS [72, 74]. According to a study, amino-acid substitution S408N in the CSKN1E gene might protect the body form DSPS and non-24-hour sleep-wake syndrome development [58]. Both the shorter and longer VNTR (variable number tandem repeat) alleles were found (PER3-4/4) associated with DSPS [75]. Per3 gene is involved in delayed sleep phase syndrome and extreme diurnal preference.

Mutation in hPER2 has a notable association with advanced sleep phase syndrome, whereas some haplotypes of hPER3 have shown an association with delayed sleep phase syndrome. These conditions arise because of the alterations into casein kinase Iε (CKIε) phosphorylation of the target clock proteins, affecting morningness and eveningness. Polymorphisms in 3′ flanking region of the human clock homolog (3111T/C, hClock) is associated with eveningness and morningness such that 3111T allele has lower evening preference than 3111C allele [72]. As shiftwork has direct impacts on sleep, it may affect the normal functioning of genes to cause one of the two sleep abnormalities ASPS and DSPS.

*5.6. Circadian Rhythm Sleep Disorder.* Circadian rhythm sleep disorder refers to the abnormalities brought about by circadian misalignment due to variations in sleep-wake pattern. The shift in sleep-wake time is commonly caused by shiftwork, jet lag, light exposure, and insufficient sleep period. Circadian misalignment alters neuroendocrine physiology, impairs glucose tolerance, and reduces insulin sensitivity [42, 62, 76, 77]. Sleep deprivation problem rises with the incompatibility between circadian rhythms and working periods [78]. Both circadian alterations and sleep deprivations lead to fatigue, impairments in vigilance and

attention, sleepiness [79], sleep deficiency, impaired physiological function, and aberrant behavior [76, 80–82]. A subset of insomnias including non-24-hour sleep-wake syndrome is also linked with circadian rhythm sleep disorder [81]. Exposure to bright light in shiftwork and working in consecutive shifts if maintained for longer time may change the system at the genetic level. This alteration will further disrupt the normal functions of gene-related behavior, sleep, and circadian system. Such modifications will ultimately lead to health-harming conditions.

## 6. Shiftwork-Related Alterations in Nervous System to Develop Psychiatric Disorders

Adversely affected nervous system by working in continuously rotating shifts may accelerate the rate of psychiatric disorder occurrence (Figure 1). Being major causes of disabilities [83], psychiatric disorders including bipolar disorder (BD), schizophrenia (SZ), and major depression disorder (MDD) impose enormous medical burdens [84–86]. The genetic alteration and uncontrolled expression of genes primarily causing the aforementioned disorders [83, 87–89] are associated with irregular continuous changing of shifts during work. Night shift work contributed toward several psychiatric disorders through circadian misalignment, sleep deprivation, and light-induced melatonin suppression [73]. Disrupted-in-schizophrenia-1 (DISC1) is an important genetic factor in serious mental disorders including SZ, BD, and MDD [90]. We will further discuss the previously mentioned psychiatric disorders one by one to provide detailed information.

*6.1. Bipolar Disorder.* BD is considered one of the severely disabling disorders. Distinctive distortions of emotion regulation make BD a severe psychiatric condition. It mainly affects emotional and social behavior with light effect on perception and thought. Being a multifactorial disorder, its risk is influenced by genetic and environmental factors [91]. Genetic and pathophysiological factors involved in the development of BD are largely unknown [92]. BD increases the risk of schizophrenia and major depression disorder which is an indication for shared genetic basis between these disorders. Although heritability has been proven, more studies are needed to investigate the genetic mechanisms [93, 94]. In BD, important genetic variants that could be affected by shiftwork either directly or indirectly are Del, ANK3 rs1938526, COMT Val158Met, DAOA, BRD1/ZBED4, BDNF Val66Met, BRD1, ASMT, CAMTA1, CCDC132, CHES1, DGKH, DRD4, HTR1A C1019G, SLC6A4, 5-HTTLPR PARD3B, PDLIM5, STin2 VNTR, KLHL3, LYPD5, MAOA T941G, MTHFR A1298C and C677T, TPH1 STAB1, HTR3B, and WNK2 [95]. They are expressed in brain and associated with CREB (cyclic adenosine monophosphate response element-binding protein). *KCNH7* R394H (rs78247304) mutation is linked to BD. *ANK3*, *CACNA1C*, an intron variant of *CACNA1C* (rs79398153), and a missense mutation of *ANK3* (N2643S) were confirmed being involved in BD [95, 96]. All these mutations are possibly enhanced in people working in several shifts or exposing to irregular light (jet lag).

*6.2. Major Depression Disorder.* MDD, a leading cause of loss in work productivity, is considered a fatal disorder. It is considered one of the most prevalent disorders that affect females more than males [97, 98]. Genetic and environmental risk factors mainly contribute to cause MDD. It is a neuroprogressive disorder in which each persisting episode increases impairment in function and sensitivity for upcoming episodes. A decrease in the GR messenger ribonucleic acid levels in hippocampus, frontal cortex, and amygdala has been observed in the patients suffering from MDD. Susceptibility to further episodes is increased by repeated illness which causes the permanent alteration in the normal functions of neurons [99] and genes [100] including MTHFR C677T and 5-HTTLPR [101]. Some of the important genetic variants associated with MDD that could be affected by shiftwork are APOE, SLC6A4, ACE, GNB3, HS6ST3, HTR1A, LHFPL2, PDE11A, DISC1, MAOA, SLC6A3 (DAT1), SLC25A21, VGLL4 BDNF, P2RX7, TPH2, PDE9A, and GRIK3 [95, 100]. Neutral amino acid transporter (SLC6A15) is a susceptibility gene for MDD [102].

## 7. Shiftwork Affects Expressions of Oncogenes to Develop Cancer

Malignancies are majorly developed by mitigation in pineal hormone melatonin by bright light at night [1]. Reduced production of melatonin, phase shift, and sleep disruption, caused by exposure to light at night, might be the possible mechanisms that cause cancer and related disorders [103]. Shiftwork causes cancer (Figure 1) via altering the clock-controlled gene expression that regulate tumor suppression [1, 104]. Per1 negatively affects the growth of tumor cell and per2 functions as tumor repressor [105]. Long-term shiftwork negatively affects IFN$\gamma$ (interferon gamma) [106] The association of shiftwork-affected immune system with sleep deprivation increases inflammatory markers thereby causing malignancies and metabolic and cardiovascular disorders [1]. Circadian rhythm disruption by shiftwork or bright light exposure at night increases the rate of cancer and decreases the nocturnal rise in melatonin [19].

*7.1. Breast Cancer.* The disrupted circadian rhythm or circadian clock through shiftwork effects the development of breast cancer (Figure 2) [107]. Cell proliferation and apoptosis is regulated by approximately 7% of clock-controlled genes, including myelocytomatosis viral oncogene human recombinant (C-Myc), murine double minute oncogene (Mdm-2), and growth arrest and DNA damage-inducible alpha protein (Gadd 45a) and those encoding p53, caspases, and cyclins. Per1 is reduced in cancer tissue and its inhibition blunts apoptosis, whereas Per2 represses tumor in breast cancer and induces estradiol (E2) in mammary cells. Coexpression of Per2 with cytochrome inhibits growth of breast cancer cells. Deficiency of Per2 causes deregulation of Myc, cyclin A, cyclin D1, Mdm-2, and Gadd45a, while its dysfunctionality impairs apoptosis mediated by p53 by activating c-Myc signaling pathway. Overexpressing and downregulating Per1 or Per2 inhibits and accelerates growth

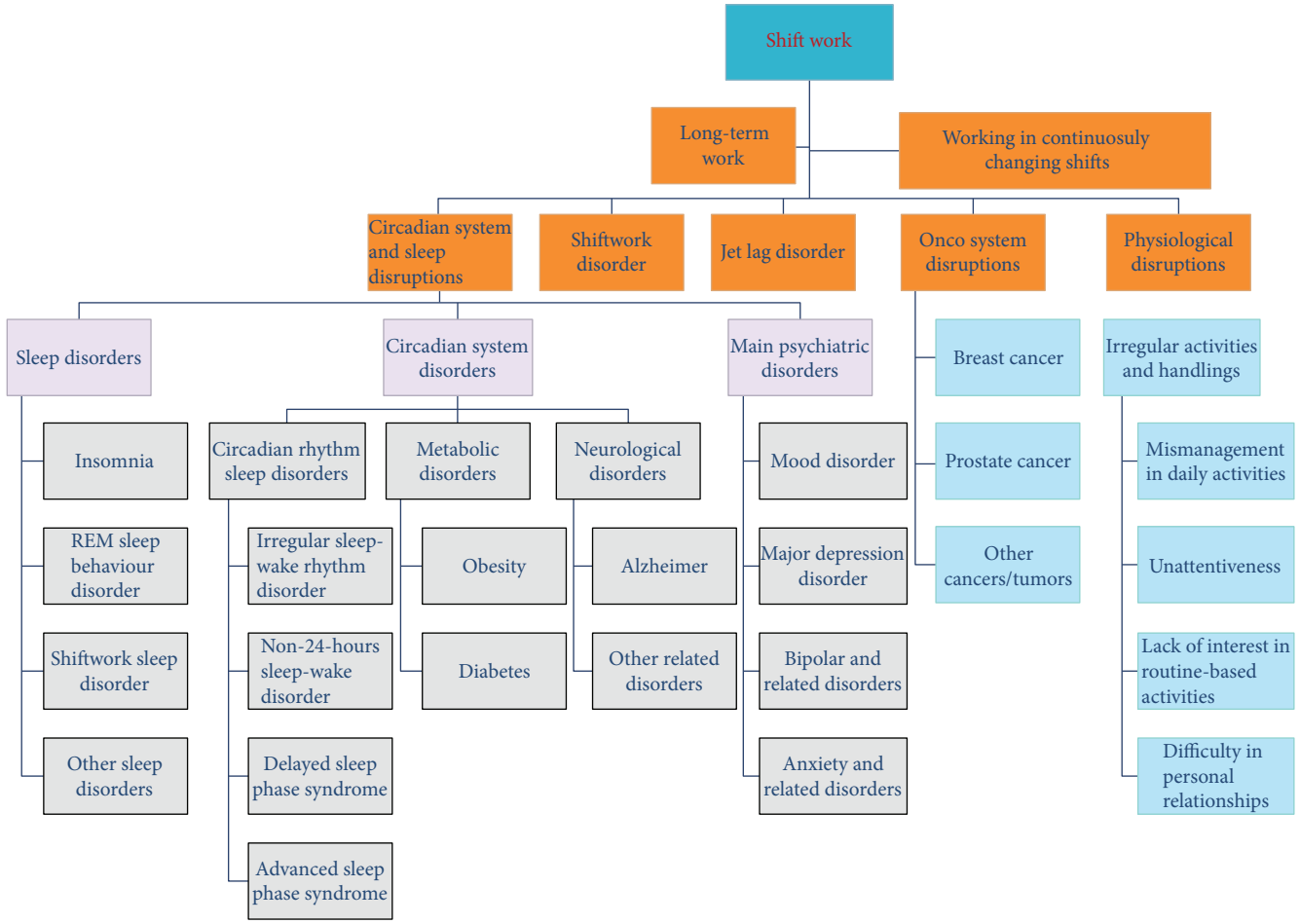



Figure 2: All disorders that could possibly be caused by shiftwork.

of cancer cells, respectively [1]. Telomere shortening is correlated with consecutive night shift work for a long time where it increases breast cancer risk [108].

*7.2. Colonic Cancer.* Per1- and Per2-mediated tumor-repressing effect is more specific to early light and early dark phase. Per2 mutations and deregulation favor development and increases cell proliferation, respectively [1]. Altered light schedules develop mutations in PER1, PER2, and PER3 which in turn promote colonic adenoma and colonic cancer [19, 109].

*7.3. Prostate Cancer.* Shiftwork-mediated sleep disruption is associated with elevated prostate-specific antigen (PSA), indicating an increased risk of developing prostate cancer [110]. Disrupted circadian rhythms through jet lag inhibits p53, enhances Myc expression, and induces tumorigenesis in prostate tissues targeted by endocrine [1, 111, 112].

*7.4. Ovarian Cancer.* Functional analysis suggested that variation, in circadian genes including *BMAL1*, *CRY2*, *CSNK1E*, *NPAS2*, *PER3*, *REV1*, and downstream transcription factors *KLF10* and *SENP3* through disruption of hormonal pathways or changes in light/dark schedules, is associated with ovarian cancer. Silencing the expression of *BMAL1* activates p53 to prevent cell cycle arrest which indicates that *BMAL1* gene may regulate the p53 tumor suppressor pathway. Per2 inhibition, reduces estrogen receptor $\alpha$ (ER$\alpha$) response to E2 by overexpression or enhances E2 activation [112, 113]. Its lower expression along with lower expression of *BMAL1* and *CRY1* promotes lower survival of cells in ovarian cancer [114].

*7.5. Lung Cancer.* Not only systemic but also somatic disruption of circadian rhythms mediated by jet lag alone affects c-Myc and enhances cell proliferation. Per2 and BMAL1 lose the ability of inhibiting tumor progression and hence promotes lung cancer [115].

*7.6. Shiftwork Tolerance.* The effects of shiftwork regarding tolerance and responsiveness are concerned with certain factors. The youngest and oldest individuals are affected more than individuals with middle age [116]. Females due to low tolerance develop more problems, like sleep disruptions, higher risk of mortality, disability, fatigue, and obesity, while males show more cognitive disturbances [117, 118]. In case of circadian rhythms, a low score on morningness [119] and languidness and a high score on flexibility [120] are associated positively with shiftwork tolerance.

## 8. Conclusion and Prospects

Shiftwork has gained central importance due to its detrimental effects on health. A large population of the world, including regular travelers, night shift workers, continuously faces jet lag conditions. Working in frequently rotating shifts causes several medical issues to arise, including cancer, psychiatric disorders, quick return accidents, and metabolic disorders. These conditions lasting longer may bring irreversible changes in the body, leaving no choice for affective recovery. It is impossible to avoid working in rotating shifts or prevent oneself from light exposure.

The health problems related to shiftwork are developed by disruptions in genetic expressions. To prevent or mitigate the adverse effects of shiftwork-related disorders, unveiling of genetic mechanisms and determination of related pathways are needed.

These relations and interactions could be studied in a better way through jet lag, circadian rhythms, and sleep behaviors.

The effects of shiftwork may involve a series of genes and factors majorly involved in circadian rhythm, sleep homeostasis, and the specific disorder prevention. It is considered that circadian genes have noticeable impacts on cancer controlling genes, psychiatric disorder causing genes, and metabolic disorder-related genes. But it is still needed to be investigated whether shiftwork directly affects the genes related to certain disorders or it elongates its impact via targeting other genes such as clock genes. The proper targeted medications are possible to be developed only if the mechanisms and factors causing and promoting the shiftwork effects to cause disorders have been identified.

## Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant no. 31372381) and the Henan Scientific and Technological Innovation Team Fund (17454). The authors are thankful to the Chinese Academy of Science and The World Academy of Science (CAS-TWAS) scholarship program.

## References

[1] E. L. Haus and M. H. Smolensky, "Shift work and cancer risk: potential mechanistic roles of circadian disruption, light at night, and sleep deprivation," *Sleep Medicine Reviews*, vol. 17, no. 4, pp. 273–284, 2013.

[2] I. B. Saksvik, B. Bjorvatn, H. Hetland, G. M. Sandal, and S. Pallesen, "Individual differences in tolerance to shift work – a systematic review," *Sleep Medicine Reviews*, vol. 15, no. 4, pp. 221–235, 2011.

[3] R. Salgado-Delgado, M. Angeles-Castellanos, N. Saderi, R. M. Buijs, and C. Escobar, "Food intake during the normal activity phase prevents obesity and circadian desynchrony in a rat model of night work," *Endocrinology*, vol. 151, no. 3, pp. 1019–1029, 2010.

[4] J. L. Barclay, J. Husse, B. Bode et al., "Circadian desynchrony promotes metabolic disruption in a mouse model of shiftwork," *PLoS One*, vol. 7, no. 5, article e37150, 2012.

[5] E. Flo, S. Pallesen, N. Magerøy et al., "Shift work disorder in nurses – assessment, prevalence and related health problems," *PLoS One*, vol. 7, no. 4, article e33981, 2012.

[6] C. A. Czeisler, J. K. Walsh, T. Roth et al., "Modafinil for excessive sleepiness associated with shift-work sleep disorder," *The New England Journal of Medicine*, vol. 353, no. 5, pp. 476–486, 2005.

[7] A. J. Davidson, M. T. Sellix, J. Daniel, S. Yamazaki, M. Menaker, and G. D. Block, "Chronic jet-lag increases mortality in aged mice," *Current Biology*, vol. 16, no. 21, pp. R914–R916, 2006.

[8] T. Kubo, K. Ozasa, K. Mikami et al., "Prospective cohort study of the risk of prostate cancer among rotating-shift workers: findings from the Japan Collaborative Cohort Study," *American Journal of Epidemiology*, vol. 164, no. 6, pp. 549–555, 2006.

[9] R. G. Stevens, J. Hansen, G. Costa et al., "Considerations of circadian impact for defining 'shift work' in cancer studies: IARC Working Group Report," *Occupational & Environmental Medicine*, vol. 68, no. 2, pp. 154–162, 2010.

[10] M. Sallinen and G. Kecklund, "Shift work, sleep, and sleepiness - differences between shift schedules and systems," *Scandinavian Journal of Work, Environment & Health*, vol. 36, no. 2, pp. 121–133, 2010.

[11] A.-C. Bara and S. Arber, "Working shifts and mental health – findings from the British household panel survey (1995–2005)," *Scandinavian Journal of Work, Environment & Health*, vol. 35, no. 5, pp. 361–367, 2009.

[12] M. Härmä and G. Kecklund, "Shift work and health – how to proceed?," *Scandinavian Journal of Work, Environment & Health*, vol. 36, no. 2, pp. 81–84, 2010.

[13] J.-U. Rehman, K. Brismar, U. Holmback, T. Akerstedt, and J. Axelsson, "Sleeping during the day: effects on the 24-h patterns of IGF-binding protein 1, insulin, glucose, cortisol, and growth hormone," *European Journal of Endocrinology*, vol. 163, no. 3, pp. 383–390, 2010.

[14] J. Arendt, "Shift work: coping with the biological clock," *Occupational Medicine*, vol. 60, no. 1, pp. 10–20, 2010.

[15] C. A. Crispim, J. Waterhouse, A. R. Dâmaso et al., "Hormonal appetite control is altered by shift work: a preliminary study," *Metabolism*, vol. 60, no. 12, pp. 1726–1735, 2011.

[16] E. Van Cauter and K. L. Knutson, "Sleep and the epidemic of obesity in children and adults," *European Journal of Endocrinology*, vol. 159, Supplement 1, pp. S59–S66, 2008.

[17] J. L. Marino, V. L. Holt, C. Chen, and S. Davis, "Shift work, hCLOCK T3111C polymorphism, and endometriosis risk," *Epidemiology*, vol. 19, no. 3, pp. 477–484, 2008.

[18] Y. Suwazono, M. Dochi, K. Sakata et al., "A longitudinal study on the effect of shift work on weight gain in male Japanese workers," *Obesity*, vol. 16, no. 8, pp. 1887–1893, 2008.

[19] S. Davis and D. K. Mirick, "Circadian disruption, shift work and the risk of cancer: a summary of the evidence and studies in Seattle," *Cancer Causes & Control*, vol. 17, no. 4, pp. 539–545, 2006.

[20] A. Knutsson and H. Bøggild, "Gastrointestinal disorders among shift workers," *Scandinavian Journal of Work, Environment & Health*, vol. 36, no. 2, pp. 85–95, 2010.

[21] S. Puttonen, T. Oksanen, J. Vahtera et al., "Is shift work a risk factor for rheumatoid arthritis? The Finnish Public Sector study," *Annals of the Rheumatic Diseases*, vol. 69, no. 4, pp. 679-680, 2010.

[22] M. F. Eldevik, E. Flo, B. E. Moen, S. Pallesen, and B. Bjorvatn, "Insomnia, excessive sleepiness, excessive fatigue, anxiety, depression and shift work disorder in nurses having less than 11 hours in-between shifts," *PLoS One*, vol. 8, no. 8, article e70882, 2013.

[23] R. Tamagawa, B. Lobb, and R. Booth, "Tolerance of shift work," *Applied Ergonomics*, vol. 38, no. 5, pp. 635–642, 2007.

[24] U. Albrecht, "Timing to perfection: the biology of central and peripheral circadian clocks," *Neuron*, vol. 74, no. 2, pp. 246–260, 2012.

[25] A. Iwamoto, M. Kawai, M. Furuse, and S. Yasuo, "Effects of chronic jet lag on the central and peripheral circadian clocks in CBA/N mice," *Chronobiology International*, vol. 31, no. 2, pp. 189–198, 2014.

[26] Y. Yamaguchi, T. Suzuki, Y. Mizoro et al., "Mice genetically deficient in vasopressin V1a and V1b receptors are resistant to jet lag," *Science*, vol. 342, no. 6154, pp. 85–90, 2013.

[27] W. Nakamura, S. Yamazaki, N. N. Takasu, K. Mishima, and G. D. Block, "Differential response of *Period 1* expression within the suprachiasmatic nucleus," *The Journal of Neuroscience*, vol. 25, no. 23, pp. 5481–5487, 2005.

[28] E. Abdullah, A. Idris, and A. Saparon, "PAPR reduction using SCS-SLM technique in STFBC MIMO-OFDM," *Journal of Engineering and Applied Science*, vol. 12, pp. 3218–3221, 2017.

[29] S. Sookoian, C. Gemma, T. Fernández Gianotti et al., "Effects of rotating shift work on biomarkers of metabolic syndrome and inflammation," *Journal of Internal Medicine*, vol. 261, no. 3, pp. 285–292, 2007.

[30] C. A. Landis, "Physiological and behavioral aspects of normal sleep," in *Sleep Disorders and Sleep Promotion in Nursing Practice*, N. Redeker and G. McEnany, Eds., pp. 1–18, Springer Publishing Company, LLC, New York, NY, USA, 2011.

[31] U. Ambrosius, S. Lietzenmaier, R. Wehrle et al., "Heritability of sleep electroencephalogram," *Biological Psychiatry*, vol. 64, no. 4, pp. 344–348, 2008.

[32] L. De Gennaro, C. Marzano, F. Fratello et al., "The electroencephalographic fingerprint of sleep is genetically determined: a twin study," *Annals of Neurology*, vol. 64, no. 4, pp. 455–460, 2008.

[33] T. Bjorness and R. Greene, "Adenosine and sleep," *Current Neuropharmacology*, vol. 7, no. 3, pp. 238–245, 2009.

[34] A. Sehgal and E. Mignot, "Genetics of sleep and sleep disorders," *Cell*, vol. 146, no. 2, pp. 194–207, 2011.

[35] C. Cirelli, "The genetic and molecular regulation of sleep: from fruit flies to humans," *Nature Reviews Neuroscience*, vol. 10, no. 8, pp. 549–560, 2009.

[36] D. M. Raizen, J. E. Zimmerman, M. H. Maycock et al., "Lethargus is a *Caenorhabditis elegans* sleep-like state," *Nature*, vol. 451, no. 7178, pp. 569–572, 2008.

[37] S. Langmesser, P. Franken, S. Feil, Y. Emmenegger, U. Albrecht, and R. Feil, "cGMP-dependent protein kinase type I is implicated in the regulation of the timing and quality of sleep and wakefulness," *PLoS One*, vol. 4, no. 1, article e4238, 2009.

[38] D. Bushey, R. Huber, G. Tononi, and C. Cirelli, "*Drosophila hyperkinetic* mutants have reduced sleep and impaired memory," *The Journal of Neuroscience*, vol. 27, no. 20, pp. 5384–5393, 2007.

[39] K. Koh, W. J. Joiner, M. N. Wu, Z. Yue, C. J. Smith, and A. Sehgal, "Identification of SLEEPLESS, a sleep-promoting factor," *Science*, vol. 321, no. 5887, pp. 372–376, 2008.

[40] A. C. Keene, E. R. Duboué, D. M. McDonald et al., "*Clock* and *cycle* limit starvation-induced sleep loss in *Drosophila*," *Current Biology*, vol. 20, no. 13, pp. 1209–1215, 2010.

[41] K. V. Allebrandt, N. Amin, B. Müller-Myhsok et al., "A $K_{ATP}$ channel gene effect on sleep duration: from genome-wide association studies to function in *Drosophila*," *Molecular Psychiatry*, vol. 18, no. 1, pp. 122–132, 2011.

[42] C. M. Depner, E. R. Stothard, and K. P. Wright, "Metabolic consequences of sleep and circadian disorders," *Current Diabetes Reports*, vol. 14, no. 7, p. 507, 2014.

[43] R. R. Markwald, E. L. Melanson, M. R. Smith et al., "Impact of insufficient sleep on total daily energy expenditure, food intake, and weight gain," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, no. 14, pp. 5695–5700, 2013.

[44] T. Miyagawa, M. Kawashima, N. Nishida et al., "Variant between *CPT1B* and *CHKB* associated with susceptibility to narcolepsy," *Nature Genetics*, vol. 40, no. 11, pp. 1324–1328, 2008.

[45] H. Toyoda, Y. Honda, S. Tanaka et al., "Narcolepsy susceptibility gene *CCR3* modulates sleep-wake patterns in mice," *PLos One*, vol. 12, no. 11, article e0187888, 2017.

[46] J. Winkelmann, D. Czamara, B. Schormair et al., "Correction: genome-wide association study identifies novel restless legs syndrome susceptibility loci on 2p14 and 16q12.1," *PLoS Genetics*, vol. 7, 2011.

[47] H. Stefansson, D. B. Rye, A. Hicks et al., "A genetic risk factor for periodic limb movements in sleep," *The New England Journal of Medicine*, vol. 357, no. 7, pp. 639–647, 2007.

[48] J. Winkelmann, P. Lichtner, B. Schormair et al., "Variants in the neuronal nitric oxide synthase (nNOS, NOS1) gene are associated with restless legs syndrome," *Movement Disorders*, vol. 23, no. 3, pp. 350–358, 2008.

[49] L. Imeri and M. R. Opp, "How (and why) the immune system makes us sleep," *Nature Reviews Neuroscience*, vol. 10, no. 3, pp. 199–210, 2009.

[50] J. Rihel, D. A. Prober, A. Arvanites et al., "Zebrafish behavioral profiling links drugs to biological targets and rest/wake regulation," *Science*, vol. 327, no. 5963, pp. 348–351, 2010.

[51] H. Funato, C. Miyoshi, T. Fujiyama et al., "Forward-genetics analysis of sleep in randomly mutagenized mice," *Nature*, vol. 539, no. 7629, pp. 378–383, 2016.

[52] L. K. Barger, S. M. W. Rajaratnam, C. P. Cannon et al., "Short sleep duration, obstructive sleep apnea, shiftwork, and the risk of adverse cardiovascular events in patients after an acute coronary syndrome," *Journal of the American Heart Association*, vol. 6, no. 10, article e006959, 2017.

[53] A. J. R. Lim, Z. Huang, S. E. Chua, M. S. Kramer, and E. L. Yong, "Sleep duration, exercise, shift work and polycystic ovarian syndrome-related outcomes in a healthy population: a cross-sectional study," *PLoS One*, vol. 11, no. 11, article e0167048, 2016.

[54] S. Lee, H.-R. Kim, J. Byun, and T. Jang, "Sleepiness while driving and shiftwork patterns among Korean bus drivers," *Annals of Occupational and Environmental Medicine*, vol. 29, no. 1, 2017.

[55] E. Espitia-Bautista, M. Velasco-Ramos, I. Osnaya-Ramírez, M. Ángeles-Castellanos, R. M. Buijs, and C. Escobar, "Social jet-lag potentiates obesity and metabolic syndrome when combined with cafeteria diet in rats," *Metabolism*, vol. 72, pp. 83–93, 2017.

[56] H. Oike, M. Sakurai, K. Ippoushi, and M. Kobori, "Time-fixed feeding prevents obesity induced by chronic advances of light/dark cycles in mouse models of jet-lag/shift work," *Biochemical and Biophysical Research Communications*, vol. 465, no. 3, pp. 556–561, 2015.

[57] J. Bass and J. S. Takahashi, "Circadian integration of metabolism and energetics," *Science*, vol. 330, no. 6009, pp. 1349–1354, 2010.

[58] J. S. Takahashi, H.-K. Hong, C. H. Ko, and E. L. McDearmon, "The genetics of mammalian circadian order and disorder: implications for physiology and disease," *Nature Reviews Genetics*, vol. 9, no. 10, pp. 764–775, 2008.

[59] L. Zhang, B. Y. Chung, B. C. Lear et al., "$DN1_p$ circadian neurons coordinate acute light and PDF inputs to produce robust daily behavior in *Drosophila*," *Current Biology*, vol. 20, no. 7, pp. 591–599, 2010.

[60] K. A. Lamia, K.-F. Storch, and C. J. Weitz, "Physiological significance of a peripheral tissue circadian clock," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 39, pp. 15172–15177, 2008.

[61] F. W. Turek, C. Joshu, A. Kohsaka et al., "Obesity and metabolic syndrome in circadian *Clock* mutant mice," *Science*, vol. 308, no. 5724, pp. 1043–1045, 2005.

[62] F. A. J. L. Scheer, M. F. Hilton, C. S. Mantzoros, and S. A. Shea, "Adverse metabolic and cardiovascular consequences of circadian misalignment," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 11, pp. 4453–4458, 2009.

[63] A. Reinberg and I. Ashkenazi, "Internal desynchronization of circadian rhythms and tolerance to shift work," *Chronobiology International*, vol. 25, no. 4, pp. 625–643, 2008.

[64] C. M. Pariante, "Risk factors for development of depression and psychosis," *Annals of the New York Academy of Sciences*, vol. 1179, no. 1, pp. 144–152, 2009.

[65] C. Campino, F. Valenzuela, E. Arteaga et al., "Melatonin reduces cortisol response to ACTH in humans," *Revista Médica de Chile*, vol. 136, no. 11, pp. 1390–1397, 2008.

[66] C. M. Pariante, "Pituitary volume in psychosis: the first review of the evidence," *Journal of Psychopharmacology*, vol. 22, 2 Supplement, pp. 76–81, 2008.

[67] M. Vogel, T. Braungardt, W. Meyer, and W. Schneider, "The effects of shift work on physical and mental health," *Journal of Neural Transmission*, vol. 119, no. 10, pp. 1121–1132, 2012.

[68] P. Boscolo, M. Di Gioacchino, M. Reale, R. Muraro, and L. Di Giampaolo, "Work stress and innate immune response," *International Journal of Immunopathology and Pharmacology*, vol. 24, pp. 51S–54S, 2011.

[69] S. R. Pandi-Perumal, V. Srinivasan, D. W. Spence, and D. P. Cardinali, "Role of the melatonin system in the control of sleep," *CNS Drugs*, vol. 21, no. 12, pp. 995–1018, 2007.

[70] S. A. Rahman, S. Marcu, L. Kayumov, and C. M. Shapiro, "Altered sleep architecture and higher incidence of subsyndromal depression in low endogenous melatonin secretors," *European Archives of Psychiatry and Clinical Neuroscience*, vol. 260, no. 4, pp. 327–335, 2010.

[71] L. Di Milia, S. Waage, S. Pallesen, and B. Bjorvatn, "Shift work disorder in a random population sample – prevalence and comorbidities," *PLoS One*, vol. 8, no. 1, article e55306, 2013.

[72] K. Mishima, T. Tozawa, K. Satoh, H. Saitoh, and Y. Mishima, "The 3111T/C polymorphism of *hClock* is associated with evening preference and delayed sleep timing in a Japanese population sample," *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, vol. 133B, no. 1, pp. 101–104, 2005.

[73] M. R. Smith and C. I. Eastman, "Shift work: health, performance and safety problems, traditional countermeasures, and innovative management strategies to reduce circadian misalignment," *Nature and Science of Sleep*, vol. 4, pp. 111–132, 2012.

[74] J. D. Carpen, S. N. Archer, D. J. Skene, M. Smits, and M. Schantz, "A single-nucleotide polymorphism in the $5'$-untranslated region of the *hPER2* gene is associated with diurnal preference," *Journal of Sleep Research*, vol. 14, no. 3, pp. 293–297, 2005.

[75] D. S. Pereira, S. Tufik, F. M. Louzada et al., "Association of the length polymorphism in the human Per3 gene with the delayed sleep-phase syndrome: does latitude have an influence upon it?," *Sleep*, vol. 28, no. 1, pp. 29–32, 2005.

[76] R. R. Markwald and K. P. Wright Jr., "Circadian misalignment and sleep disruption in shift work: Implications for fatigue and risk of weight gain and obesity," in *Sleep Loss and Obesity*, Springer, New York, NY, USA, 2012.

[77] O. M. Buxton, S. W. Cain, S. P. O'Connor et al., "Adverse metabolic consequences in humans of prolonged sleep restriction combined with circadian disruption," *Science Translational Medicine*, vol. 4, no. 129, article 129ra43, 2012.

[78] C. L. Drake and K. P. Wright, "Shift work, shift-work disorder, and jet lag," *Principles and Practice of Sleep Medicine*, vol. 1, pp. 784–798, 2011.

[79] J. S. Ruggiero and N. S. Redeker, "Effects of napping on sleepiness and sleep-related performance deficits in night-shift workers: a systematic review," *Biological Research for Nursing*, vol. 16, no. 2, pp. 134–142, 2014.

[80] B. Marcheva, K. M. Ramsey, E. D. Buhr et al., "Disruption of the clock components CLOCK and BMAL1 leads to hypoinsulinaemia and diabetes," *Nature*, vol. 466, no. 7306, pp. 627–631, 2010.

[81] R. L. Sack, D. Auckley, R. R. Auger et al., "Circadian rhythm sleep disorders: part I, basic principles, shift work and jet lag disorders," *Sleep*, vol. 30, no. 11, pp. 1460–1483, 2007.

[82] K. P. Wright Jr., R. K. Bogan, and J. K. Wyatt, "Shift work and the assessment and management of shift work disorder (SWD)," *Sleep Medicine Reviews*, vol. 17, no. 1, pp. 41–54, 2013.

[83] Y. Zhao and F. X. Castellanos, "Annual research review: discovery science strategies in studies of the pathophysiology of child and adolescent psychiatric disorders - promises and limitations," *The Journal of Child Psychology and Psychiatry*, vol. 57, no. 3, pp. 421–439, 2016.

[84] Z. Hawi, T. D. R. Cummins, J. Tong et al., "The molecular genetic architecture of attention deficit hyperactivity disorder," *Molecular Psychiatry*, vol. 20, no. 3, pp. 289–297, 2015.

[85] S. S. Jeste and D. H. Geschwind, "Disentangling the heterogeneity of autism spectrum disorder through genetic findings," *Nature Reviews Neurology*, vol. 10, no. 2, pp. 74–81, 2014.

[86] M. P. Milham, "Open neuroscience solutions for the connectome-wide association era," *Neuron*, vol. 73, no. 2, pp. 214–218, 2012.

[87] R. T. de Sousa, A. A. Loch, A. F. Carvalho et al., "Genetic studies on the tripartite glutamate synapse in the pathophysiology and therapeutics of mood disorders," *Neuropsychopharmacology*, vol. 42, no. 4, pp. 787–800, 2017.

[88] M. Burmeister, M. G. McInnis, and S. Zöllner, "Psychiatric genetics: progress amid controversy," *Nature Reviews Genetics*, vol. 9, no. 7, pp. 527–540, 2008.

[89] A. B. Amstadter, H. H. Maes, C. M. Sheerin, J. M. Myers, and K. S. Kendler, "The relationship between genetic and environmental influences on resilience and on common internalizing and externalizing psychiatric disorders," *Social Psychiatry and Psychiatric Epidemiology*, vol. 51, no. 5, pp. 669–678, 2016.

[90] N. Sawamura, T. Ando, Y. Maruyama et al., "Nuclear DISC1 regulates CRE-mediated gene transcription and sleep homeostasis in the fruit fly," *Molecular Psychiatry*, vol. 13, no. 12, pp. 1138–1148, 2008.

[91] T. Kieseppä, T. Partonen, J. Haukka, J. Kaprio, and J. Lönnqvist, "High concordance of bipolar I disorder in a nationwide sample of twins," *The American Journal of Psychiatry*, vol. 161, no. 10, pp. 1814–1821, 2004.

[92] T. W. Mühleisen, M. Leber, T. G. Schulze et al., "Genome-wide association study reveals two new risk loci for bipolar disorder," *Nature Communications*, vol. 5, 2014.

[93] N. Craddock and P. Sklar, "Genetics of bipolar disorder: successful start to a long journey," *Trends in Genetics*, vol. 25, no. 2, pp. 99–105, 2009.

[94] Psychiatric GWAS Consortium Bipolar Disorder Working Group, "Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near *ODZ4*," *Nature Genetics*, vol. 43, pp. 977–983, 2012.

[95] J. M. Gatt, K. L. O. Burton, L. M. Williams, and P. R. Schofield, "Specific and common genes implicated across major mental disorders: a review of meta-analysis studies," *Journal of Psychiatric Research*, vol. 60, pp. 1–13, 2015.

[96] T. Kato, "Whole genome/exome sequencing in mood and psychotic disorders," *Psychiatry and Clinical Neurosciences*, vol. 69, no. 2, pp. 65–76, 2015.

[97] R. C. Kessler, W. T. Chiu, O. Demler, K. R. Merikangas, and E. E. Walters, "Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the national comorbidity survey replication," *Archives of General Psychiatry*, vol. 62, no. 6, pp. 617–627, 2005.

[98] K. Koido, T. Traks, R. Balõtšev et al., "Associations between *LSAMP* gene polymorphisms and major depressive disorder and panic disorder," *Translational Psychiatry*, vol. 2, no. 8, article e152, 2012.

[99] S. Moylan, M. Maes, N. R. Wray, and M. Berk, "The neuroprogressive nature of major depressive disorder: pathways to disease evolution and resistance, and therapeutic implications," *Molecular Psychiatry*, vol. 18, no. 5, pp. 595–606, 2013.

[100] D. F. Levinson, "The genetics of depression: a review," *Biological Psychiatry*, vol. 60, no. 2, pp. 84–92, 2006.

[101] S. López-León, A. C. J. W. Janssens, A. M. González-Zuloeta Ladd et al., "Meta-analyses of genetic studies on major depressive disorder," *Molecular Psychiatry*, vol. 13, no. 8, pp. 772–785, 2008.

[102] M. A. Kohli, S. Lucae, P. G. Saemann et al., "The neuronal transporter gene *SLC6A15* confers risk to major depression," *Neuron*, vol. 70, no. 2, pp. 252–265, 2011.

[103] L. Fritschi, D. C. Glass, J. S. Heyworth et al., "Hypotheses for mechanisms linking shiftwork and cancer," *Medical Hypotheses*, vol. 77, no. 3, pp. 430–436, 2011.

[104] S. Khan, M. W. Ullah, R. Siddique et al., "Role of recombinant DNA technology to improve life," *International Journal of Genomics*, vol. 2016, Article ID 2405954, 14 pages, 2016.

[105] S. Xiang, S. B. Coffelt, L. Mao, L. Yuan, Q. Cheng, and S. M. Hill, "Period-2: a tumor suppressor gene in breast cancer," *Journal of Circadian Rhythms*, vol. 6, no. 1, p. 4, 2008.

[106] Y. Zhu, R. G. Stevens, A. E. Hoffman et al., "Epigenetic impact of long-term shiftwork: pilot evidence from circadian genes and whole-genome methylation analysis," *Chronobiology International*, vol. 28, no. 10, pp. 852–861, 2011.

[107] E. Filipski, P. F. Innominato, M. Wu et al., "Effects of light and food schedules on liver and tumor molecular clocks in mice," *Journal of the National Cancer Institute*, vol. 97, no. 7, pp. 507–517, 2005.

[108] J. Samulin Erdem, H. Ø. Notø, Ø. Skare et al., "Mechanisms of breast cancer risk in shift workers: association of telomere shortening with the duration and intensity of night work," *Cancer Medicine*, vol. 6, no. 8, pp. 1988–1997, 2017.

[109] P. A. Wood, X. Yang, A. Taber et al., "*Period 2* mutation accelerates $Apc^{Min/+}$ tumorigenesis," *Molecular Cancer Research*, vol. 6, no. 11, pp. 1786–1793, 2008.

[110] E. E. Flynn-Evans, L. Mucci, R. G. Stevens, and S. W. Lockley, "Shiftwork and prostate-specific antigen in the national health and nutrition examination survey," *Journal of the National Cancer Institute*, vol. 105, no. 17, pp. 1292–1297, 2013.

[111] S. Lee, L. A. Donehower, A. J. Herron, D. D. Moore, and L. Fu, "Disrupting circadian homeostasis of sympathetic signaling promotes tumor development in mice," *PLoS One*, vol. 5, no. 6, article e10995, 2010.

[112] X. Yang, P. A. Wood, E.-Y. Oh, J. Du-Quiton, C. M. Ansell, and W. J. M. Hrushesky, "Down regulation of circadian clock gene period 2 accelerates breast cancer growth by altering its daily growth rhythm," *Breast Cancer Research and Treatment*, vol. 117, no. 2, pp. 423–431, 2009.

[113] S. Gery, N. Komatsu, L. Baldjyan, A. Yu, D. Koo, and H. P. Koeffler, "The circadian gene per1 plays an important role in cell growth and DNA damage control in human cancer cells," *Molecular Cell*, vol. 22, no. 3, pp. 375–382, 2006.

[114] S. B. Reeder, H. H. Hu, and C. B. Sirlin, "Proton density fat-fraction: a standardized MR-based biomarker of tissue fat concentration," *Journal of Magnetic Resonance Imaging*, vol. 36, no. 5, pp. 1011–1014, 2012.

[115] T. Papagiannakopoulos, M. R. Bauer, S. M. Davidson et al., "Circadian rhythm disruption promotes lung tumorigenesis," *Cell Metabolism*, vol. 24, no. 2, pp. 324–331, 2016.

[116] A. Blanch, B. Torrelles, A. Aluja, and J. A. Salinas, "Age and lost working days as a result of an occupational accident: a study in a shiftwork rotation system," *Safety Science*, vol. 47, no. 10, pp. 1359–1363, 2009.

[117] H. Admi, O. Tzischinsky, R. Epstein, P. Herer, and P. Lavie, "Shift work in nursing: is it really a risk factor for nurses'

health and patients' safety?," *Nursing Economics*, vol. 26, no. 4, pp. 250–257, 2008.

[118] F. Tüchsen, K. B. Christensen, and T. Lund, "Shift work and sickness absence," *Occupational Medicine*, vol. 58, no. 4, pp. 302–304, 2008.

[119] M. Takahashi, T. Tanigawa, N. Tachibana et al., "Modifying effects of perceived adaptation to shift work on health, wellbeing, and alertness on the job among nuclear power plant operators," *Industrial Health*, vol. 43, no. 1, pp. 171–178, 2005.

[120] L. Di Milia, P. A. Smith, and S. Folkard, "A validation of the revised circadian type inventory in a working sample," *Personality and Individual Differences*, vol. 39, no. 7, pp. 1293–1305, 2005.

Hindawi
International Journal of Genomics
Volume 2018, Article ID 8124950, 14 pages
https://doi.org/10.1155/2018/8124950

*Research Article*

# Ensemble Methods with Voting Protocols Exhibit Superior Performance for Predicting Cancer Clinical Endpoints and Providing More Complete Coverage of Disease-Related Genes

**Runyu Jing,[1] Yu Liang,[2] Yi Ran,[3] Shengzhong Feng,[1] Yanjie Wei,[1] and Li He[3]**

[1]*Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China*
[2]*College of Chemistry, Sichuan University, Chengdu 610064, China*
[3]*Biogas Appliance Quality Supervision and Inspection Center, Biogas Institute of Ministry of Agriculture, Chengdu, Sichuan, China*

Correspondence should be addressed to Yanjie Wei; yj.wei@siat.ac.cn and Li He; helibiogas@126.com

Received 27 July 2017; Revised 6 November 2017; Accepted 14 November 2017; Published 10 January 2018

Academic Editor: Zhichao Liu

Copyright © 2018 Runyu Jing et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In genetic data modeling, the use of a limited number of samples for modeling and predicting, especially well below the attribute number, is difficult due to the enormous number of genes detected by a sequencing platform. In addition, many studies commonly use machine learning methods to evaluate genetic datasets to identify potential disease-related genes and drug targets, but to the best of our knowledge, the information associated with the selected gene set was not thoroughly elucidated in previous studies. To identify a relatively stable scheme for modeling limited samples in the gene datasets and reveal the information that they contain, the present study first evaluated the performance of a series of modeling approaches for predicting clinical endpoints of cancer and later integrated the results using various voting protocols. As a result, we proposed a relatively stable scheme that used a set of methods with an ensemble algorithm. Our findings indicated that the ensemble methodologies are more reliable for predicting cancer prognoses than single machine learning algorithms as well as for gene function evaluating. The ensemble methodologies provide a more complete coverage of relevant genes, which can facilitate the exploration of cancer mechanisms and the identification of potential drug targets.

## 1. Introduction

With the development of genetic sequencing technology, genetic information could be recorded as gene expression data. Data mining, such as using machine learning methods, is commonly used to reveal latent correlations between diseases and gene expression. Supervised or semisupervised machine learning algorithms have been proposed to predict the clinical outcomes of cancers [1–4] in the context of tumorigenesis. The support vector machine (SVM) [5–8] and artificial neural network (ANN) [9–11] algorithms were the most commonly used approaches for predicting prognoses. In addition, the Bayesian probability model [12–14] and the fuzzy neural network [15] were also used for cancer prognosis prediction. The microarray quality control (MAQC) project thoroughly investigated the performance of models for the prediction of clinical outcomes of breast cancer, multiple myeloma, and neuroblastoma and were common practices for microarray-based model construction and validation [16]. The network-based approaches have seen a recent widespread use for the identification of cancer-related genes and have revealed the molecular mechanisms of various cancers [17–22]. However, to the best of our knowledge, no studies have examined multiple algorithms with two different kinds of expression data and their ensemble performance with a limited number of samples, which might be crucial when using them in practical. In most cases, the number of available samples is restricted due to the

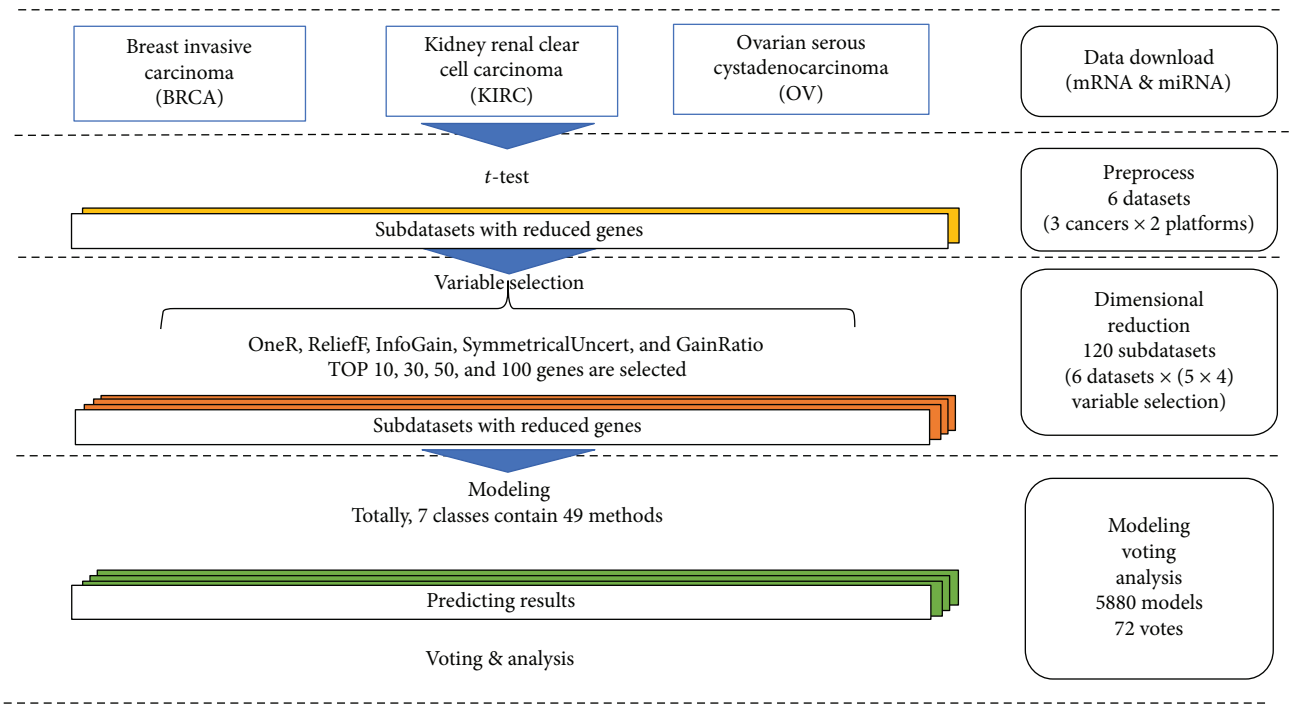



Figure 1: Work flow of the whole process. First, the datasets were downloaded from the GDC (Genomic Data Commons) database. Next, the downloaded mRNA and microRNA sequencing data are united by the usable information. The *t*-test was used afterwards to determine the significantly expressed genes. Five selection methods were used to select the cancer-associated genes and the subdatasets generated according to the ranks. Finally, the prediction results were integrated by a voting protocol. Note that every subdataset was divided into two pieces for cross-validation and independent test in the ratio 4 : 1 before variable selection. Only the datasets for cross-validation will be used for variable selection and modeling.

cost, privacy, and other reasons. A limited sample number causes a predictive model to be more sensitive to the dataset distribution, and the lack of prior knowledge simultaneously reduces the overall predicting performance which could be reflected by MCC (Matthews correlation coefficient). Moreover, a single machine learning algorithm provides insufficient coverage of the disease-related genes because it only uses genes that show the greatest difference of expression profiles between the phenotypic statuses compared when the genes have a similar function. Thus, identifying a stable predictive model using a limited number of samples becomes a challenge.

To address this problem, the present study thoroughly investigated the performance of single models among different datasets and proposed a strategy to combine multiple predictive models as well as the datasets into a final ensemble for clinical prediction. Compared with a single machine learning algorithm, an ensemble scheme could not only perform more reliably when predicting clinical endpoints but could also provide broader coverage of disease-related genes, which will be beneficial for further downstream analysis in such applications as the identification of potential drugs.

## 2. Materials and Methods

The workflow of this study is listed in Figure 1. The datasets were carefully generated such that the scale and representation of the samples in the different datasets are consistent such that the predictions were comparable. This section describes the data and methods used.

*2.1. GDC Data.* All data were downloaded from the NCI's Genomic Data Commons (GDC) [23] by using the official web-based API (https://gdc.cancer.gov/developers/gdc-application-programming-interface-api). The genomic data were from the official normalized microRNA and RNA sequence expression data because it was restricted by multiplatform coverage and accessibility; a portion of the cancer data was excluded from this study. For example, the number of available samples of neuroblastoma in GDC is 1127 (this number might change if the database is updated), but only approximately half of these samples have associated RNA sequence data, and no microRNA sequence data are available. Finally, the freely accessible data for breast invasive carcinoma (BRCA), ovarian serous cystadenocarcinoma (OV), and kidney renal clear cell carcinoma (KIRC) were downloaded and used for modeling. The clinical information was downloaded in the XML format. The relationships of samples from different platforms (such as mRNA, microRNA, and clinical) were identified by the official MetaData file in the JSON format. The detailed distribution can be found in Table 1.

*2.2. Preprocessing.* Since we wanted to make the results from two sequencing platforms (e.g., mRNA and microRNA) comparable and able for voting at last, the selection of samples was determined by the integration of different platforms.

Table 1: Scales of the three datasets.

| Disease name | Sample number in modeling dataset | Sample number in independent test dataset | Number of kept genes | |
|---|---|---|---|---|
| BRCA | 558 | 141 | mRNA | 24585 |
| | | | miRNA | 722 |
| KIRC | 112 | 29 | mRNA | 9119 |
| | | | miRNA | 190 |
| OV | 66 | 18 | mRNA | 4390 |
| | | | miRNA | 83 |

Only samples that contained clinical information and expression data for both mRNA and microRNA were retained for subsequent analysis.

The sample label was determined by the clinical information. For OV and KIRC, the label was determined by the survival time. A sample was deemed positive if the recorded survival time was larger than one year and if the patient was still alive (based on the clinical data). Similarly, a sample was deemed negative if the associated patient was dead and if the recorded survival time was less than one year. The sample label for BRCA was determined by the estrogen receptor (ER) status (negative or positive), which was also recorded in the clinical XML file. Therefore, if the required information for a sample could not be found in the clinical data or its status did not satisfy the criteria (e.g., the patient was alive but the survival time was less than one year), it was excluded.

The sample number was reduced by the clinical information. Student's *t*-test was used to subsequently reduce the mRNA and microRNA numbers. Only genes that had significant expression with a $p$ value less than 0.05 were retained in a dataset. The ratio of positive to negative samples was kept in an appropriate range to reduce classification bias. In this study, the range of this ratio was 0.5 to 2. For example, if a dataset contained 22 positive and 50 negative samples, 6 negative samples were randomly removed to adjust the ratio so it fell within the required range. The eliminated datasets were divided into two parts for cross-validation and independent tests in a 4 : 1 ratio. The scale of the datasets is listed in Table 1. Only the datasets for cross-validation will be used for variable selection and the 5-fold cross-validation; the datasets for independent testing will not participate in modeling. And only the independent prediction will be used for further ensemble analysis and comparison.

*2.3. Machine Learning Methods.* 49 modeling methods in WEKA [24] (version 3.8.1) were investigated in this study. The methods were divided into seven different classes by the developers of WEKA according to specific features of the methods (Table 2). The different method classes had different features. In the *functions* class, most of the methods use a functional solution for modeling the data, such as *LibSVM* [25] and logistic regression [26], and in most cases, few mechanisms are available for ensemble learning, such as voting or resampling. However, in the *meta* class, the methods use resampling and voting for classification and regression, and the methods in the other classes are considered model units, such as *AdaBoost* [27] and *Bagging* [28]. The methods from the *bayes* class are from the probability and graph theory, and most of them, including *NaïveBayes* [29], *BayesNetwork* [30], and *BayesianLogisticRegression* [31], are sensitive to sample number. Similarly, methods in the *rules* class use rules (such as decision table) for classification [32]. The methods in the *lazy* class are instance-based and could be optimized for better efficiency using a lazy algorithm [33]. Most of the methods in the *trees* class are based on the classification and regression tree algorithm, but the way they are carried out is different. Many other mechanisms are integrated into the *trees* such as resampling used in a random forest [34]. Finally, the *misc* class contains methods for which it is difficult to assign to another class. Only two methods fell into this class in this study, namely, the *VFI*, an ensemble method based on a voting protocol [35], and the *HyperPipes*, based on an algorithm that finds similarities among attributes. Considering running time, comparability, and reducing the risk of overfitting, only default parameters are used for modeling.

Because the sample number was limited, fewer genes should be considered to avoid overfitting. In this study, five variable selection methods were used for dimensional reduction: *OneR*, *ReliefF*, *InfoGain*, *SymmetricalUncert*, and *GainRatio*. *OneR* is executed by using the *OneR* classifier, which is based on measuring the error between the attributes and the response values [36]. *ReliefF* uses a resampling mechanism for evaluating the attributes [37]. The other methods are from the information theory, and the associated formulas are

$$\begin{aligned}
\mathrm{GainRatio} = (\mathrm{class}, \mathrm{attribute}) &= \frac{H(\mathrm{class}) - H(\mathrm{class}|\mathrm{attribute})}{H(\mathrm{attribute})},\\
\mathrm{InfoGain}(\mathrm{class}, \mathrm{attribute}) &= H(\mathrm{class}) - H(\mathrm{class}|\mathrm{attribute}),\\
\mathrm{SymmetricalUncert}(\mathrm{class}, \mathrm{attribute}) &= 2\frac{H(\mathrm{class}) - H(\mathrm{class}|\mathrm{attribute})}{H(\mathrm{class})}\\
&\quad + H(\mathrm{attribute}),
\end{aligned} \tag{1}$$

where the "$H()$" in the formula is the information entropy (Shannon entropy) [38] and "class" denotes the values of a label. According to their formulas, *GainRatio* could be considered as a normalization of *InfoGain*. However, the information entropies of all of the attributes, such as the expression of mRNAs and microRNAs, are different, so both methods are used in this study. By using the ranking mechanism in WEKA, in every subdataset for cross-validation, the attributes can be ranked, and the top 10, 30, 50, and 100 ranked attributes are selected for cross-validation and modeling. Note that since the number of microRNAs in OV is limited (83, which is less than 100), the numbers of the attributes in the subdatasets are 10, 30, 50, and 83. Totally, we investigated a total of 5 variable selection methods × 4 subdatasets × 49 modeling methods = 980 predictive models.

Table 2: Methods used.

| Class | Method names |
|---|---|
| *bayes* | NaiveBayes, BayesianLogisticRegression, BayesNet, ComplementNaiveBayes, DMNBtext, NaiveBayesMultinomial, NaiveBayesMultinomialUpdateable |
| *functions* | Logistic, MultilayerPerceptron, RBFNetwork, SimpleLogistic, SPegasos, VotedPerceptron, LibSVM |
| *lazy* | IB1, IBk, KStar, LWL |
| *meta* | AdaBoostM1, Bagging, Dagging, Decorate, END, FilteredClassifier, LogitBoost, MultiBoostAB, MultiClassClassifier |
| *misc* | HyperPipes, VFI |
| *rules* | ConjunctiveRule, DecisionTable, DTNB, NNge, OneR, PART, Ridor, ZeroR |
| *trees* | ADTree, BFTree, DecisionStump, FT, J48, J48graft, LADTree, LMT, NBTree, RandomForest, RandomTree, REPTree |

The names, including the class names, are from WEKA. The results and discussion are based on the classes.

*2.4. Integrating the Predictions by Voting.* After generating hundreds of models, it is possible to combine their predictions. As previously mentioned, the prediction performances are ranged in the datasets. To integrate the ranged predictions and find a stable modeling method for genetic datasets, we used a voting protocol in this study to identify the datasets.

All of the weights were the same, except in the information theory methods such as *InfoGain*, *SymmetricalUncert*, and *GainRatio*. The weights in the information theory methods were modified to 1/3 when voting the predictions from the subdatasets of the OV-miRNA group according to the prediction distribution. More details may be found in "the coverage and reliability of selected genes" and "the distributions of the predictions."

Three voting schemes were used to arrive at a comprehensive conclusion. All of the methods were first used for voting. Then, some of the methods which performed poorly for all datasets were eliminated. Finally, mRNA and miRNA datasets were combined for voting.

*2.5. Measurement Methods.* Since there were many predictions, box plots were used to reflect the stability of the different classes. The quartiles Q1 and Q3, the interquartile range (IQR), and the whiskers (the lower whisker is Q1 −1.5 IQR, and upper whisker is Q3 +1.5 IQR) in the box plot are discussed. Two types of box plots were used to present the results in different angles, one based on the WEKA classes of modeling methods and another based on the scale of the subdatasets. Because the ratio of positive to negative samples was biased, the Matthews correlation coefficient (MCC) was used as the criterion for the plots. The MCC is one of the criteria used to evaluate the prediction performance, and the associated formula is

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FN}) \times (\mathrm{TP} + \mathrm{FP}) \times (\mathrm{TN} + \mathrm{FP}) \times (\mathrm{TN} + \mathrm{FN})}}, \quad (2)$$

where TP, TN, FP, and FN are the number of true-positive predictions, true-negative predictions, false-positive predictions, and false-negative predictions, respectively.

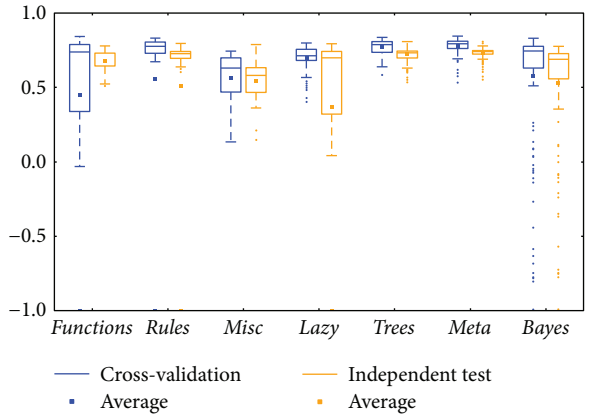



Figure 2: MCC of the BRCA-mRNA group by the *functions* class.

## 3. Results

Figure 1 is a flow chart that shows the modeling and integration of the preprocessed datasets. The associated results are listed below. Since there were 5880 modules in total (6 datasets × 980 predictive models per datasets), figures instead of tables were used to present the results (Figures 2–13). The individual cross-validation and independent test results were together listed in the Supplementary file "ModelingResults.xlsx" (available here).

*3.1. Modeling Results.* The prediction performance was ascertained in two ways: by the modeling method class defined by WEKA (see Table 2) and by the different subdatasets generated by the different variable selection methods. Therefore, a total of 3 cancers × 2 sequence methods × 2 kinds of plots = 12 figures were generated for the modeling results.

The *meta* class and *trees* class methods performed better than those in other classes for the two types of BRCA genomics data (mRNA in Figure 2 and miRNA in Figure 4), as evidenced by the best medians and averages. The box for the *trees* class had higher whiskers, but the box for the *meta* class had a smaller IQR.

The KIRC mRNA (Figure 6) and miRNA (Figure 8) datasets showed diverse prediction distributions. In the KIRC-mRNA group, the distributions were similar in the two BRCA groups but the *lazy* class performed in a similar manner in the *meta* and *trees* classes. The distributions were

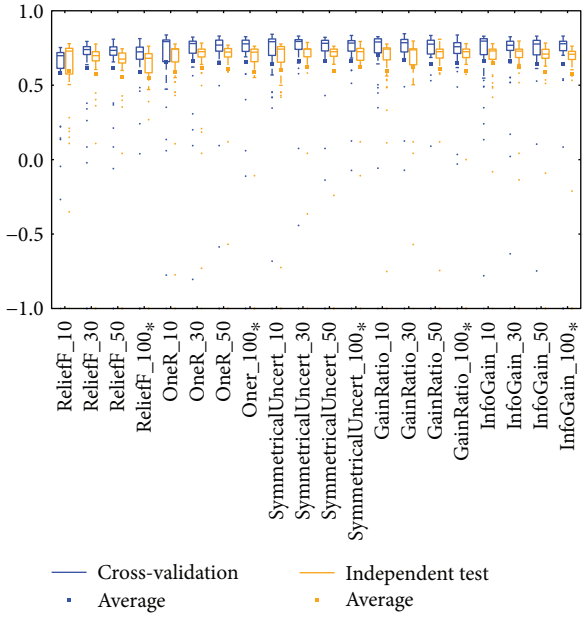



Figure 3: MCC of the BRCA-mRNA group by reduced datasets. *Note that the subdatasets from OV-miRNA have at most 83 micro-RNAs and thus the scale "100" of OV means 83.

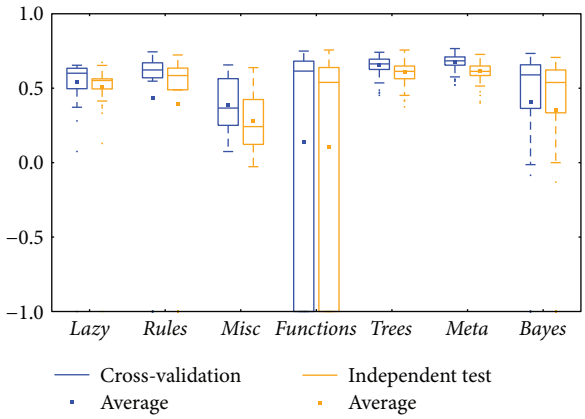



Figure 4: MCC of the BRCA-miRNA group by the *functions* class.

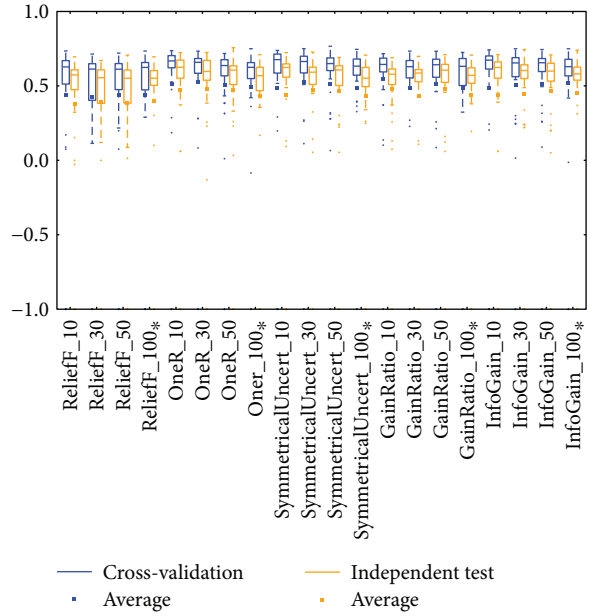



Figure 5: MCC of the BRCA-miRNA group by reduced datasets. *Note that the subdatasets from OV-miRNA have at most 83 micro-RNAs and thus the scale "100" of OV means 83.

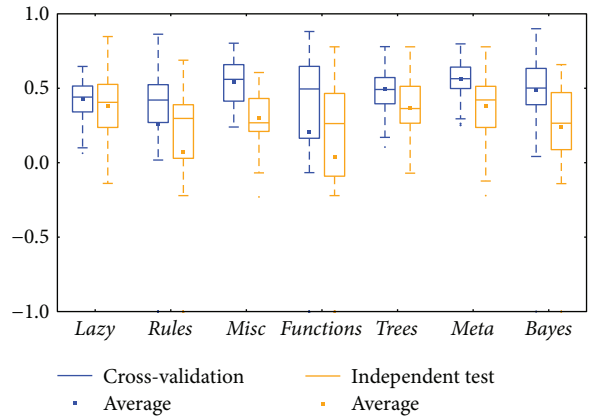



Figure 6: MCC of the KIRC-mRNA group by the *functions* class.

totally different in the KIRC-miRNA class. The *misc* class had the best distribution in the box plot, and the others were far worse. However, most of the medians and averages in the KIRC-miRNA group were less than those in the KIRC-mRNA group.

The situation was similar in the OV-mRNA group (Figure 10) and the OV-miRNA group (Figure 12). The *misc* class was best in the OV-mRNA group datasets, but the *meta* class had relatively better prediction performance for the OV-miRNA group. The *rules* class and *functions* class had very poor distributions in both groups.

Comparing between cross-validation and independent test, most of them were similar; the differences between the whiskers were not huge except for two classes: *functions* and *lazy* (could be found in Figures 2, 8, and 12). The *functions* class often has wider ranges between the two whiskers in cross-validation and is still larger than its independent test. On the contrary, the *lazy* class usually has narrower ranges in cross-validation than its independent test.

Distributions based on different datasets were clearer. In most cases, the dataset (mRNA or microRNA) that contained more attributes had a better distribution (i.e., a smaller IQR or a higher median). However, the peak was sometimes found in a smaller dataset. In more detail, the two BRCA groups (Figures 3 and 5) seemed to be insensitive to the attribute number and the distributions especially medians are very similar. The boxes were larger in the KIRC-miRNA group (Figure 9). The prediction results were sensitive to the attribute number, and the upper whisker increased as the attributes increased, but the lower whisker simultaneously decreased. In comparison, the distributions of the boxes in the KIRC-mRNA group (Figure 7) were better when the attribute number increased to 30 and were relatively stable afterwards. Differently in the OV-mRNA (Figure 11) and

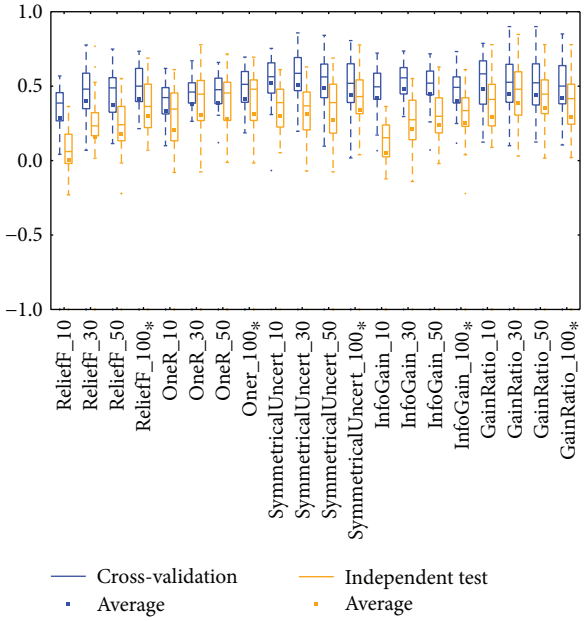



Figure 7: MCC of the KIRC-mRNA group by reduced datasets. *Note that the subdatasets from OV-miRNA have at most 83 micro-RNAs and thus the scale "100" of OV means 83.

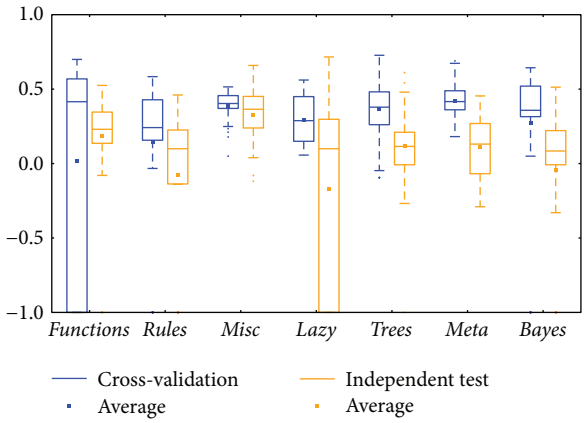



Figure 8: MCC of the KIRC-miRNA group by the *functions* class.

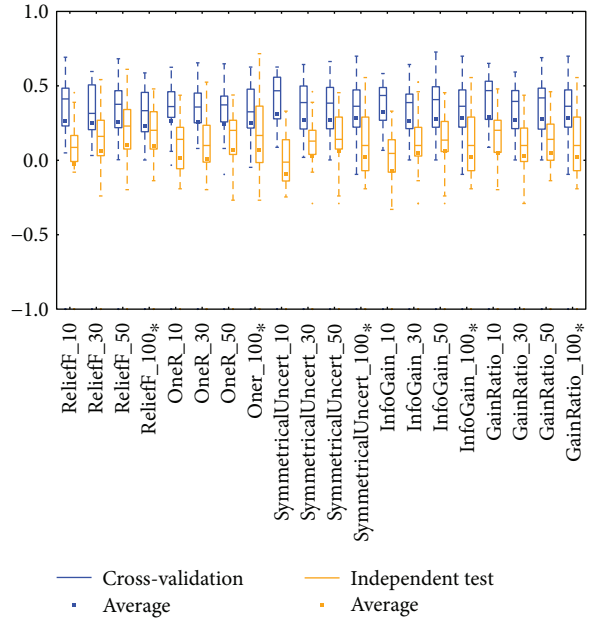



Figure 9: MCC of the KIRC-miRNA group by reduced datasets. *Note that the subdatasets from OV-miRNA have at most 83 micro-RNAs and thus the scale "100" of OV means 83.

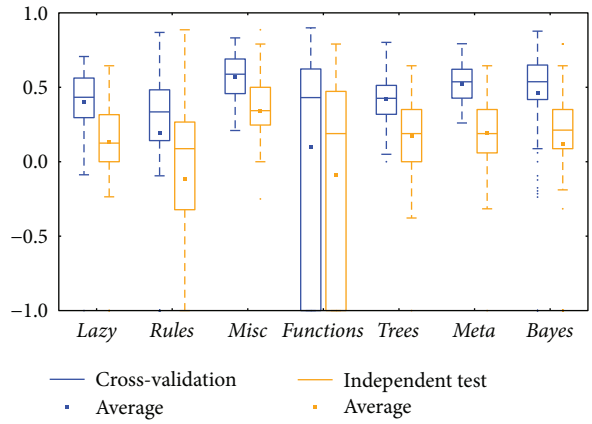



Figure 10: MCC of the OV-mRNA group by the *functions* class.

the OV-miRNA groups (Figure 13), the boxes contracted when attributes were added.

*3.2. Voting Results from All Methods.* The performance of all of the voted methods is listed in the column "Vote from all methods" of Tables 3, 4, and 5. There is no doubt that the overall voting performance would be lower than optimum, since not all of the voting methods are sufficiently good for the 6 datasets; nevertheless, most of the MCC achieved by voting are better than the average MCC. In addition, a part of the voting performance reached the upper bound (i.e., the maximum), for example, in the *bayes* class in the OV-miRNA group. In more detail, similar to the box plots, the voting performance based on the BRCA datasets was similar and near the upper whisker except for *misc*. The range was larger in the other datasets (e.g., KIRC-mRNA, KIRC-miRNA, OV-mRNA, and OV-miRNA). Different classes in turn, including *bayes*, *functions*, *meta*, *lazy*, *misc*, and *trees*, showed the top three best prediction performance. The *rules* class always ranked lower in the overall voting test, but the *bayes* class showed good voting performance even though its distribution, as indicated by the box plot, was not stable.

*3.3. Voting Results from Eliminated Methods.* The filtering rule was based on the distribution of the prediction results. Values that fell out of the range indicated by the whiskers in a box plot were considered to be outliers. Similarly, in our study, a method with an MCC below the lower whisker was considered as an outlier. There were 5 value selection methods, and each generated 4 subdatasets. If 6 datasets (3 cancers × 2 sequencing techniques) were considered, a modeling method was used: 5 variable selection methods × 4

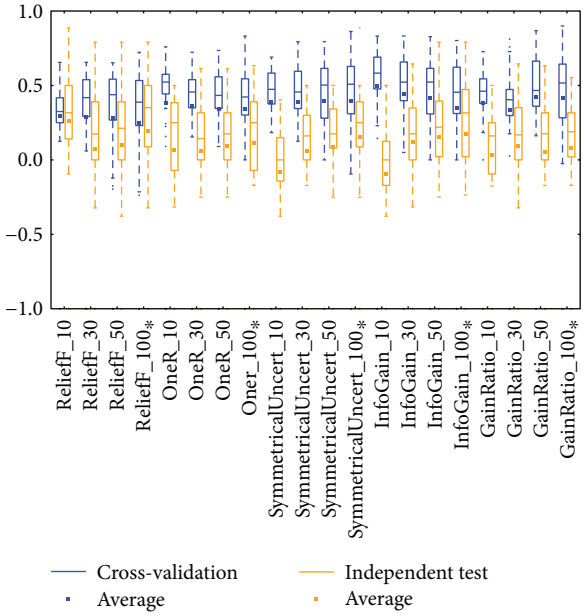



Figure 11: MCC of the OV-mRNA group by reduced datasets. *Note that the subdatasets from OV-miRNA have at most 83 micro-RNAs and thus the scale "100" of OV means 83.

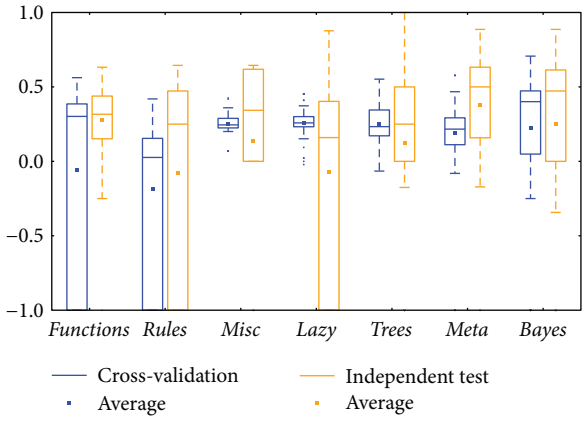



Figure 12: MCC of the OV-miRNA group by the *functions* class.

subdatasets × 6 datasets = 120 times. Therefore, if a method was designated as an outlier more than 6 times (5% of 120), it was not considered in this voting test.

The details of eliminated methods are listed in Table 6. The associated prediction performance was listed in the column "Vote after filtering" in Tables 3, 4, and 5. One or more methods were eliminated in all of the classes except the *meta*. It was evident that *ZeroR* had the most counts, and it was accordingly eliminated. Since only default parameters were used, *LibSVM* and *SPegasos*, which could be considered an optimized *SVM*, were designated as outliers with high counts. Similarly, the *BayesianLogisticRegression* method required parameter optimization and was thus eliminated by many counts. No methods in the *meta* class were eliminated, which indicated that methods based on resampling and an ensemble

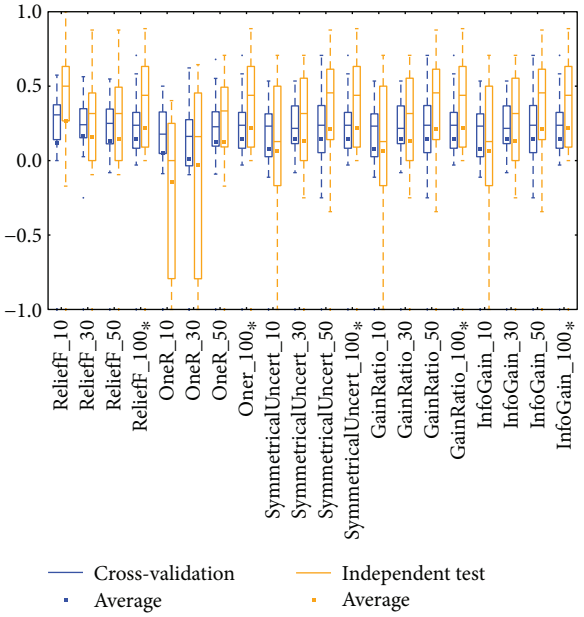



Figure 13: MCC of the OV-miRNA group by reduced datasets. In the 12 box plots, the line in the box is the median. The upper and lower boundaries of the box are Q1 and Q3. The boundaries of the dotted line are the whiskers. *The subdatasets from OV-miRNA have at most 83 microRNAs, and thus, the scale "100" of OV means 83. *Note that the subdatasets from OV-miRNA have at most 83 micro-RNAs and thus the scale "100" of OV means 83.

mechanism were stable and could address varied datasets even they were not always the best.

*3.4. Voting Results from the Combined Datasets.* As mentioned in the Preprocessing, the datasets for one cancer comprised the same samples. Therefore, voting from both mRNA and miRNA datasets is possible. The prediction performance is in the last row of Tables 3, 4, and 5. According to these three tables, the effects of the combination were different in the different classes. No class always benefited from the combination, but the MCCs determined by voting were not worse than the lowest MCC (Tables 4 and 5) and sometimes better than both (Table 3).

## 4. Discussion

In this section, the modeling results will be discussed. Therefore, the discussion is comprehensive, and the gene selection and the modeling results are discussed separately.

*4.1. Coverage of Selected Genes.* Five variable selection methods were used in this study, and different mRNA/miRNA datasets were separately generated. If the subdatasets generated are similar, combining multiple variable selection methods is worthless. Therefore, it is necessary to analyze the contents of the subdatasets and determine the similarity of the datasets. Moreover, the importance of the mRNA was evaluated by using a 3rd party database.

Table 3: Voting results of BRCA.

| Platforms | Class | Vote from all methods | Vote after filtering | MCC avg | MCC max |
|---|---|---|---|---|---|
| BRCA-mRNA | *bayes* | 0.7620 | 0.7620 | 0.5327 | 0.7766 |
| | *functions* | 0.7252 | 0.7766 | 0.3696 | 0.7938 |
| | *lazy* | 0.7620 | 0.7423 | 0.6797 | 0.7793 |
| | *meta* | 0.7274 | 0.7274 | 0.7295 | 0.8085 |
| | *misc* | 0.5821 | 0.6466 | 0.5449 | 0.7869 |
| | *rules* | 0.7274 | 0.7274 | 0.5069 | 0.7967 |
| | *trees* | 0.7274 | 0.7274 | 0.7207 | 0.8085 |
| | *overall* | 0.7274 | 0.7274 | 0.5985 | 0.8085 |
| BRCA-miRNA | *bayes* | 0.7237 | 0.6895 | 0.3544 | 0.7067 |
| | *functions* | 0.6733 | 0.6908 | 0.1080 | 0.7566 |
| | *lazy* | 0.6214 | 0.6214 | 0.5067 | 0.6736 |
| | *meta* | 0.6278 | 0.6278 | 0.6142 | 0.7269 |
| | *misc* | 0.3203 | 0.5773 | 0.2816 | 0.6383 |
| | *rules* | 0.5990 | 0.6278 | 0.3933 | 0.7234 |
| | *trees* | 0.6427 | 0.6602 | 0.6066 | 0.7566 |
| | *overall* | 0.6405 | 0.6908 | 0.4427 | 0.7566 |
| BRCA-mRNA and BRCA-miRNA | *bayes* | 0.7423 | 0.7252 | 0.4436 | 0.7766 |
| | *functions* | 0.6555 | 0.7915 | 0.2388 | 0.7938 |
| | *lazy* | 0.7595 | 0.7595 | 0.5932 | 0.7793 |
| | *meta* | 0.7595 | 0.7595 | 0.6719 | 0.8085 |
| | *misc* | 0.5624 | 0.6756 | 0.4133 | 0.7869 |
| | *rules* | 0.7080 | 0.7746 | 0.4501 | 0.7967 |
| | *trees* | 0.7595 | 0.7595 | 0.6636 | 0.8085 |
| | *overall* | 0.7407 | 0.7746 | 0.5206 | 0.8085 |

All of the measurements in the tables are MCCs, and the vote after filtering is the MCC based on the eliminated methods. The "avg" is the average of the MCCs.

The coverage fraction is the number of the genes which are used more than once of all genes, and accordingly, the formula is

$$\text{Coverage} = \frac{\text{NumOfSelectedGenes} - \text{NumOfIndependentGenes}}{\text{NumOfSelectedGenes}}. \quad (3)$$

According to Tables 7 and 8, the overall coverage was approximately 40% in the datasets from mRNA. However, the overall coverage of the microRNA datasets was much larger, due to dataset scale limitations. Especially in OV, the finally selected microRNAs were the same because there were only 83 microRNAs; therefore, all of them were selected and ranked in the top 100, and the OV-miRNA datasets must be carefully considered when the predictions are integrated. Conversely, the coverage of the two mRNA subdatasets was not large. According to the statistical results for the mRNA datasets, the frequency of shared mRNA increased as the coverage fraction increased. The shared mRNAs were usually ranked lower by the variable selection methods, which means the most commonly used mRNAs were not recognized as crucial genes that would be ranked higher. The reason might be an insufficient sample number to determine the relationships, since there are many mRNAs. Another probable reason is that the gene expression data could not be directly associated with a disease since cancer is a complex group of diseases, so not only one or a small number of gene are correlated, such as in coexpression [39].

*4.2. Reliability of the Selected mRNAs.* To determine if the shared mRNA is important, the Human Protein Atlas database [40] was used to evaluate its importance. The Human Protein Atlas contains a map from mRNA to tissue generated by an antibody-based approach. The gene reliability is recorded in the database, so that the mRNAs selected in this study could be evaluated by using the database. The evaluation had two steps. First, the fraction of Hits/Total in Table 9 was used to determine the number of selected mRNAs in the dataset. Next, the mRNAs' reliability was verified by using the associated record in the "reliability (IH)" table. An mRNA was considered reliable only if the record was designated as "approved" or "supported."

Table 9 shows that the fractions of hits in the database were approximately 60%, 55%, and 35% for BRCA, KIRC, and OV, respectively. However, the reliable hits ranged near 70% for the three datasets. The reliabilities based on more than 10 hits were always around 70% no matter which dataset was used. Few differences in reliability were found for the three datasets. The fraction of the reliable mRNAs from OV is relatively larger, and smaller for BRCA. Since the number of records in the Human Protein Atlas is still limited, a more

Table 4: Voting results of KIRC.

| Platforms | Class | Vote from all methods | Vote after filtering | MCC avg | MCC max |
|---|---|---|---|---|---|
| KIRC-mRNA | *bayes* | 0.4216 | 0.3672 | 0.2401 | 0.6590 |
| | *functions* | 0.6292 | 0.7162 | 0.0399 | 0.7785 |
| | *lazy* | 0.4682 | 0.6292 | 0.3801 | 0.8474 |
| | *meta* | 0.5261 | 0.5261 | 0.3780 | 0.7785 |
| | *misc* | 0.4176 | 0.4105 | 0.2991 | 0.6058 |
| | *rules* | 0.4385 | 0.6110 | 0.0723 | 0.6885 |
| | *trees* | 0.5421 | 0.5421 | 0.3649 | 0.7785 |
| | *overall* | 0.6110 | 0.5421 | 0.2535 | 0.8474 |
| KIRC-miRNA | *bayes* | 0.2368 | 0.1805 | −0.0433 | 0.5131 |
| | *functions* | 0.0889 | 0.0530 | −0.1698 | 0.7162 |
| | *lazy* | 0.4371 | 0.3410 | 0.1865 | 0.5249 |
| | *meta* | 0.1667 | 0.1667 | 0.1105 | 0.4542 |
| | *misc* | 0.4606 | 0.4795 | 0.3230 | 0.6590 |
| | *rules* | 0.0889 | 0.2689 | −0.0795 | 0.4606 |
| | *trees* | 0.0530 | 0.1667 | 0.1165 | 0.6110 |
| | *overall* | 0.1667 | 0.1667 | 0.0323 | 0.7162 |
| KIRC-mRNA and KIRC-miRNA | *bayes* | 0.4795 | 0.4795 | 0.0984 | 0.6590 |
| | *functions* | 0.2605 | 0.6885 | −0.0649 | 0.7785 |
| | *lazy* | 0.5514 | 0.4371 | 0.2833 | 0.8474 |
| | *meta* | 0.2300 | 0.2300 | 0.2442 | 0.7785 |
| | *misc* | 0.4105 | 0.5131 | 0.3110 | 0.6590 |
| | *rules* | 0.2605 | 0.5249 | −0.0036 | 0.6885 |
| | *trees* | 0.4371 | 0.5249 | 0.2407 | 0.7785 |
| | *overall* | 0.4371 | 0.5249 | 0.1429 | 0.8474 |

All of the measurements in the tables are MCCs, and the vote after filtering is the MCC based on the eliminated methods. The "avg" is the average of the MCCs.

reliable discussion and conclusion must await further analysis based on more samples and records.

The records in the Human Protein Atlas will be updated, and more mRNAs will become available. Therefore, the fraction of hits and the reliability will change accordingly and the shared mRNAs that are currently not recorded in this database might be worthy of study. In addition, the reliabilities of the genes were similar whether a gene was independent or shared, and this indicates that the selected genes were representative for a prognosis but might have a redundant function; therefore, the shared genes are not significantly different from the independent genes in reliability.

*4.3. Reliability of the Modules.* The results from cross-validation could be used as the reference to evaluate the reliability by comparing the results with independent test. As an empirical conclusion, in most cases, the results from cross-validation could be better than an independent test due to various reasons such as the overfitting and batch effects, but if the difference is not too large, the modules could be identified as reliable. Reflected in the figures, except a few classes such as the *functions* class and *lazy* class, the MCC from cross-validation had a relatively better predicting performance (e.g., higher median and average or narrower IQR) than independent test. According to the whiskers, most of the classes had small differences, but there was still a lot of the modules that had a large difference which could be reflected by the outliers. Thus, it is still risky to get the unreliable prediction if we only use the modules which have good performance in cross-validation for predicting. However, since most of the methods were reliable, combining the methods together becomes useful and necessary to reduce the risk.

*4.4. Distributions of the Predictions.* A balanced ratio of positive samples to negative samples is an important factor for prediction. The BCRA datasets had the largest sample numbers; therefore, the IQRs were the smallest, which indicates that the MCCs were concentrated toward the median. The boxes become wider for the KIRC, OV-mRNA, and OV-miRNA datasets. The sample number should be guaranteed before modeling if a stable prediction is to be obtained. However, sometimes, many reasons such as cost, privacy, and difficulty limit the sample number, so it is insufficient to confirm the prediction stability. Such predictions should be considered very carefully because overfitting may have occurred.

A basic way to avoid overfitting is to reduce the attribute number for modeling, and that is why 4 subdatasets (i.e., datasets with 10, 30, 50, and 100 samples) were used for modeling. As shown in the Modeling Results, especially in Figures 9 and 11, more attributes relatively improved the

Table 5: Voting results of OV.

| Platforms | Class | Vote from all methods | Vote after filtering | MCC avg | MCC max |
|---|---|---|---|---|---|
| OV-mRNA | *bayes* | 0.4725 | 0.4725 | 0.1217 | 0.7906 |
| | *functions* | 0.1250 | 0.4725 | −0.0890 | 0.7906 |
| | *lazy* | 0.3162 | 0.3162 | 0.1328 | 0.6447 |
| | *meta* | 0.1890 | 0.1890 | 0.1923 | 0.6447 |
| | *misc* | 0.6139 | 0.7500 | 0.3433 | 0.8864 |
| | *rules* | 0.1250 | 0.3162 | −0.1160 | 0.8864 |
| | *trees* | 0.1890 | 0.1890 | 0.1734 | 0.6447 |
| | *overall* | 0.3162 | 0.3162 | 0.0890 | 0.8864 |
| OV-miRNA | *bayes* | 1.0000 | 0.8771 | 0.2508 | 0.8864 |
| | *functions* | 0.5000 | 0.6139 | −0.0740 | 0.8771 |
| | *lazy* | 0.6447 | 0.6447 | 0.2756 | 0.6325 |
| | *meta* | 0.7559 | 0.7559 | 0.3786 | 0.8864 |
| | *misc* | 0.7559 | 0.6139 | 0.1384 | 0.6447 |
| | *rules* | 0.3430 | 0.7559 | −0.0783 | 0.6447 |
| | *trees* | 0.5000 | 0.7559 | 0.1258 | 1.0000 |
| | *overall* | 0.7559 | 0.7559 | 0.1432 | 1.0000 |
| OV-mRNA And OV-miRNA | *bayes* | 0.4725 | 0.4725 | 0.1863 | 0.8864 |
| | *functions* | −1.0000 | 0.7500 | −0.0815 | 0.8771 |
| | *lazy* | 0.7500 | 0.8771 | 0.2042 | 0.6447 |
| | *meta* | 0.3162 | 0.3162 | 0.2854 | 0.8864 |
| | *misc* | 0.6139 | 0.7500 | 0.2408 | 0.8864 |
| | *rules* | 0.3430 | 0.3162 | −0.0971 | 0.8864 |
| | *trees* | 0.3162 | 0.4725 | 0.1496 | 1.0000 |
| | *overall* | 0.3162 | 0.4725 | 0.1161 | 1.0000 |

All of the measurements in the tables are MCCs, and the vote after filtering is the MCC based on the eliminated methods. The "avg" is the average of the MCCs.

Table 6: Methods eliminated as outliers.

| Class | Names | Counts |
|---|---|---|
| *bayes* | BayesianLogisticRegression | 78 |
| | DMNBtext | 24 |
| *functions* | SPegasos | 113 |
| | VotedPerceptron | 19 |
| | LibSVM | 106 |
| *lazy* | KStar | 8 |
| *meta* | *none* | / |
| *misc* | HyperPipes | 12 |
| *rules* | ConjunctiveRule | 56 |
| | OneR | 9 |
| | ZeroR | 120 |
| *trees* | BFTree | 14 |
| | DecisionStump | 20 |

The counts are the number of methods whose MCC is lower than the lower whisker in the box plot.

overall prediction performance. However, the improvement was still limited by the sample number. All of the BRCA subdatasets had the relatively smallest IQR compared to the others that had the same attribute number. This limitation might apply not only to modeling but also to many other studies which must use the samples as a template to measure the correlations among samples and genes. For example, in a gene set enrichment analysis, genes should be eliminated by statistical methods such as a *t*-test, fold-change, or FDR. In the analysis, the genes are independent from each other when calculating the correlations and thus the validity of the identified genes is only affected by the sample number. If the sample number is too small (less than 10), the *t*-test result is not reliable.

Limited sample and attribute numbers make a prediction sensitive to the datasets. The various distributions in Figures 2, 4, 6, 8, 10, and 12 indicate that the methods in the *meta* class are relatively stable for prediction, which is reasonable because the methods in the *meta* class use other methods for ensemble learning, so they are not sensitive to different dataset distributions as other types of methods are. The methods in the *trees* class were similar but had a more varied performance than the *meta* class methods because the classification and regression tree algorithms can be as simple as REPTree or as complex as random forest. Therefore, the boxes of the *trees* class usually had a larger IQR than the *meta* class. However, the *misc* class performed best for two datasets, but only two methods were contained in this class. Based on the algorithm, the methods in the *misc* class were much different and thus could have a much more variable performance for different datasets. Except for the *rules*

Table 7: Coverage of the selected genes from mRNA.

| Disease | ShareNum | Subdata scale | | | |
|---|---|---|---|---|---|
| | | 10 | 30 | 50 | 100 |
| BRCA | 5 | 0 | 2 | 2 | 5 |
| | 4 | 4 | 5 | 10 | 25 |
| | 3 | 4 | 13 | 21 | 45 |
| | 2 | 2 | 16 | 26 | 42 |
| | 1 | 18 | 49 | 85 | 156 |
| | Total | 28 | 85 | 144 | 273 |
| | Coverage fraction | 35.7% | 42.4% | 40.1% | 42.9% |
| KIRC | 5 | 0 | 0 | 0 | 2 |
| | 4 | 0 | 3 | 3 | 8 |
| | 3 | 2 | 6 | 14 | 28 |
| | 2 | 6 | 22 | 34 | 71 |
| | 1 | 32 | 76 | 128 | 232 |
| | Total | 40 | 107 | 179 | 341 |
| | Coverage fraction | 20% | 29% | 28.5% | 32% |
| OV | 5 | 0 | 1 | 2 | 5 |
| | 4 | 1 | 2 | 7 | 16 |
| | 3 | 1 | 8 | 16 | 55 |
| | 2 | 9 | 27 | 34 | 43 |
| | 1 | 25 | 59 | 96 | 160 |
| | Total | 36 | 97 | 155 | 279 |
| | Coverage fraction | 30.6% | 39.2% | 38.1% | 42.7% |

The coverage fraction is the number of the genes which are used more than once in all of the genes, and accordingly, the formula is Coverage = NumOfSelectedGenes − NumOfIndependentGenes/NumOfSelectedGenes. The ShareNum is the number of a gene used in the subdatasets. For example, in Table 7, the value in the OV-miRNA group with ShareNum 3 and data scale 10 is 1; it means that there is one gene which is used by 3 subdatasets and each subdataset has 10 microRNAs as the attributes.

Table 8: Coverage of the datasets from miRNA.

| Disease | ShareNum | Subdata scale | | | |
|---|---|---|---|---|---|
| | | 10 | 30 | 50 | 100* |
| BRCA | 5 | 1 | 2 | 2 | 8 |
| | 4 | 3 | 8 | 17 | 47 |
| | 3 | 3 | 11 | 23 | 31 |
| | 2 | 3 | 9 | 12 | 26 |
| | 1 | 18 | 57 | 79 | 127 |
| | Total | 28 | 87 | 133 | 239 |
| | Coverage fraction | 35.7% | 34.5% | 40.6% | 46.9% |
| KIRC | 5 | 1 | 14 | 16 | 35 |
| | 4 | 3 | 6 | 13 | 53 |
| | 3 | 5 | 7 | 19 | 12 |
| | 2 | 2 | 7 | 11 | 9 |
| | 1 | 14 | 21 | 39 | 59 |
| | Total | 25 | 55 | 98 | 168 |
| | Coverage fraction | 44% | 61.8% | 60.2% | 64.9% |
| OV | 5 | 0 | 5 | 17 | 83 |
| | 4 | 3 | 15 | 27 | 0 |
| | 3 | 7 | 10 | 6 | 0 |
| | 2 | 3 | 5 | 12 | 0 |
| | 1 | 11 | 25 | 15 | 0 |
| | Total | 24 | 60 | 77 | 83 |
| | Coverage fraction | 54.2% | 58.3% | 80.5% | 100% |

The coverage fraction means the number of the genes which are used for more than 1 times in all of the genes, and accordingly, the formula is Coverage = NumOfSelectedGenes − NumOfIndependentGenes/NumOfSelectedGenes. The ShareNum is the number of a gene that is used for the subdatasets. For example, in Table 8, the value in the OV-miRNA group with ShareNum 3 and data scale 10 is 7; it means that there are 7 genes which are used by 3 subdatasets and each subdataset has 10 microRNAs as the attributes. *The subdatasets from the OV-miRNA group have at most 83 microRNAs, and thus, the scale "100" of OV-miRNA means 83.

class, the methods in the *bayes* class also had a relatively inferior performance, which might have been caused by the low sample number, because probability-based methods are sensitive on a modeling scale. The performance of methods in the *functions* class was more skewed; the boxes usually had a good upper whisker but a poor lower whisker. One reason for these might be that the algorithms in this class are also different from those in the *misc* class. There were 6 methods in the *functions* class, and thus the prediction performance varied widely. Another reason might be parameter optimization; less parameter optimization would affect all of the methods, but the methods in the *functions* class might be the most affected because they are much more sensitive to the parameters when only default parameters were used.

The high similarity of the gene data used would lead to a similar prediction performance, but when the similarity is lower than 50%, as reflected by the coverage in Table 7 or Table 8, the associated box plots were not significantly similar (as shown in Figures 3, 5, 7, 9, and 11).

According to the coverage of mRNA and microRNA used and shown in Tables 7 and 8, the coverage among the datasets was not large, but most of the medians from different subdatasets for the same cancer were similar, and this is reflected in the boxplots (Figures 3, 5, 7, 9, and 11), which indicates that some of the information contained in the genes was duplicated. In other words, the genes eliminated by the variable selection methods are representative for modeling, but not comprehensive. The duplicated genes could arise from similar genetics or pathology; for example, they could have the same genetic regulation pathway or simply be coexpressed, so that only one would be sufficient for modeling. Additionally, duplication could indicate that variable selection and machine learning methods are not sufficient to find out all of the disease-correlated genes. On the one hand, current machine learning methods can only determine some disease-associated genes, so further study might be necessary. On the other hand, the voting scheme provided in this study could be helpful for evaluating the relationship between cancer and genes.

As previously mentioned, the predictions using different datasets differed, meaning that we cannot determine which method is best for all datasets. The separate use of different

Table 9: Reliability of selected mRNAs.

| Names | Data scale | Reliable/Hits/ShareNum | | | |
|---|---|---|---|---|---|
| | | 10 | 30 | 50 | 100* |
| BRCA | 5 | 0 | 1/2/2 | 1/2/2 | 4/5/5 |
| | 4 | 4/4/4 | 4/5/5 | 8/9/10 | 13/17/25 |
| | 3 | 0/2/4 | 2/6/13 | 5/10/21 | 13/21/45 |
| | 2 | 0/0/2 | 5/6/16 | 8/12/26 | 18/29/42 |
| | 1 | 8/10/18 | 27/33/49 | 39/58/85 | 68/97/156 |
| | Total | 12/16/28 | 39/52/85 | 61/91/144 | 116/169/273 |
| | Hits/Total | 57.1% | 61.2% | 63.2% | 61.9% |
| | Reliable/Hits | 75% | 75% | 67% | 68.6% |
| KIRC | 5 | 0 | 0 | 0 | 0/0/2 |
| | 4 | 0 | 0/0/3 | 0/0/3 | 2/4/8 |
| | 3 | 2/2/2 | 2/2/6 | 4/6/14 | 12/17/28 |
| | 2 | 2/2/6 | 10/17/22 | 19/27/34 | 26/40/71 |
| | 1 | 15/18/32 | 31/42/76 | 45/63/128 | 98/129/232 |
| | Total | 19/22/40 | 43/61/107 | 68/96/179 | 138/190/341 |
| | Hits/Total | 55% | 57% | 53.6% | 55.7% |
| | Reliable/Hits | 86.4% | 70.5% | 70.8% | 72.6% |
| OV | 5 | 0 | 1/1/1 | 1/1/2 | 2/2/5 |
| | 4 | 1/1/1 | 0/1/2 | 2/3/7 | 2/4/16 |
| | 3 | 0/0/1 | 1/1/8 | 1/2/16 | 12/16/55 |
| | 2 | 0/2/9 | 5/8/27 | 6/10/34 | 14/18/43 |
| | 1 | 7/8/25 | 13/21/59 | 27/34/96 | 50/64/160 |
| | Total | 8/11/36 | 20/32/97 | 37/50/155 | 80/104/279 |
| | Hits/Total | 35% | 33% | 33.3% | 37.3% |
| | Reliable/Hits | 72.7 | 62.5% | 74% | 76.9% |

The table that records the hits and reliability in the "Human Protein Atlas" database. The "ShareNum" is the same in Table 8, and the "Hits" is the number of mRNAs recorded in the "Human Protein Atlas" database. The "Reliable" is the number of reliable hits. The reliability is measured by using the associated record in the "reliability (IH)" table. An mRNA is considered reliable only if the record is "approved" or "supported." "Hits/Total" and "Reliable/Hits" are calculated simply by using the row "Total" for division. For example, the last two elements in the last column are 37.3% and 76.9%. They are calculated by the element in the associated row in "Total," such as 80/104/279, where 37.3% = 104/279 and 76.9% = 80/104. *The subdatasets from the OV-miRNA group have at most 83 microRNAs, and thus, the scale "100" of OV means 83.

modeling methods will result in a loss of information, so using ensemble methods to integrate the modeling results is necessary.

*4.5. Effects of Voting.* A sufficiently large fraction of coverage in the data space is necessary for good voting performance, and the average accuracy must not be too low. Therefore, if only the box plots are considered, the expected performance of the *bayes* class is much lower. However, the result is anomalous in that the *bayes* class showed a good ensemble classification performance (it was ranked in top three) for 5 datasets. On the other hand, the *meta* and *trees* classes were not as good as the *bayes* class even though they had relatively similar distributions. One reason is many ensemble algorithms are contained in those two classes, so that the voting had already been accomplished, so the prediction results were concentrated near the median or the average. It is not difficult to see that in most cases, the voting results of the two classes were near the average. Another reason is that only default parameters were used in the entire test. Many ensemble methods must optimize the submethods for ensemble learning, and the resampling methods also must be optimized.

As a comparison, the voting performance from the eliminated methods was similar. However, the voting results were not always better than the original results. The *bayes* class was negatively affected by the filter. The MCCs based on the *bayes* class were not larger than prior to voting. The *rules* and *functions* classes were benefited by the filter, and the MCCs were improved for most of the datasets. The *trees* class was slightly benefited in the BRCA-miRNA group but was not globally affected as the *meta* class was. The other classes, including the overall voting, were affected positively or negatively by different datasets. The biased effects could indicate that the methods were sensitive to the datasets; the prediction performance of a method changed greatly when the dataset changed. On the other hand, the overall performance was not affected too greatly. One reason was the *meta* class methods, which showed a stable prediction ability, were not eliminated, and thus, the overall results remained stable.

Another reason might be that the performance effects were polarized. For example, the *bayes* class that showed the best comprehensive performance was negatively affected but the *rules* class was benefited. Therefore, the overall differences from the two voting mechanisms could be less.

The biased effects mean that a filter might not be that necessary if there is no prior knowledge. However, the voting by combining the mRNA and miRNA datasets could produce better performance if the sample size is sufficiently large, as shown in Table 3. Moreover, as shown in Tables 4 and 5, even the sample size is insufficient, the results will not become worse. Since the methods are not weighted, the results support that most of the methods will produce the same prediction, so combining the two datasets will be beneficial.

## 5. Conclusions

The purpose of this study was to discover a reliable way to predict unknown data to reduce the risk of error prediction when not enough samples were used for modeling. The distribution of the modeling performance indicated that the best methods were different for different datasets; therefore, the methods were integrated using a voting protocol. Finally, we proposed a better way to model different gene expression datasets. In conclusion, no prior knowledge exists; a comparison of the prediction results for three cancers indicates that the methods in the *bayes* class show a good ensemble performance, even though the individual methods are not as stable as those in the *meta* or *trees* classes. The *meta* and *trees* classes already contain many ensemble methods; therefore, their performance is stable but, again, not good for ensemble twice. Therefore, using the methods in the *bayes* class as a group and one of the algorithms in the *meta* class might be a practical approach for a dataset without sufficient prior knowledge. If prior knowledge exists for a cancer, the methods and datasets used can be more specific. For example, this study indicates that using miRNA as an attribute for modeling the OV data could yield a better result than using mRNA, if we knew that at first, some of the negative effects could be avoided. We hope that the scheme can facilitate related studies of genetic data modeling and elucidate important genes to enhance the reliability of the final model.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Acknowledgments

This project was supported by the National Natural Science Foundation of China under Grant no. 21575094, Sichuan Province Key Laboratory of Special Waste Water Treatment under Grant no. SWWT2015-3, Fundamental Research Funds for Central Non-profit Scientific Institution under Grant no. 12017206030202209, the National Key Research and Development Program of China under Grant no. 2016YFB0201305, Guangdong Provincial Department of Science and Technology under Grant no. 2016B090918122, the Science Technology and Innovation Committee of Shenzhen Municipality under Grant nos. JCYJ20160331190 123578 and GJHZ20170314154722613, and Youth Innovation Promotion Association of the Chinese Academy of Sciences, to Yanjie Wei.

## Supplementary Materials

As mentioned in Results, a file named "ModelingResults.xlsx" was provided as the supplementary material for recording the 5880 modeling results which were used for generating Figures 2–12. *(Supplementary Materials)*

## References

[1] A. Bashiri, M. Ghazisaeedi, R. Safdari, L. Shahmoradi, and H. Ehtesham, "Improving the prediction of survival in cancer patients by using machine learning techniques: experience of gene: a narrative review," *Iranian Journal of Public Health*, vol. 46, no. 2, pp. 165–172, 2017.

[2] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and Structural Biotechnology Journal*, vol. 13, pp. 8–17, 2015.

[3] S.-W. Chang, S. Abdul-Kareem, A. F. Merican, and R. B. Zain, "Oral cancer prognosis based on clinicopathologic and genomic markers using a hybrid of feature selection and machine learning methods," *BMC Bioinformatics*, vol. 14, no. 1, p. 170, 2013.

[4] Y. Ishibashi, N. Hanyu, K. Nakada et al., "Profiling gene expression ratios of paired cancerous and normal tissue predicts relapse of esophageal squamous cell carcinoma," *Cancer Research*, vol. 63, no. 16, pp. 5159–5164, 2003.

[5] X. Xu, Y. Zhang, L. Zou, M. Wang, and A. Li, "A gene signature for breast cancer prognosis using support vector machine," in *2012 5th International Conference on BioMedical Engineering and Informatics*, pp. 928–931, Chongqing, China, 2012.

[6] L. He, Y. Wang, Y. Yang, L. Huang, and Z. Wen, "Identifying the gene signatures from gene-pathway bipartite network guarantees the robust model performance on predicting the cancer prognosis," *BioMed Research International*, vol. 2014, Article ID 424509, 10 pages, 2014.

[7] L. Jiang, L. Huang, Q. Kuang et al., "Improving the prediction of chemotherapeutic sensitivity of tumors in breast cancer via optimizing the selection of candidate genes," *Computational Biology and Chemistry*, vol. 49, pp. 71–78, 2014.

[8] F. Xie, M. He, L. He et al., "Bipartite network analysis reveals metabolic gene expression profiles that are highly associated with the clinical outcomes of acute myeloid leukemia," *Computational Biology and Chemistry*, vol. 67, pp. 150–157, 2017.

[9] L. P. Petalidis, A. Oulas, M. Backlund et al., "Improved grading and survival prediction of human astrocytic brain tumors by artificial neural network analysis of gene expression microarray data," *Molecular Cancer Therapeutics*, vol. 7, no. 5, pp. 1013–1024, 2008.

[10] Y.-C. Chen, W.-C. Ke, and H.-W. Chiu, "Risk classification of cancer survival using ANN with gene expression data from multiple laboratories," *Computers in Biology and Medicine*, vol. 48, pp. 1–7, 2014.

[11] F. Sato, Y. Shimada, F. M. Selaru et al., "Prediction of survival in patients with esophageal carcinoma using artificial neural networks," *Cancer*, vol. 103, no. 8, pp. 1596–1605, 2005.

[12] O. Gevaert, F. D. Smet, D. Timmerman, Y. Moreau, and B. D. Moor, "Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks," *Bioinformatics*, vol. 22, no. 14, pp. e184–e190, 2006.

[13] A. Moslemi, H. Mahjub, M. Saidijam, J. Poorolajal, and A. R. Soltanian, "Bayesian survival analysis of high-dimensional microarray data for mantle cell lymphoma patients," *Asian Pacific Journal of Cancer Prevention*, vol. 17, no. 1, pp. 95–100, 2016.

[14] H. Wang, L. Huang, R. Jing et al., "Identifying oncogenes as features for clinical cancer prognosis by Bayesian nonparametric variable selection algorithm," *Chemometrics and Intelligent Laboratory Systems*, vol. 146, pp. 464–471, 2015.

[15] T. Ando, M. Suguro, T. Hanai, T. Kobayashi, H. Honda, and M. Seto, "Fuzzy neural network applied to gene expression profiling for predicting the prognosis of diffuse large B-cell lymphoma," *Japanese Journal of Cancer Research*, vol. 93, no. 11, pp. 1207–1212, 2002.

[16] L. Shi, G. Campbell, W. D. Jones et al., "The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models," *Nature Biotechnology*, vol. 28, no. 8, pp. 827–838, 2010.

[17] C. Park, J. Ahn, H. Kim, and S. Park, "Integrative gene network construction to analyze cancer recurrence using semi-supervised learning," *PloS One*, vol. 9, no. 1, article e86309, 2014.

[18] J. Hou, J. Aerts, B. den Hamer et al., "Gene expression-based classification of non-small cell lung carcinomas and survival prediction," *PloS One*, vol. 5, no. 4, article e10312, 2010.

[19] J. Xu, R. Jing, Y. Liu, Y. Dong, Z. Wen, and M. Li, "A new strategy for exploring the hierarchical structure of cancers by adaptively partitioning functional modules from gene expression network," *Scientific Reports*, vol. 6, no. 1, 2016.

[20] F. M. Lopes, R. M. Cesar Jr., and L. D. F. Costa, "Gene expression complex networks: synthesis, identification, and analysis," *Journal of Computational Biology*, vol. 18, no. 10, pp. 1353–1367, 2011.

[21] O. Rozenblatt-Rosen, R. C. Deo, M. Padi et al., "Interpreting cancer genomes using systematic host network perturbations by tumour virus proteins," *Nature*, vol. 487, no. 7408, pp. 491–495, 2012.

[22] T. Wang, J. Gu, J. Yuan, R. Tao, Y. Li, and S. Li, "Inferring pathway crosstalk networks using gene set co-expression signatures," *Molecular BioSystems*, vol. 9, no. 7, pp. 1822–1828, 2013.

[23] R. L. Grossman, A. P. Heath, V. Ferretti et al., "Toward a shared vision for cancer genomic data," *The New England Journal of Medicine*, vol. 375, no. 12, pp. 1109–1112, 2016.

[24] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[25] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1–27, 2011.

[26] S. Le Cessie and J. C. Van Houwelingen, "Ridge estimators in logistic regression," *Applied Statistics*, vol. 41, no. 1, pp. 191–201, 1992.

[27] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *Proceedings of the Thirteenth International Conference on Machine Learning*, vol. 96, pp. 148–156, Bari, Italy, 1996.

[28] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[29] G. H. John and P. Langley, "Estimating continuous distributions in Bayesian classifiers," in *UAI'95 Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pp. 338–345, Montreal, Quebec, Canada, 1995.

[30] A. McCallum and K. Nigam, "A comparison of event models for naive bayes text classification," in *Learning for Text Categorization: Papers from the 1998 AAAI Workshop*, vol. 752, pp. 41–48, Stanford, CA, USA, 1998.

[31] A. Genkin, D. D. Lewis, and D. Madigan, "Large-scale Bayesian logistic regression for text categorization," *Technometrics*, vol. 49, no. 3, pp. 291–304, 2007.

[32] R. Kohavi, "The power of decision tables," in *ECML '95 Proceedings of the 8th European Conference on Machine Learning*, pp. 174–189, Heraclion, Crete, Greece, 1995.

[33] E. Frank, M. Hall, and B. Pfahringer, "Locally weighted naive bayes," in *UAI'03 Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*, pp. 249–256, Acapulco, Mexico, 2002.

[34] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[35] G. Demiröz and H. Güvenir, "Classification by voting feature intervals," in *ECML '97 Proceedings of the 9th European Conference on Machine Learning*, pp. 85–92, Prague, Czech Republic, 1997.

[36] R. C. Holte, "Very simple classification rules perform well on most commonly used datasets," *Machine Learning*, vol. 11, no. 1, pp. 63–90, 1993.

[37] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *ML92 Proceedings of the Ninth International Workshop on Machine Learning*, pp. 249–256, San Francisco, CA, USA, 1992.

[38] C. E. Shannon, "A mathematical theory of communication, part I, part II," *Bell System Technical Journal*, vol. 27, no. 4, pp. 623–656, 1948.

[39] Z. Wen, Z. Wang, S. Wang et al., "Discovery of molecular mechanisms of traditional Chinese medicinal formula Si-Wu-Tang using gene expression microarray and connectivity map," *PloS One*, vol. 6, no. 3, article e18278, 2011.

[40] M. Uhlén, L. Fagerberg, B. M. Hallström et al., "Tissue-based map of the human proteome," *Science*, vol. 347, no. 6220, article 1260419, 2015.

Hindawi
International Journal of Genomics
Volume 2017, Article ID 3538568, 11 pages
https://doi.org/10.1155/2017/3538568

*Research Article*

# A New Network-Based Strategy for Predicting the Potential miRNA-mRNA Interactions in Tumorigenesis

**Jiwei Xue, Fanfan Xie, Junmei Xu, Yuan Liu, Yu Liang, Zhining Wen, and Menglong Li**

*College of Chemistry, Sichuan University, Chengdu 610064, China*

Correspondence should be addressed to Zhining Wen; w_zhining@163.com and Menglong Li; liml@scu.edu.cn

Received 29 April 2017; Accepted 10 July 2017; Published 2 August 2017

Academic Editor: Brian Wigdahl

Copyright © 2017 Jiwei Xue et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

MicroRNA (miRNA) plays an important role in the degradation and inhibition of mRNAs and is a kind of essential drug targets for cancer therapy. To facilitate the clinical cancer research, we proposed a network-based strategy to identify the cancer-related miRNAs and to predict their targeted genes based on the gene expression profiles. The strategy was validated by using the data sets of acute myeloid leukemia (AML), breast invasive carcinoma (BRCA), and kidney renal clear cell carcinoma (KIRC). The results showed that in the top 20 miRNAs ranked by their degrees, 90.0% (18/20), 70.0% (14/20), and 70.0% (14/20) miRNAs were found to be associated with the cancers for AML, BRCA, and KIRC, respectively. The KEGG pathways and GO terms enriched with the genes that were predicted as the targets of the cancer-related miRNAs were significantly associated with the biological processes of cancers. In addition, several genes, which were predicted to be regulated by more than three miRNAs, were identified to be the potential drug targets annotated by using the human protein atlas database. Our results demonstrated that the proposed strategy can be helpful for predicting the miRNA-mRNA interactions in tumorigenesis and identifying the cancer-related miRNAs as the potential drug targets.

## 1. Introduction

MicroRNAs (miRNAs) are a class of endogenous small noncoding RNA molecule with a length of ~22 nucleotides, which regulate gene expression posttranscriptionally [1]. miRNAs can combine with mRNAs to form the RNA-induced silencing complex (RISC) and degrade the mRNAs or inhibit the translation of the target genes [2]. The "seed sequence" with a length of 2 ~ 8 nt at the 5′ end of the miRNA plays an important role in target recognition by binding to the complementary sequences in the untranslated regions (3′-UTRs) of mRNAs [3]. A single miRNA may have the capability to target multiple mRNAs [4, 5] and participates in multiple signaling pathways and biological processes in mammals. It has been reported that miRNAs are involved in numerous cancer-relevant processes such as cell growth, proliferation, apoptosis, migration, and metabolism [6, 7]. The aberrant expression of miRNAs is related to different types of diseases and cancers, such as coronary artery disease [8], gastric cancer [9], lung cancer [10], and breast cancer [11].

Based on the increasing number of studies, miRNAs are being explored as the diagnostic and prognostic biomarkers and as the therapeutic targets for cancer treatment [12]. Previous studies revealed that miRNAs mainly acted as the oncogenic targets or tumor suppressors in the gene regulatory networks [13]. Therefore, two miRNA-based therapeutic strategies were proposed to restore or inhibit miRNA function through miRNA mimics and inhibitors (anti-miRs) [14]. As reported, numerous tumor-suppressive miRNAs and oncogenic miRNAs are promising drug candidates for the treatment of cancers and other diseases [15]. Although most of the miRNA-targeted drugs are still in the preclinical trials, antimiR-122, which is a LNA- (locked nucleic acid-) modified antisense inhibitor, has reached phase II trials for treating hepatitis [16] and the mimics of miR-34, which were encapsulated in lipid nanoparticles, have reached phase I clinical trials for the cancer treatment [17, 18]. Therefore, it

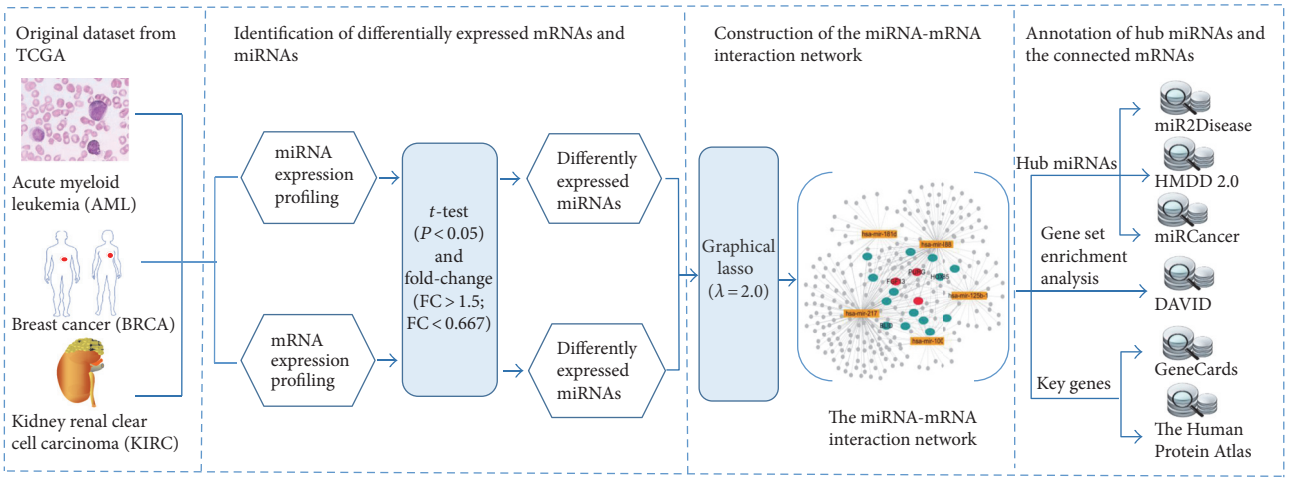



Figure 1: The overview of the study design.

Table 1: The annotation of the top 20 miRNAs in AML.

| Cancer type | | Number of genes | Disease |
|---|---|---|---|
| AML | hsa-mir-556 | 170 | — |
| | hsa-mir-217* | 163 | B-cell chronic lymphocytic leukemia, pancreatic neoplasms, nasopharyngeal carcinoma |
| | hsa-mir-636 | 159 | Myelodysplastic syndromes, multiple myeloma |
| | hsa-mir-320c-1 | 147 | Hepatocellular carcinoma, interstitial cystitis |
| | hsa-mir-639 | 145 | Lung cancer, gastric cancer, breast cancer |
| | hsa-mir-873 | 145 | Glioblastoma, endometriosis |
| | hsa-mir-573 | 138 | Pancreatic cancer, esophageal cancer, breast cancer |
| | hsa-mir-216b | 116 | Lung neoplasms, nasopharyngeal neoplasms, colorectal neoplasms |
| | hsa-mir-605 | 109 | Stomach neoplasms, ovarian cancer |
| | hsa-mir-188* | 103 | B-cell chronic lymphocytic leukemia, salivary gland neoplasms, rectal neoplasms |
| | hsa-mir-1468 | 89 | — |
| | hsa-mir-296 | 52 | Glioma, prostate cancer, urinary bladder neoplasms |
| | hsa-mir-488 | 49 | Melanoma, ovarian neoplasms, prostatic neoplasms |
| | hsa-mir-125b-1* | 40 | Acute myeloid leukemia, breast neoplasms, hepatocellular carcinoma |
| | hsa-mir-502 | 36 | Colonic neoplasms, ovarian neoplasms, hepatocellular carcinoma |
| | hsa-mir-551a | 32 | Stomach neoplasms, ovarian cancer |
| | hsa-mir-100* | 30 | Acute myeloid leukemia, precursor cell lymphoblastic leukemia-lymphoma, endometrial neoplasms |
| | hsa-mir-501 | 29 | Melanoma, atrophic muscular disorders |
| | hsa-mir-520a | 26 | Hodgkin's lymphoma, stomach neoplasms, colorectal neoplasms |
| | hsa-mir-181d* | 25 | Acute myeloid leukemia, acute promyelocytic leukemia, glioblastoma |

*The miRNA was directly associated with AML. ¯No description of the miRNA was found in the disease-related miRNA database.

is essential to identify the key miRNA candidates for the development of miRNA-based therapeutics of the cancers. In recent years, numerous databases, such as miRBase [19], miRanda [20], DIANA-TarBase [21], and HMDD v2.0 [22], have been developed to investigate the key role of miRNAs in the biological processes and reveal the miRNA-mRNA interaction mechanisms. However, considering the fact that a single miRNA will simultaneously target multiple genes, the miRNA-based therapeutics, which were designed to modulate miRNA expression levels, will affect hundreds of genes. It would be harmful for the patient to randomly regulate the hundreds of transcripts [23]. Thus, it is important to provide an exhaustive analysis of the key miRNAs and the miRNA-mRNA interactions before applying the miRNA-based therapeutics to the clinical trials.

In our study, we proposed a strategy by using the graphical lasso algorithm [24] to discover the key miRNAs and the miRNA-mRNA interaction in tumorigenesis based on the

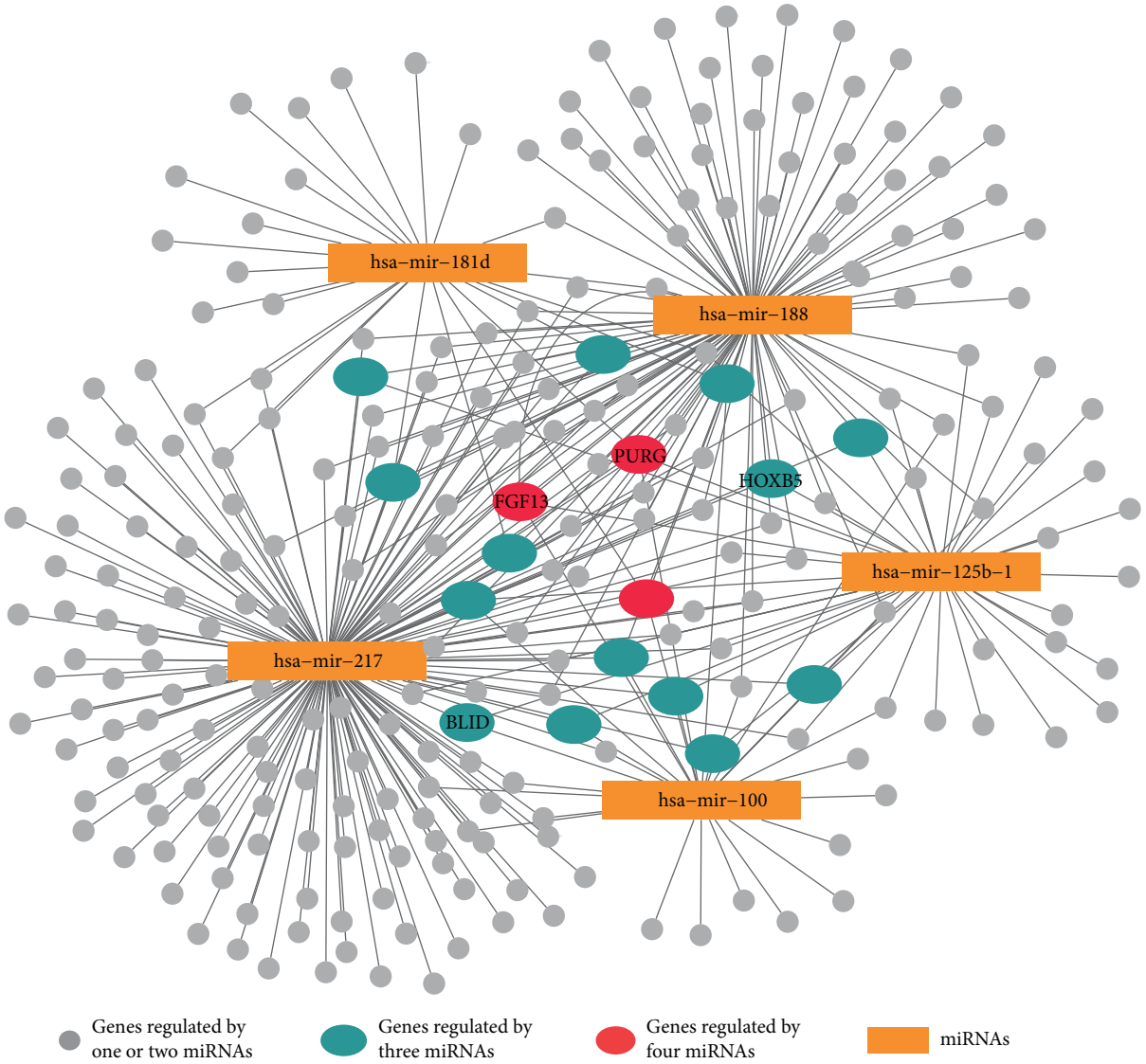



Figure 2: The miRNA-mRNA interaction subnetwork in AML. The five miRNAs in the network were reported to be associated with AML. In the figure, 14 mRNAs (cyan dots) and 3 mRNAs (red dots) were predicted to be connected with three and four miRNAs, respectively. The genes correlated with cancers were marked with their gene symbols.

expression levels of miRNAs and mRNAs. A bipartite network with the miRNAs as hubs was constructed to explore the interactions between the miRNAs and mRNAs, and the top 20 miRNAs ranked by their degrees in the network were verified by using three miRNA disease association databases, namely, miRCancer [25], miR2Disease [26], and HMDD v2.0 [22]. Moreover, the gene set enrichment analysis was conducted for the genes that were predicted as the targets in the network by using Database for the Annotation, Visualization, and Integrated Discovery (DAVID) v6.7 [27]. The proposed strategy was validated by using three cancer data sets. Our results showed that for both three data sets, most of the top 20 miRNAs as well as their targeted genes in the network were highly associated with cancers. In addition, the genes, which were predicted to be regulated by more than three cancer-related miRNAs in our study, had been reported as the potential drug targets in previous studies, indicating the satisfactory performance of our proposed strategy on predicting the cancer-related miRNAs and the interactions between miRNAs and their targeted genes.

## 2. Materials and Methods

*2.1. Datasets.* The miRNA expression data, the mRNA expression data, and the clinical data of three types of cancers, namely, acute myeloid leukemia (AML) [28], breast invasive carcinoma (BRCA) [29], and kidney renal clear cell carcinoma (KIRC) [30], were downloaded from the Cancer Genome Atlas (TCGA, https://cancergenome.nih.gov/) data portal. The miRNA-seq data in three data sets were generated by an Illumina Genome Analyzer in the Baylor College Human Genome Sequencing Center (BCGSC). The mRNA-seq data of AML (downloaded on November 7, 2016) were generated by an Illumina Genome Analyzer

Table 2: The annotation of the top 20 miRNAs in BRCA.

| Cancer type | | Number of genes | Disease |
|---|---|---|---|
| BRCA | hsa-mir-1269 | 381 | Lung cancer, colorectal cancer, hepatocellular carcinoma |
| | hsa-mir-934 | 368 | — |
| | hsa-mir-2115 | 325 | — |
| | hsa-mir-618 | 305 | — |
| | hsa-mir-1251 | 286 | — |
| | hsa-mir-9-3* | 282 | Breast neoplasms, stomach neoplasms, glioblastoma |
| | hsa-mir-105-2 | 268 | Biliary tract neoplasms, hepatocellular carcinoma |
| | hsa-mir-767 | 268 | Melanoma, rhinitis, allergy, perennial |
| | hsa-mir-449a* | 264 | Breast cancer, adenocarcinoma, colonic neoplasms, ovarian neoplasms |
| | hsa-mir-885 | 261 | Leukemia |
| | hsa-mir-105-1 | 253 | Biliary tract neoplasms, hepatocellular carcinoma |
| | hsa-mir-135a-1* | 251 | Breast neoplasms, colorectal neoplasms, non-small-cell lung carcinoma |
| | hsa-mir-3662 | 246 | Gastric cancer, head and neck cancer |
| | hsa-mir-138-1 | 242 | Oral squamous cell carcinoma, renal cell carcinoma, urinary bladder neoplasms |
| | hsa-mir-376a-2 | 234 | Adrenocortical carcinoma, glioblastoma, lung neoplasms |
| | hsa-mir-137* | 233 | Breast neoplasms, malignant melanoma, glioblastoma multiforme |
| | hsa-mir-3190 | 232 | — |
| | hsa-mir-138-2 | 231 | Papillary thyroid carcinoma, oral squamous cell carcinoma, pituitary adenoma |
| | hsa-mir-372 | 231 | Colorectal cancer, acute myeloid leukemia, stomach neoplasms |
| | hsa-mir-3926-2 | 231 | — |

*The miRNA was directly associated with BRCA. ¯No description of the miRNA was found in the disease-related miRNA database.

in the Baylor College Human Genome Sequencing Center (BCGSC). The mRNA-seq data of the BRCA (downloaded on December 15, 2014) and KIRC (downloaded on November 6, 2016) were produced by an Illumina HiSeq 2000 sequencer of the University of North Carolina (UNC). For the three data sets, the read counts for each miRNA and mRNA (data in level 3) were considered the expression level of the miRNA and the mRNA, respectively. In total, we collected 149, 829, and 253 samples for the data sets of AML, BRCA, and KIRC, respectively.

*2.2. Study Design.* In our study, the graphical lasso algorithm was proposed to construct the miRNA-mRNA interaction network. Figure 1 showed the overview of our study design. Three cancer data sets, namely, AML, BRCA, and KIRC, were downloaded from the TCGA database, and the differentially expressed miRNA and mRNAs were separately identified for each of the data sets by using the fold change ranking combined with a nonstringent $P$ value cutoff. Based on the expression profiles of the differentially expressed miRNAs and mRNAs, the interaction network was constructed by the graphical lasso algorithm, including the connections among the miRNAs and the mRNAs, as well as the connections between miRNAs and mRNAs. The miRNAs and their connected mRNAs in the network were extracted and regrouped into subnetworks, representing the interactions between miRNAs and mRNAs.

To validate whether the cancer-related miRNAs and their key targeted genes can be well characterized by our miRNA-mRNA interaction network or not, we annotated the top 20 miRNAs, which were ranked by their degrees (the number of connections), by using three disease-related miRNA databases, namely, miRCancer [25], miR2Disease [26], and HMDD v2.0 [22], for each of the data sets. Meanwhile, the gene set enrichment analysis was conducted with the targeted genes of the cancer-specific miRNAs by using DAVID v6.7. We checked whether or not the significantly enriched Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways and Gene Ontology (GO) terms were associated with cancers. In addition, we mainly discussed the functions of those genes that were predicted as the targets of more than three miRNAs.

*2.3. Identification of Differentially Expressed mRNAs and miRNAs.* To identify the differentially expressed mRNAs and miRNAs, we firstly divided the samples into two groups for each of the cancer types according to the clinical endpoints. For AML data set, the patients were subdivided into high-risk and low-risk groups according to their survival time. The patients with the survival days longer than one year were assigned to the low-risk group, and the patients with the survival days less than or equal to one year were assigned to the high-risk group. For BRCA data set, the patients were divided into the estrogen receptor- (ER-) positive group and the ER-negative group according to their estrogen receptor status [29]. As to the KIRC data set, the patients in the pathological stages I and II were assigned into the low-risk group and the patients in stages III and IV were assigned into the high-risk group. Then, for all the data sets, Student's $t$-test $P$ value was calculated for each of the miRNAs and mRNAs

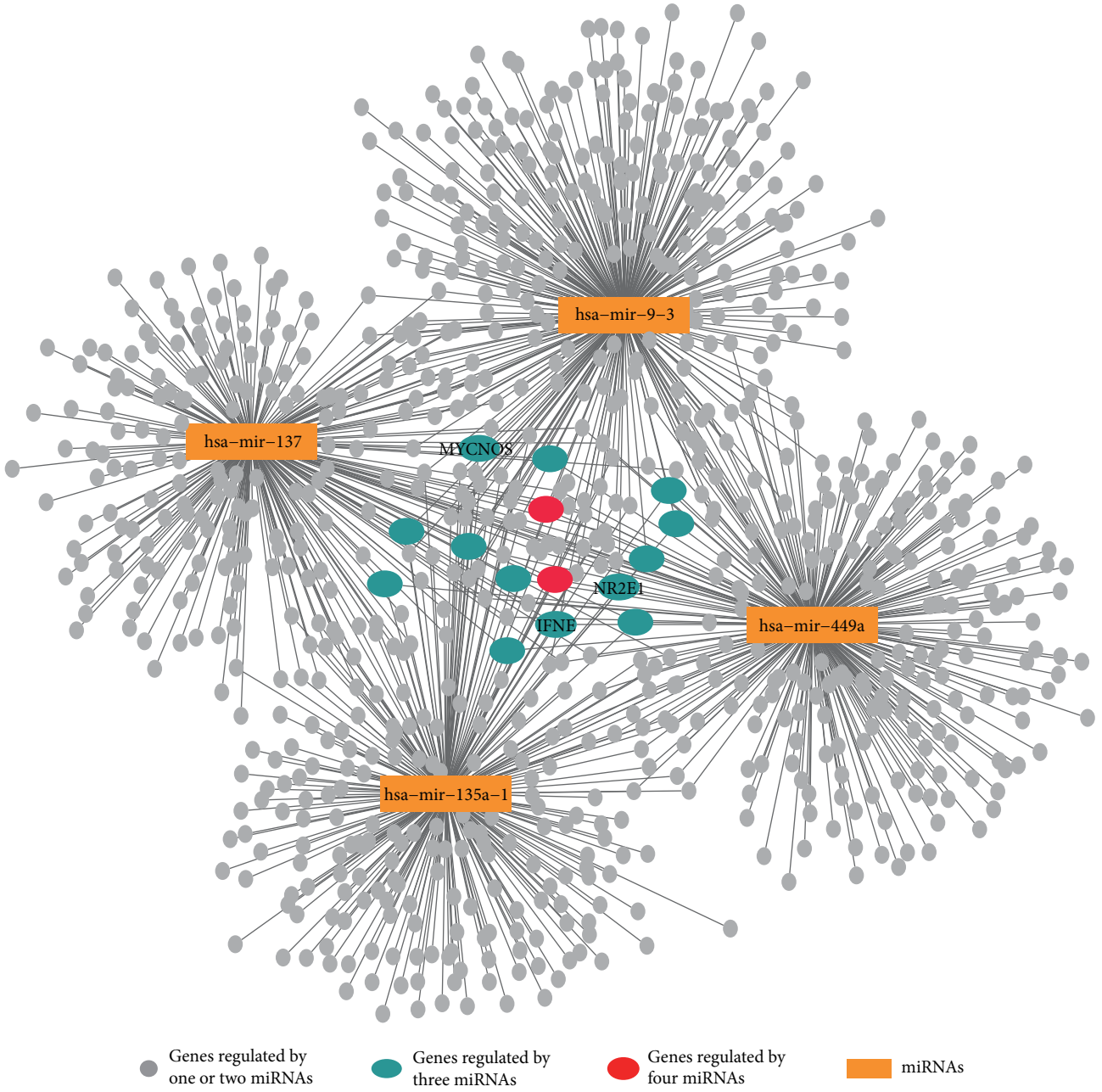



Figure 3: The miRNA-mRNA interaction subnetwork in BRCA. The four miRNAs in the network were reported to be associated with AML. In the figure, 13 mRNAs (cyan dots) and 2 mRNAs (red dots) were predicted to be connected with three and four miRNAs, respectively. The genes correlated with cancers were marked with their gene symbols.

by comparing the expression profiles of the miRNAs and mRNAs between the patient groups. We kept the miRNAs and mRNAs with $P < 0.05$ and calculated the fold changes of them between the compared patient groups, respectively. Finally, the miRNAs and the mRNAs with fold change greater than 1.5 (FC > 1.5) or less than 0.667 (FC < 0.667) were considered the differentially expressed miRNAs and mRNAs, respectively.

*2.4. Construction of the miRNA-mRNA Interaction Network.* As reported, Gaussian graphical models (GGMs) have been widely used to identify the dependent relationship among the variables and to be applied on the biological network inference [31, 32]. In GGMs, the conditional dependence of the two nodes was estimated by an inverse covariance matrix. A nonzero number in the inverse covariance matrix indicates a connection between two nodes [33]. The network inference actually is the estimation of the inverse covariance matrix, and numerous algorithms have been proposed to solve this problem [34]. Notably, based on the GGMs, a more reasonable approach named graphical lasso was proposed to directly estimate a sparse inverse covariance matrix by using the L1 (lasso) penalty [24, 35].

Table 3: The annotation of the top 20 miRNAs in KIRC.

| miRNA | Number of genes | Disease |
|---|---|---|
| hsa-mir-1291* | 344 | Renal cell carcinoma, ovarian cancer, kidney cancer |
| hsa-mir-558 | 243 | Pancreatic cancer, gastric cancer |
| hsa-mir-3924 | 237 | — |
| hsa-mir-376a-1 | 233 | Salivary gland neoplasms, lung neoplasms, adrenocortical carcinoma |
| hsa-mir-653 | 229 | — |
| hsa-mir-485 | 227 | Ependymoma, non-small-cell lung carcinoma, leukemia |
| hsa-mir-200b* | 216 | Renal cell carcinoma, diabetic nephropathies, pancreatic neoplasms |
| hsa-mir-134* | 215 | Renal cell carcinoma, lupus nephritis, glioblastoma |
| hsa-mir-1246 | 214 | Colorectal neoplasms, esophageal neoplasms |
| hsa-mir-346 | 212 | Lupus nephritis, hepatocellular carcinoma |
| hsa-mir-2110 | 210 | Hepatocellular carcinoma, colorectal neoplasms |
| hsa-mir-365-2 | 210 | — |
| hsa-mir-153-1 | 201 | Endometrial neoplasms, glioblastoma, rectal neoplasms |
| ‘hsa-mir-374c | 191 | — |
| hsa-mir-376b | 190 | Adrenocortical carcinoma, uterine leiomyoma, epithelial ovarian cancer |
| hsa-mir-218-2* | 184 | Renal cell carcinoma, lung cancer, urinary bladder neoplasms |
| hsa-mir-300 | 181 | Urinary bladder neoplasms, ovarian neoplasms, heart failure |
| hsa-mir-1303 | 179 | Colorectal neoplasms, hepatocellular carcinoma |
| hsa-mir-676 | 174 | — |
| hsa-mir-1237 | 156 | — |

*The miRNA was directly associated with KIRC. ¯No description of the miRNA was found in the disease-related miRNA database.

We assume a designed $n \times m$ matrix where $n$ indicates the number of samples and $m$ is the number of genes or miRNAs. Let $\boldsymbol{\theta} = \Sigma^{-1}$ and let $\mathbf{S}$ be the empirical covariance matrix; the problem of estimating $\boldsymbol{\theta}$ is converted to maximize the penalized log-likelihood:

$$\log\det\boldsymbol{\theta} - \mathrm{tr}(\mathrm{S}\boldsymbol{\theta}) - \rho||\boldsymbol{\theta}||_1, \quad (1)$$

where tr indicates the trace. $||\boldsymbol{\theta}||_1$ is the L1 norm of the matrix, which is the maximum value of the sum of the absolute values of the elements in each of the columns in $\boldsymbol{\theta}$, and $\rho$ is a nonnegative tuning parameter, which controls the sparseness of the network.

In fact, the graphical lasso gets a $\boldsymbol{\theta}_{m\times m}$ matrix to construct the network by using an $n \times m$ matrix as an input. We have two matrices $\mathbf{X}_{n\times j}$ ($j$ miRNA expression profiles of $n$ samples) and $\mathbf{Y}_{n\times k}$ ($k$ mRNA expression profiles of $n$ samples). Therefore, we integrated these two matrices into the matrix $\mathbf{Z}_{n\times(j+k)}$, which were used to construct an interaction network including the connections among the miRNAs and the mRNAs, as well as the connections between the miRNAs and mRNAs. In our study, only the differentially expressed miRNAs and the mRNAs were used to construct the interaction network and the penalty parameter $\rho$ was set to 2.0 for all the data sets. We mainly concentrated on the interactions between the miRNAs and the mRNA in the network.

## 3. Results

*3.1. Most of the Top 20 miRNAs Were Highly Associated with Cancers.* For AML data set, 34 differentially expressed miRNAs and 798 differentially expressed mRNAs were identified from 706 miRNAs and 20,319 mRNAs, respectively. Considering the miRNAs as the hubs of the miRNA-mRNA interaction network, we selected the top 20 miRNAs ranked by their degrees and listed them in Table 1. It can be seen from the table that 90% (18/20) miRNAs were associated with the cancers after being annotated by the three disease-related miRNA databases. Among the cancer-related miRNAs, five miRNAs, namely, hsa-mir-217, hsa-mir-188, hsa-mir-125b-1, hsa-mir-100, and hsa-mir-181d, were reported to be associated with the acute myeloid leukemia. Figure 2 showed the subnetworks including these five miRNAs as hubs and their targeted mRNAs.

For the data sets of BRCA and KIRC, we identified 266 and 54 differentially expressed miRNAs from 1043 and 1046 miRNAs, respectively, and identified 6021 and 1647 differentially expressed mRNAs from 20,502 and 20,503 mRNAs, respectively. The top 20 miRNAs ranked by their degrees in the miRNA-mRNA interaction network of the BRCA data set were listed in Table 2. It can be seen that 70% (14/20) miRNAs were annotated to be associated with cancers and four out of them, namely, hsa-mir-9-3, hsa-mir-449a, hsa-mir-135a-1, and hsa-mir-137, were breast cancer-specific miRNAs. Table 3 showed the top 20 miRNAs that were obtained from the interaction network of KIRC. 14 out of 20 (70%) miRNAs were reported to be associated with cancers, and four out of them, namely, hsa-mir-1291, hsa-mir-200b, hsa-mir-134, and hsa-mir-218-2 were directly associated with the renal cell carcinoma. The subnetworks

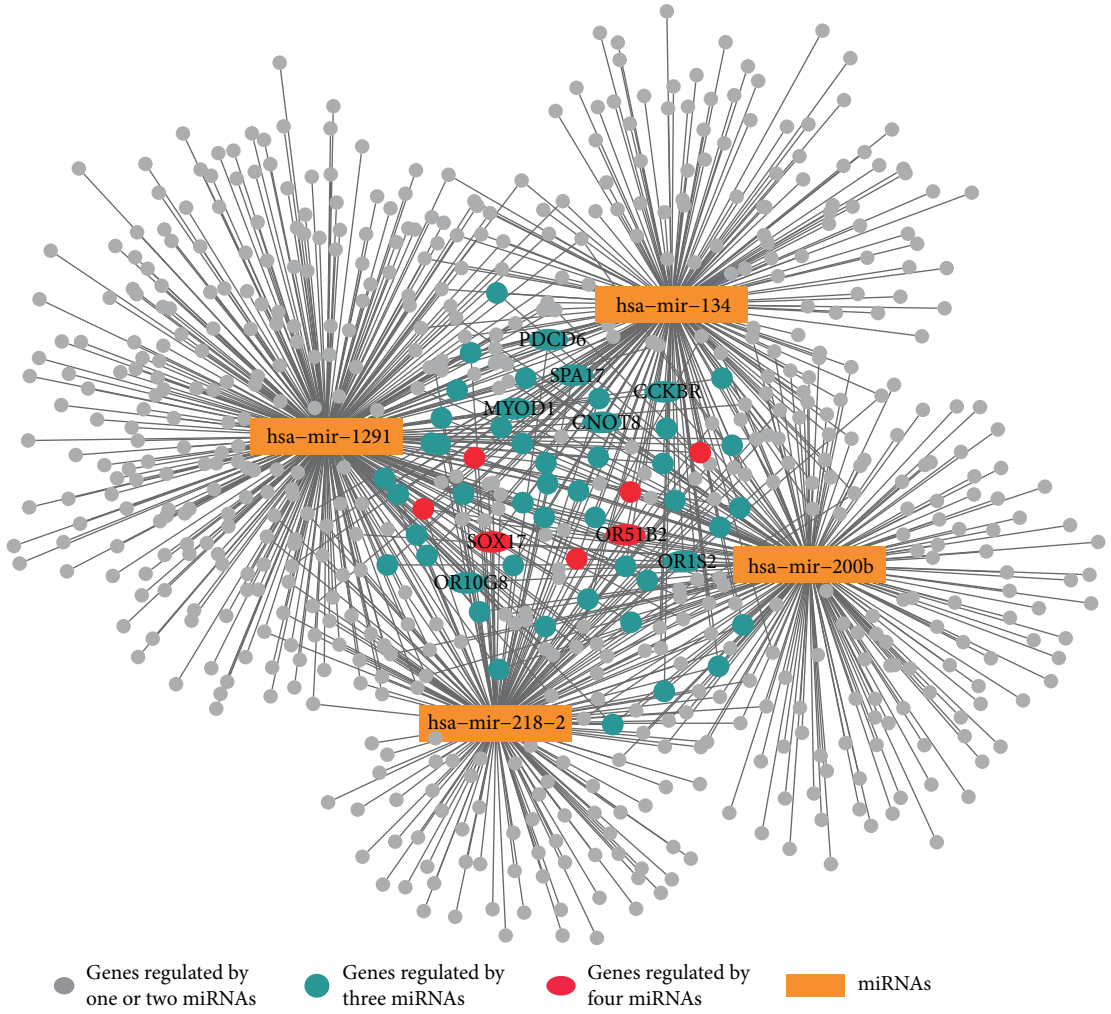



Figure 4: The miRNA-mRNA interaction subnetwork in KIRC. The four miRNAs in the network were reported to be associated with AML. In the figure, 49 mRNAs (cyan dots) and 7 mRNAs (red dots) were predicted to be connected with three and four miRNAs, respectively. The genes correlated with cancers were marked with their gene symbols.

Table 4: The top 5 KEGG pathways enriched with the genes connected with the cancer-specific miRNAs.

| Cancer type | KEGG pathways | $P$ value |
|---|---|---|
| AML | hsa00980: metabolism of xenobiotics by cytochrome | 0.0241 |
| | hsa00982: drug metabolism | 0.0263 |
| | hsa04740: olfactory transduction** | 0.0407 |
| BRCA | hsa04080: neuroactive ligand-receptor interaction* | $P < 0.0001$ |
| | hsa00140: steroid hormone biosynthesis ** | 0.0120 |
| | hsa03320: PPAR signaling pathway** | 0.0176 |
| | hsa04610: complement and coagulation cascades* | 0.0176 |
| | hsa00150: androgen and estrogen metabolism** | 0.0246 |
| KIRC | hsa05322: systemic lupus erythematosus | 0.0003 |
| | hsa04060: cytokine-cytokine receptor interaction** | 0.0021 |
| | hsa04740: olfactory transduction* | 0.0122 |
| | hsa05034: alcoholism | 0.0219 |
| | hsa00350: tyrosine metabolism | 0.0224 |

**The pathway was directly associated with the corresponding cancer type. *The pathway was associated with other cancers.

Table 5: The top 5 GO terms enriched with the genes connected with the cancer-specific miRNAs.

| Cancer type | Category | Term | *P* value |
|---|---|---|---|
| AML | GOTERM_BP_4 | GO:0009887 ~ organ morphogenesis | $P < 0.0001$ |
| | | GO:0048705 ~ skeletal system morphogenesis | $P < 0.0001$ |
| | | GO:0001501 ~ skeletal system development | $P < 0.0001$ |
| | | GO:0003002 ~ regionalization** | $P < 0.0001$ |
| | | GO:0048704 ~ embryonic skeletal system morphogenesis** | $P < 0.0001$ |
| | GOTERM_MF_4 | GO:0043565 ~ sequence-specific DNA binding** | 0.0055 |
| | | GO:0003700 ~ transcription factor activity | 0.0075 |
| | | GO:0008236 ~ serine-type peptidase activity | 0.0286 |
| | | GO:0004888 ~ transmembrane receptor activity* | 0.0301 |
| BRCA | GOTERM_BP_4 | GO:0019226 ~ transmission of nerve impulse | $P < 0.0001$ |
| | | GO:0007268 ~ synaptic transmission** | $P < 0.0001$ |
| | | GO:0007417 ~ central nervous system development** | $P < 0.0001$ |
| | | GO:0044057 ~ regulation of system process* | $P < 0.0001$ |
| | | GO:0009888 ~ tissue development* | $P < 0.0001$ |
| | GOTERM_MF_4 | GO:0030594 ~ neurotransmitter receptor activity** | $P < 0.0001$ |
| | | GO:0015267 ~ channel activity* | $P < 0.0001$ |
| | | GO:0015075 ~ ion transmembrane transporter activity* | $P < 0.0001$ |
| | | GO:0008188 ~ neuropeptide receptor activity** | $P < 0.0001$ |
| | | GO:0005179 ~ hormone activity* | $P < 0.0001$ |
| KIRC | GOTERM_BP_4 | GO:0006954 ~ inflammatory response | $P < 0.0001$ |
| | | GO:0007186 ~ G protein-coupled receptor signaling pathway** | $P < 0.0001$ |
| | | GO:0050707 ~ regulation of cytokine secretion* | $P < 0.0001$ |
| | | GO:0050663 ~ cytokine secretion* | 0.0002 |
| | | GO:0050715 ~ positive regulation of cytokine secretion | 0.0003 |
| | GOTERM_MF_4 | GO:0005125 ~ cytokine activity* | 0.0002 |
| | | GO:0004930 ~ G protein-coupled receptor activity** | 0.0002 |
| | | GO:0001664 ~ G protein-coupled receptor binding** | 0.0004 |
| | | GO:0005126 ~ cytokine receptor binding* | 0.0007 |
| | | GO:0004984 ~ olfactory receptor activity* | 0.0013 |

**The Go term was directly associated with the corresponding cancer type. *The Go term was associated with other cancers.

of the specific cancer-related miRNAs and their targeted mRNAs for the data sets of BRCA and KIRC were shown in Figures 3 and 4, respectively.

*3.2. The mRNAs Targeted by the Cancer-Specific miRNAs Were Significantly Associated with the Biological Process of Cancers.* The gene set enrichment analysis was conducted to investigate the gene functions by using the mRNAs, which were predicted as the targets of the cancer-specific miRNAs. For the data sets of AML, BRCA, and KIRC, 255, 853, and 670 targeted mRNAs were used for the gene set enrichment analysis, respectively. The top 5 significantly enriched KEGG pathways were listed in Table 4. For the data sets of AML, BRCA, and KIRC, there were one, five, and two signaling pathways, respectively, which were reported to be associated with cancers. Likewise, the top 5 significantly enriched GO terms related to the biological process and the molecular functions were listed in Table 5. There were four, nine, and eight GO terms for the data sets of AML, BRCA, and KIRC, respectively, which were associated with the tumorigenesis of the cancers.

When focusing on the mRNAs that were predicted to be the targets of multiple miRNAs, we found 14, 13, and 49 mRNAs targeted by three miRNAs in the miRNA-mRNA interaction networks of AML, BRCA, and KIRC, respectively. Moreover, three, two, and seven mRNAs were predicted to be targeted by four miRNAs in the networks of AML, BRCA, and KIRC, respectively. Figures 2, 3, and 4 showed the mRNAs targeted by three miRNAs (cyan dots) and four miRNAs (red dots). We also annotated these genes by using the GeneCards database v4.4.2 (http://www.genecards.org/) and found four, three, and nine genes from the networks of AML, BRCA, and KIRC, respectively, which were reported to be associated with cancers. The HUGO gene symbols of the cancer-related genes were marked in Figures 2, 3, and 4.

Table 6: The annotation of the key genes connected with more than three cancer-specific miRNAs in the miRNA-mRNA interaction networks.

| Cancer type | Gene | Gene description | Protein class |
|---|---|---|---|
| AML | ASPG | Asparaginase | Enzymes, predicted intracellular proteins |
| BRCA | AQP2 | Aquaporin 2 (collecting duct) | Disease-related genes, potential drug targets, predicted membrane proteins, transporters |
| KIRC | CNOT8 | CCR4-NOT transcription complex subunit 8 | Enzymes, plasma proteins, predicted intracellular proteins |
| | CTPS1 | CTP synthase 1 | Disease-related genes, enzymes, potential drug targets, predicted intracellular proteins |
| | IFNAR2 | Interferon (alpha, beta, and omega) receptor 2 | Cancer-related genes, FDA-approved drug targets, predicted intracellular proteins, predicted membrane proteins |
| | MOCS2 | Molybdenum cofactor synthesis 2 | Disease-related genes, enzymes, potential drug targets, predicted intracellular proteins |
| | PRSS37 | Protease, serine 37 | Enzymes, predicted secreted proteins |
| | VCP | Valosin-containing protein | Disease-related genes, enzymes, plasma proteins, potential drug targets, predicted intracellular proteins, transporters |

## 4. Discussion

In this study, we proposed a new strategy to construct the miRNA-mRNA interaction network based on the expression profiles of miRNAs and mRNAs. The connections between miRNAs and mRNAs were created by the graphical lasso algorithm. We applied the strategy to the three cancer data sets and successfully identified a number of cancer-related miRNAs and their targeted mRNAs.

For the AML data set, 90% miRNAs in the top 20 miRNAs were found to be associated with cancers (Table 1). Among these miRNAs, hsa-mir-100 was considered a potential tumor-related miRNA, which has been reported to regulate cell differentiation by targeting *RBSP3* in acute myeloid leukemia [36]. The pediatric AML patients with the upregulation of miR-100 may have poor relapse-free and overall survival [37]. Moreover, the downregulation of miR-181 family members including miR-181a, miR-181b, miR-181c, and miR-181d was associated with poor prognosis in cytogenetically normal acute myeloid leukemia [38]. For the BRCA data set, 70% miRNAs in Table 2 were associated with cancers and four of them were specifically associated with the breast cancer. has-miR-9 acted as a tumor suppressor, which can inhibit the proliferation of breast cancer cells [39]. miR-137 is a potential tumor suppressor miRNA, which negatively regulates the gene *ERRα* (estrogen-related receptor alpha) by targeting the two functional sites in the 3′-UTR of *ERRα* [40]. As to the KIRC data set, 70% miRNAs in Table 3 were associated with cancers and four of them had been reported to be associated with the development of the renal cell carcinoma. hsa-mir-134 had been reported as a tumor suppressor and can obstruct the tumor growth and metastasis by inhibiting epithelial-mesenchymal transition (EMT) in renal cell carcinoma cells [41]. miR-218 can mediate the focal adhesion pathway and inhibit the cell migration and invasion in renal cell carcinoma [42].

We also inspected the gene functions of the mRNAs, which were predicted to be the targets of the cancer-related miRNAs. The results of gene set enrichment analysis showed that the majority of the KEGG pathways (Table 4) and GO terms (Table 5) were significantly associated with the cancers. In the interaction subnetworks (Figures 2, 3, and 4), several mRNAs targeted by multiple cancer-specific miRNAs were found to have key roles in cancers. For example, the gene *SOX17*, which was predicted to be regulated by four miRNAs in the subnetwork of KIRC, was considered an important tumor suppressor with aberrant methylation for the cancers [43, 44]. In addition, the genes targeted by more than three miRNAs in the subnetworks were mapped to the Human Protein Atlas database v16.1 (http://www.proteinatlas.org), and 8 genes were annotated as the potential drug targets (Table 6).

Note that compared to the conventional drug therapies, the miRNA-targeting drugs have been regarded as a high-value therapy because miRNA may modulate multiple biological processes and pathways. However, there are a lot of challenges for utilizing miRNAs as potential therapeutic targets [45]. The miRNAs regulate tens of thousands of genes which could contribute to both efficacy and unexpected side effects. Therefore, the downstream analysis of genes and pathways regulated by miRNAs should be further elucidated and explored. Due to the complex regulatory mechanisms of the miRNAs, it is still challenging to successfully translate the miRNA-based therapy to the clinics [46]. It is a crucial step in miRNA drug discovery [47] to identify the specific miRNAs as drug targets and clarify the mechanisms of the actions for the key miRNAs. The network-based approach proposed in our study can identify the key miRNAs as well as their targeted mRNAs, which were also significantly associated with the biological process of cancers. It would be helpful for providing the complementary support to the miRNA-targeting drug discovery. In addition, the potential mRNA target could be enriched by integrating the protein structure information and medicinal chemistry. Furthermore, the accumulative information about the side effect and off target relationship from available public resources, such as PharmGKB [48] and the Comparative Toxicogenomics Database (CTD) [49], could be utilized to prioritize the genes regulated by miRNAs for therapeutic target discovery.

## 5. Conclusions

The network-based strategy proposed in our study can efficiently construct the miRNA-mRNA interaction network in tumorigenesis, which included the important cancer-related miRNAs and their targeted genes. The miRNAs and the targeted genes predicted by using the interaction networks may be considered the potential candidates of drug targets in the cancer research.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Acknowledgments

This work was supported by the grants from the National Natural Science Foundation of China (no. 21575094 and no. 21375090).

## References

[1] Y. Huang, X. J. Shen, Q. Zou, and Q. L. Zhao, "Biological functions of MicroRNAs," *Journal of Physiology and Biochemistry*, vol. 67, no. 1, pp. 129–139, 2011.

[2] S. H. Chan and L. H. Wang, "Regulation of cancer metastasis by microRNAs," *Journal of Biomedical Science*, vol. 22, no. 1, p. 9, 2015.

[3] A. E. Pasquinelli, "MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship," *Nature Reviews Genetics*, vol. 13, no. 4, pp. 271–282, 2012.

[4] D. M. Pereira, P. M. Rodrigues, P. M. Borralho, and C. M. Rodrigues, "Delivering the promise of miRNA cancer therapeutics," *Drug Discovery Today*, vol. 18, no. 5-6, pp. 282–289, 2013.

[5] K. B. Reddy, "MicroRNA (miRNA) in cancer," *Cancer Cell International*, vol. 15, no. 1, pp. 1–6, 2015.

[6] R. Garzon, M. Fabbri, A. Cimmino, G. A. Calin, and C. M. Croce, "MicroRNA expression and function in cancer," *Trends in Molecular Medicine*, vol. 12, no. 12, p. 580, 2006.

[7] M. D. Jansson and A. H. Lund, "MicroRNA and cancer," *Molecular Oncology*, vol. 6, no. 6, p. 590, 2012.

[8] Y. Jin, C. J. Yang, X. Xu, J. N. Cao, Q. T. Feng, and J. Yang, "MiR-214 regulates the pathogenesis of patients with coronary artery disease by targeting VEGF," *Molecular and Cellular Biochemistry*, vol. 402, no. 1, p. 111, 2015.

[9] Y. Mi, D. Zhang, W. Jiang et al., "miR-181a-5p promotes the progression of gastric cancer via RASSF6-mediated MAPK signalling activation," *Cancer Letters*, vol. 389, p. 11, 2016.

[10] C. G. Li, M. F. Pu, C. Z. Li et al., "MicroRNA-1304 suppresses human non-small cell lung cancer cell growth in vitro by targeting heme oxygenase-1," *Acta Pharmacologica Sinica*, vol. 38, no. 1, pp. 110–119, 2017.

[11] M. V. Iorio, M. Ferracin, C. G. Liu et al., "MicroRNA gene expression deregulation in human breast cancer," *Cancer Research*, vol. 65, no. 16, pp. 7065–7070, 2005.

[12] M. V. Iorio and C. M. Croce, "MicroRNA dysregulation in cancer: diagnostics, monitoring and therapeutics. A comprehensive review," *EMBO Molecular Medicine*, vol. 4, no. 3, pp. 143–159, 2012.

[13] G. A. Calin and C. M. Croce, "MicroRNA signatures in human cancers," *Nature Reviews Cancer*, vol. 6, no. 11, pp. 857–866, 2006.

[14] N. Tyagi, S. Arora, S. K. Deshmukh, S. Singh, S. Marimuthu, and A. P. Singh, "Exploiting nanotechnology for the development of microRNA-based cancer therapeutics," *Journal of Biomedical Nanotechnology*, vol. 12, no. 1, pp. 28–42, 2016.

[15] R. Rupaimoole and F. J. Slack, "MicroRNA therapeutics: towards a new era for the management of cancer and other diseases," *Nature Reviews Drug Discovery*, vol. 16, no. 3, pp. 203–222, 2017.

[16] H. L. Janssen, H. W. Reesink, E. J. Lawitz et al., "Treatment of HCV infection by targeting microRNA," *New England Journal of Medicine*, vol. 368, no. 18, pp. 1685–1694, 2013.

[17] A. G. Bader, "miR-34–a microRNA replacement therapy is headed to the clinic," *Frontiers in Genetics*, vol. 3, p. 120, 2012.

[18] M. S. Beg, A. J. Brenner, J. Sachdev et al., "Phase I study of MRX34, a liposomal miR-34a mimic, administered twice weekly in patients with advanced solid tumors," *Investigational New Drugs*, vol. 35, no. 2, pp. 1–9, 2016.

[19] A. Kozomara and S. Griffiths-Jones, "miRBase: annotating high confidence microRNAs using deep sequencing data," *Nucleic Acids Research*, vol. 42, Database issue, pp. D68–D73, 2014.

[20] B. John, A. J. Enright, A. Aravin, T. Tuschl, C. Sander, and D. S. Marks, "Human microRNA targets," *PLoS Biology*, vol. 2, no. 11, article e363, 2004.

[21] I. S. Vlachos, M. D. Paraskevopoulou, D. Karagkouni et al., "DIANA-TarBase v7. 0: indexing more than half a million experimentally supported miRNA: mRNA interactions," *Nucleic Acids Research*, vol. 43, Data issue, pp. D153–D159, 2015.

[22] Y. Li, C. Qiu, J. Tu et al., "HMDD v2. 0: a database for experimentally supported human microRNA and disease associations," *Nucleic Acids Research*, vol. 42, Database issue, pp. D1070–D1074, 2013.

[23] P. Hydbring and G. Badalian-Very, "Clinical applications of microRNAs," *F1000 Research*, vol. 2, no. 2, p. 136, 2013.

[24] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.

[25] B. Xie, Q. Ding, H. Han, and D. Wu, "miRCancer: a microRNA–cancer association database constructed by text mining on literature," *Bioinformatics*, vol. 29, no. 5, pp. 638–644, 2013.

[26] Q. Jiang, Y. Wang, Y. Hao et al., "miR2Disease: a manually curated database for microRNA deregulation in human disease," *Nucleic Acids Research*, vol. 37, Supplement 1, pp. D98–D104, 2009.

[27] G. Dennis Jr, B. T. Sherman, D. A. Hosack et al., "DAVID: database for annotation, visualization, and integrated discovery," *Genome Biology*, vol. 4, no. 9, article R60, 2003.

[28] Network TCGA, Cancer Genome Atlas Research Network, T. J. Ley et al., "Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia," *New England Journal of Medicine*, vol. 368, no. 22, pp. 2059–2074, 2013.

[29] Cancer Genome Atlas Network, D. C. Koboldt, R. S. Fulton et al., "Comprehensive molecular portraits of human breast tumors," *Nature*, vol. 490, no. 7418, pp. 61–70, 2012.

[30] The Cancer Genome Atlas Research Network, "Comprehensive molecular characterization of clear cell renal cell carcinoma," *Nature*, vol. 499, no. 7456, pp. 43–49, 2013.

[31] A. Dobra, C. Hans, B. Jones, J. R. Nevins, G. Yao, and M. West, "Sparse graphical models for exploring gene expression data," *Journal of Multivariate Analysis*, vol. 90, no. 1, pp. 196–212, 2004.

[32] G. Song, L. Han, and K. Xie, "Overlapping decomposition for Gaussian graphical modeling," *IEEE Transactions on Knowledge & Data Engineering*, vol. 27, no. 8, pp. 2217–2230, 2015.

[33] Y. Zuo, Y. Cui, G. Yu, R. Li, and H. W. Ressom, "Incorporating prior biological knowledge for network-based differential gene expression analysis using differentially weighted graphical LASSO," *BMC Bioinformatics*, vol. 18, no. 1, p. 99, 2017.

[34] X. F. Zhang, L. Ou-Yang, X. M. Zhao, and H. Yan, "Differential network analysis from cross-platform gene expression data," *Scientific Reports*, vol. 6, article 34112, 2016.

[35] J. Xu, R. Jing, Y. Liu, Y. Dong, Z. Wen, and M. Li, "A new strategy for exploring the hierarchical structure of cancers by adaptively partitioning functional modules from gene expression network," *Scientific Reports*, vol. 6, article 28720, 2016.

[36] Y. S. Zheng, H. Zhang, X. J. Zhang et al., "MiR-100 regulates cell differentiation and survival by targeting RBSP3, a phosphatase-like tumor suppressor in acute myeloid leukemia," *Oncogene*, vol. 31, no. 1, p. 80, 2012.

[37] J. Bai, A. Guo, Z. Hong, and W. Kuai, "Upregulation of microRNA-100 predicts poor prognosis in patients with pediatric acute myeloid leukemia," *Onco Targets Therapy*, vol. 5, pp. 213–219, 2012, default.

[38] Z. Li, H. Huang, Y. Li et al., "Up-regulation of a HOXA-PBX3 homeobox-gene signature following down-regulation of miR-181 is associated with adverse prognosis in patients with cytogenetically abnormal AML," *Blood*, vol. 119, no. 10, p. 2314, 2012.

[39] S. D. Selcuklu, M. T. Donoghue, K. Rehmet et al., "MicroRNA-9 inhibition of cell proliferation and identification of novel miR-9 targets by transcriptome profiling in breast cancer cells," *Journal of Biological Chemistry*, vol. 287, no. 35, pp. 29516–29528, 2012.

[40] Y. Zhao, Y. Li, G. Lou et al., "MiR-137 targets estrogen-related receptor alpha and impairs the proliferative and migratory capacity of breast cancer cells," *PloS One*, vol. 7, no. 6, article e39102, 2012.

[41] Y. Liu, M. Zhang, J. Qian et al., "miR-134 functions as a tumor suppressor in cell proliferation and epithelial-to-mesenchymal transition by targeting KRAS in renal cell carcinoma cells," *Dna & Cell Biology*, vol. 34, no. 6, pp. 429–436, 2015.

[42] T. Yamasaki, N. Seki, H. Yoshino et al., "MicroRNA-218 inhibits cell migration and invasion in renal cell carcinoma through targeting caveolin-2 involved in focal adhesion pathway," *Journal of Urology*, vol. 190, no. 3, pp. 1059–1068, 2013.

[43] I. Balgkouranidou, A. Karayiannakis, D. Matthaios et al., "Assessment of SOX17 DNA methylation in cell free DNA from patients with operable gastric cancer. Association with prognostic variables and survival," *Clinical Chemistry & Laboratory Medicine*, vol. 51, no. 7, pp. 1505–1510, 2013.

[44] D. Fu, C. Ren, H. Tan et al., "Sox17 promoter methylation in plasma DNA is associated with poor survival and can be used as a prognostic factor in breast cancer," *Medicine*, vol. 4, no. 3, pp. 143–159, 2012.

[45] M. F. Schmidt, "Drug target miRNAs: chances and challenges," *Trends in Biotechnology*, vol. 32, no. 11, pp. 578–585, 2014.

[46] J. K. Nagpal, R. Rani, B. Trink, and K. S. Saini, "Targeting miRNAs for drug discovery: a new paradigm," *Current Molecular Medicine*, vol. 10, no. 5, pp. 503–510, 2011.

[47] S. P. Nana-Sinkam and C. M. Croce, "Clinical applications for microRNAs in cancer," *Clinical Pharmacology & Therapeutics*, vol. 93, no. 1, p. 98, 2013.

[48] C. F. Thorn, T. E. Klein, and R. B. Altman, "PharmGKB: the pharmacogenomics knowledge base," *Methods in Molecular Biology*, vol. 1015, pp. 311–320, 2013.

[49] C. J. Mattingly, G. T. Colby, J. N. Forrest, and J. L. Boyer, "The Comparative Toxicogenomics Database (CTD)," *Environmental Health Perspectives*, vol. 111, no. 6, pp. 793–795, 2003.

Hindawi
International Journal of Genomics
Volume 2017, Article ID 9264034, 7 pages
https://doi.org/10.1155/2017/9264034

*Research Article*

# Uncover the Underlying Mechanism of Drug-Induced Myopathy by Using Systems Biology Approaches

**Dong Li,[1] Aixin Li,[1] Hairui Zhou,[2] Xi Wang,[1] Peng Li,[3] Sheng Bi,[1] and Yang Teng[2]**

[1]*First Affiliated Hospital of Jiamusi University, Jiamusi, Heilongjiang 154002, China*
[2]*Jiamusi University, Jiamusi, Heilongjiang 154002, China*
[3]*Jiamusi Central Hospital, Jiamusi, Heilongjiang 154002, China*

Correspondence should be addressed to Aixin Li; liaixin1981-2005@163.com

Received 12 April 2017; Accepted 8 June 2017; Published 31 July 2017

Academic Editor: Zhichao Liu

Copyright © 2017 Dong Li et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Drug-induced myopathy (DIM) is a rare side effect; however, the consequence could be fatal. There are few reports to systematically assess the underlying mechanism of DIM. In this study, we curated the comprehensive DIM drug list based on structured labeling products (SPLs) and carried out the analysis based on chemical structure space, drug protein interaction, side effect space, and transcriptomic profiling space. Some key features are enriched from each of analysis. Specifically, the similarity of DIM drugs is more significant than random chance, which shows that the chemical structure could distinguish the DIM-positive drugs from negatives. The cytochrome P450 (CYP) was identified to be shared by DIM drugs, which indicated the important role of metabolism in DIM. Three pathways including *pathways in cancer*, *MAPK signaling pathway*, and *GnRH signaling pathway* enriched based on transcriptomic analysis may explain the underlying mechanism of DIM. Although the DIM is the current focus of the study, the proposed approaches could be applied to other toxicity assessments and facilitate the safety evaluation.

## 1. Introduction

Myopathy is a muscular disease in which the muscle fibers do not function, resulting in muscular weakness. There are many causes for myopathy including inheritable genetic defects, metabolic disorder, exposure to toxins, and medication [1]. Although myopathies are not unusual in drug therapy, the consequence could be severe and may cause deaths. Rhabdomyolysis is a severe form of myopathy with muscle breakdown, which leads to myoglobinuria and may result in renal failure and death [2]. Different therapeutic drugs have been associated with myopathy. For example, the statin drugs which are used to lower cholesterol may cause myopathy when it is given in high dose [3]. Therefore, the US Food and Drug Administration (FDA) has recommended to limit the use of the highest approved dose of the cholesterol-lowering medication simvastatin (80 mg) because of increased risk of muscle damage such as myopathy (http://www.fda.gov/Drugs/DrugSafety/ucm256581.htm).

The mechanisms of drug-induced myopathies are complex and unknown. Metabolic change and immune-mediated disorder are two possible mechanisms [1]. Most of studies are based on case reports or focused on the certain therapeutic category. Bonifacio et al. [4] used biochemical experiments to uncover the key role of the AKT/mTOR signaling pathway in statin-induced myotoxicity. Mo et al. introduced a case report on statin-induced myopathy associated with concomitant use of cyclosporine. The patient's serum creatine kinase was significantly increased, which provides the evidence of a potential association between the elevation of creatine kinase and an increased risk of myopathy [5]. Other studies are focused on host factors such as genetic variation for drug-induced myopathy. Link et al. [6] found a strong association of myopathy with the rs4363657 single-nucleotide polymorphism (SNP) located within SLCO1B1 by using genome-wide association study (GWAS). Furthermore, it was reported that SLCO1B1 encodes the organic anion-transporting polypeptide OATP1B1, which could regulate the hepatic uptake of statins. Vicart et al. [7] reported a missense mutation in the alpha B-crystallin chaperone gene that causes a desmin-related myopathy in the French population. However, few reports provide the

systematic way to assess the drug-induced myopathy (DIM), which could provide a better understanding of the underlying mechanisms of drug-induced myopathy and further develop safer drugs with low risk of myopathy.

There are two factors which hindered the deciphering of the mechanism of drug-induced myopathy—the drug itself and host information. Since the myopathy is related to the metabolic levels of individuals, it is worth investigating the genetic factor such as CYP450 in the individuals with myopathy with emerging technologies such as whole-genome sequencing (WGS) or whole-exome sequencing (WES). However, it is still hard to build the causality relationship among the genetic factors and drug taken by the patients, which leads to drug-induced myopathy. Therefore, it is necessary to investigate whether the drug properties also play a role in the drug-induced myopathy.

In this study, we hypothesized that the drug properties are associated with the cause of myopathy. The association between myopathy and the diverse of drug properties including chemical structure, side effects, protein target, and transcriptomic profiling was systematically assessed. The key features were generated to facilitate the mechanistic understanding of drug-induced myopathy.

## 2. Materials and Methods

*2.1. Compilation of the Drug List.* There are public available databases providing drug side effect data, including MetaADEDB [8] and SIDER. In this study, the SIDER database (http://sideeffects.embl.de/) was employed to extract the drugs that could cause myopathy [9]. The SIDER database consists of side effect of drugs in human, which was extracted from multiple version of structured product labeling (SPLs) by using Unified Medical Language System (UMLS) MetaMap tools [9]. In the current version (SIDER2), the database contains 996 FDA-approved drugs and 4192 side effect terms. The 4192 side effect terms were further mapped and normalized into 4500 Medical Dictionary for Regulatory Activities (MedDRA) preferred terms (PTs). We limited the term specifically to "myopathy" to obtain a drug list that causes myopathy. As a result, there are 75 drugs obtained, as shown in Supplementary Table S1 available online at https://doi.org/10.1155/2017/9264034. It is worth mentioning that we did not take into consideration of different frequency ranges in the population for side effects. Furthermore, in order to investigate the relationship between myopathy and other side effects, we composed a whole drug-side effect matrix ($996 \times 4500$) with binary entities (1 denotes drug with side effect, otherwise 0).

*2.2. Retrieval of Drug Properties Information.* Drug structure data files (SDFs) were downloaded from PubChem database (https://pubchem.ncbi.nlm.nih.gov/) [10]. Then, the chemical descriptors were generated by using KNIME v2.5.1 (https://www.knime.org/). Specifically, the Extended Connectivity Fingerprints (ECFP-4) were used, which is well-established and extensively applied in chemical structure analysis.

The therapeutic categories of drugs were extracted from the WHO Anatomical Therapeutic Chemical (ATC) Classification System (http://www.whocc.no/atc_ddd_index/). The ATC code is a hieratical ontology structure with five levels. Naïve Bayesian classifier was used to assess the disproportionality of a specific ADR drug combination against the ADR distribution for all drugs in the global ADR database. In this study, we used the second level of the code that indicated the therapeutic main group. The details of drug profiles and side effect information were listed in Supplementary Table S1.

Drug protein target information was extracted from the DrugBank (version 4.3) database (http://www.drugbank.ca/) [11]. The target information in DrugBank was divided into four categories including therapeutic targets, enzymes, transporters, and carriers. In this study, only the protein targets in *Homo sapiens* species were employed.

*2.3. Drug Transcriptomic Profiling Data.* Transcriptomic profiles of the drugs were obtained from the Connectivity Map (version 02) generated by the Broad Institute of MIT (http://www.broadinstitute.org/cmap/) [12]. The Connectivity Map contains a collection of genome-wide transcriptional expression data of 1309 drugs from different cultured human cancer cells treated using Affymetrix Human Genome U133A 2.0 arrays. Since the array data are from different cancer cell lines, we used the prototype ranked list (PRL) by merging all the ranked lists referring to the same compound from different cell lines by using the Borda merging method [13]. Then, for each compound, the top 100 and down 100 regulated genes from PRL were considered as signature genes. Finally, we ranked the signature genes based on the frequency of myopathy drugs involved. The top 100 genes were used as the representative genes for drug-induced myopathy for further analysis.

*2.4. Functional Analysis.* Two types of functional analysis were used to interpret 100 drug-induced myopathy representative genes. First, the KEGG pathway analysis was performed to identify the significant pathways for 100 drug-induced myopathy representative genes with Fisher's exact test with multiple testing corrections by using DAVID tool (https://david.ncifcrf.gov/) [14]. Here, the signature pathways were considered with Benjamini-Hochberg (BH) adjusted $p$ value less than 0.05. Then, the 100 drug-induced myopathy representative genes were mapped into a protein-protein interaction network to investigate the physical connection of genes and their functional similarity. In detail, the STRING 9.1 version [15] was applied to study protein-protein-interaction (PPI) using 100 drug-induced myopathy representative genes as input, and PPI were considered with confidence scores more than 0.4.

*2.5. Myopathy-Related Side Effects.* In order to identify the significant myopathy and related side effects, the Fisher's exact test and seven confusion matrices were generated based on drugs related to myopathy and other side effect information as described in [16] and the formulas as follows:

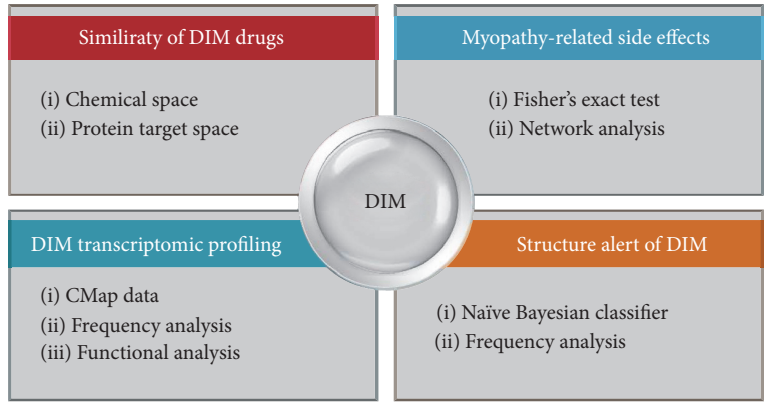



Figure 1: Flowchart of the study: (1) similarity between DIM drugs based on chemical similarity and protein distance in PPI network; (2) myopathy-related side effects based on Fisher's exact test and network analysis; (3) the DIM transcriptomic analysis based on CMap data; (4) structure alert of DIM.

| | | Side effect | |
|---|---|---|---|
| | | Yes | No |
| Myopathy | Yes | TP | TN |
| | No | FP | FN |

$$
\begin{aligned}
p &= \frac{\binom{\mathrm{TP}+\mathrm{FN}}{\mathrm{TP}}\binom{\mathrm{FP}+\mathrm{TN}}{\mathrm{FP}}}{\mathrm{TP}+\mathrm{FN}+\mathrm{FP}+\mathrm{TN}},\\
\text{Accuracy} &= \frac{\dfrac{\mathrm{TP}+\mathrm{FP}}{\mathrm{TP}+\mathrm{TN}}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}},\\
\text{Sensitivity} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}},\\
\text{Specificity} &= \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}},\\
\mathrm{MCC} &= \frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})\times(\mathrm{TP}+\mathrm{FN})\times(\mathrm{TN}+\mathrm{FP})\times(\mathrm{TN}+\mathrm{FN})}},\\
\mathrm{AUC} &= \frac{\text{Sensitivity}+\text{Specificity}}{2},\\
\mathrm{PPV} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}},\\
\mathrm{NVP} &= \frac{\mathrm{TN}}{\mathrm{FN}+\mathrm{TN}},
\end{aligned}
\tag{1}
$$

where TP (true positive) represents a list of drugs with investigated side effect and myopathy, FP (false positive) means a list drugs with investigated side effect but without the myopathy, TN (true negative) denotes a list of drugs not belong to investigated side effect but belong to myopathy, and FN (false negative) represents drugs neither involved in investigated side effect nor involved in myopathy.

For Fisher's exact test, the two-side $p$ value was used with multiple testing corrections. As a result, significant genotype/phenotype and ADR (adverse drug reaction) pair was considered if the corrected $p$ value $<0.05|\mathrm{MCC}\geq 0.2|\mathrm{Sen}\geq 0.70|\mathrm{TP}>2$. The threshold for MCC and sensitivity was decided by using empirical probability distribution function (pdf) of all the possible pairs of association; specifically, the significant thresholds located in the upper bound of the 5% quantile of the empirical pdf. The TP was decided by the consideration of the statistical power of measure, which is even high than the similar approaches [16].

*2.6. Similarity of Myopathy Drugs.* The chemical structure similarity was assessed based on ECFP-4 descriptors by using the Jaccard similarity coefficient, as shown as follows:

$$
J\left(\mathrm{drug}_i, \mathrm{drug}_j\right) = \frac{\mathrm{drug}_i \cap \mathrm{drug}_j}{\mathrm{drug}_i \cup \mathrm{drug}_j}. \tag{2}
$$

Here, the similar drug pairs were extracted with the Jaccard similarity coefficient more than 0.4.

Furthermore, the myopathy drug similarity was assessed based on their shared protein target and distance in the PPI interaction network.

## 3. Results

The workflow of this study is illustrated in Figure 1. In order to systematically assess the underlying mechanisms of drug-induced myopathy, four types of analysis were carried out including (1) similarity among of DIM drugs, (2) myopathy-associated side effects, (3) transcriptomic analysis of DIM drugs and their involved pathways, and (4) structure alert of drug-induced myopathy.

*3.1. Similarity among the DIM Drugs.* There are a total of 75 DIM drug extracted from SIDER2 database. The distribution of DIM drugs involved therapeutic categories were shown in Figure 2. As shown in Figure 2, corticosteroids drugs such as statin are more susceptible to myopathy than other therapeutic categories. It is consistent with current observation that statin drugs were more associated with myopathy [17]. In addition, central nervous system (CNS) agents (*N02*, *N06*, and *N03*) are also needed to be cared for myopathy risk. Especially, CNS drugs are taken in high dose.

In order to investigate whether the DIM-positive drugs share more commonality than DIM-negative drugs, we

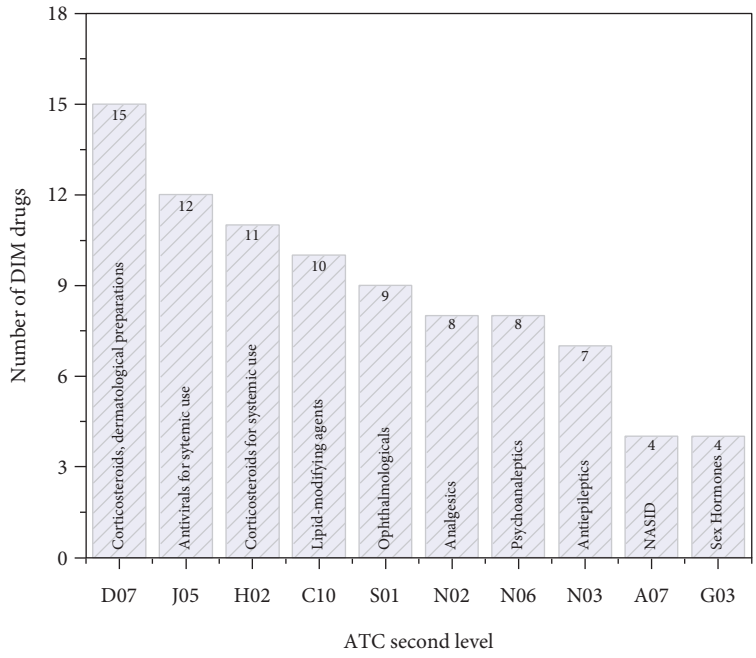



Figure 2: The therapeutic categories distribution of drug-induced myopathy (DIM) drugs. The DIM drugs were mapped to the second level of the WHO Anatomical Therapeutic Chemical (ATC) Classification System. Then, for each therapeutic category, the number of DIM drugs were counted.

carried out pair-wise similarity analysis based on chemical structures, which generated a total of 2775 similarity pairs for 75 DIM-positive drugs. We also randomly selected the same number of DIM-negative drugs to generate a negative control for comparison. The process was repeated for 10,000 times. For each time, the $F$ test was employed to investigate whether the two lists of similarity values are statistically different. The distribution of median similarity distribution values for 10,000 times randomization test is shown in Figure 3. The median similarity values of 2775 DIM-positive drug pairs are statistically larger than those of 100,000 permutation test results (Figure 3). It is demonstrated that the DIM drugs have some chemical structure similarity that could be used for differentiation from DIM-negative drugs.

In addition, we investigated the common targets among the 75 DIM drugs. Figure 4(a) lists the top 10 targets that are most frequently interacting with DIM drugs. It could be seen that the cytochrome P450s family dominated, indicating that the metabolism is playing a crucial role in the mechanism of DIM. It was reported that statin drugs that usually cause myopathy through statin metabolism via the CYP system [18]. Then, we mapped the 10 protein targets into STRING 9.3 PPI network to investigate whether they are interacted with each other (Figure 4(b)). It was observed that the 8 of 10 protein targets are strongly interacting with each other, which shows the similar underlying mechanism of DIM.

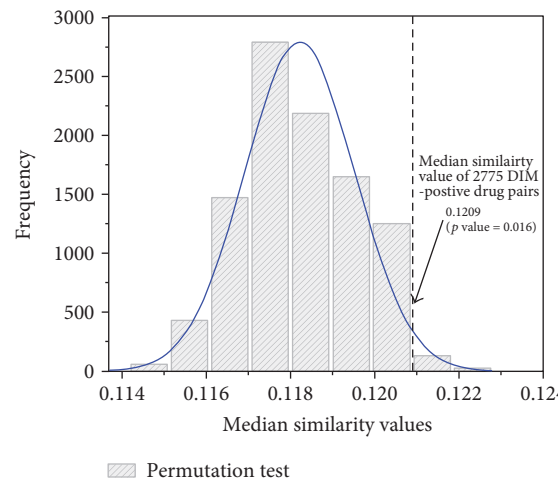
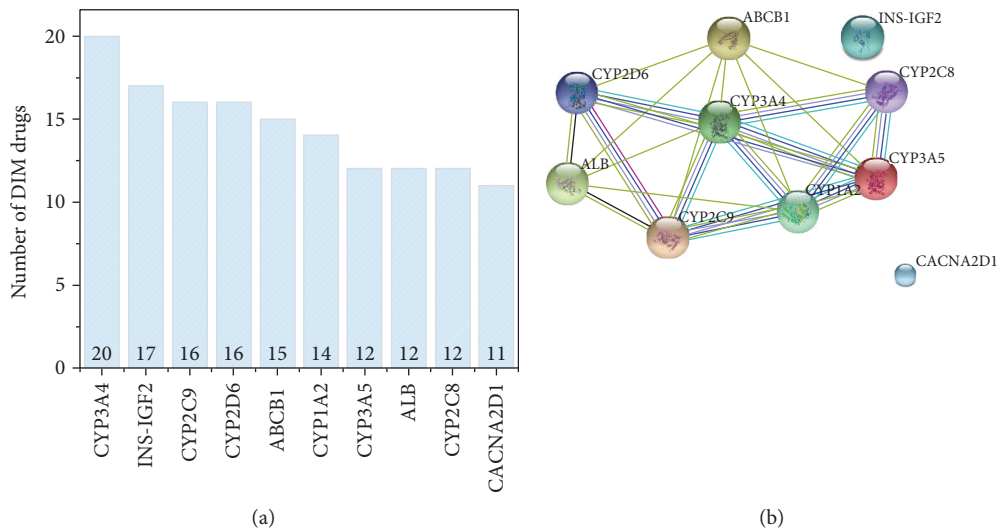



Figure 3: Permutation test for similarity of DIM drug pairs. The distribution of 75 DIM-positive drugs were drawn based on their chemical similarity. Then, the random test was carried out based on chemical similarity of negative DIM drugs, which are randomly picked up for 100,000 times. Then, the $p$ value could be calculated for assessing whether the DIM-positive drugs are more similar.

*3.2. Myopathy-Associated Side Effects.* The cooccurrent side effects may indicate the similarity of mechanisms or causality relationship between each other. Therefore, we extracted myopathy-associated side effects to investigate whether they shared the same mechanism. As mentioned in Section 2, Fisher's exact test and network visualization were used to perform the side effect similarity analysis. There are 39 side effects associated with myopathy (Figure 5). We mapped the 39 side effects (PTs) into System Organ Class (SOC) level of MedDRA to extract their organ attributes. We highlighted




(a) (b)

Figure 4: The protein space of DIM drugs. (a) The top 10 protein targets for DIM; the DIM drug and target relationship was extracted from DrugBank. (b) The STRING PPI for the top 10 protein targets; the top 10 proteins corresponding to more DIM drugs were inputted to the STRING PPI database to exact the subnetwork among the 10 proteins.

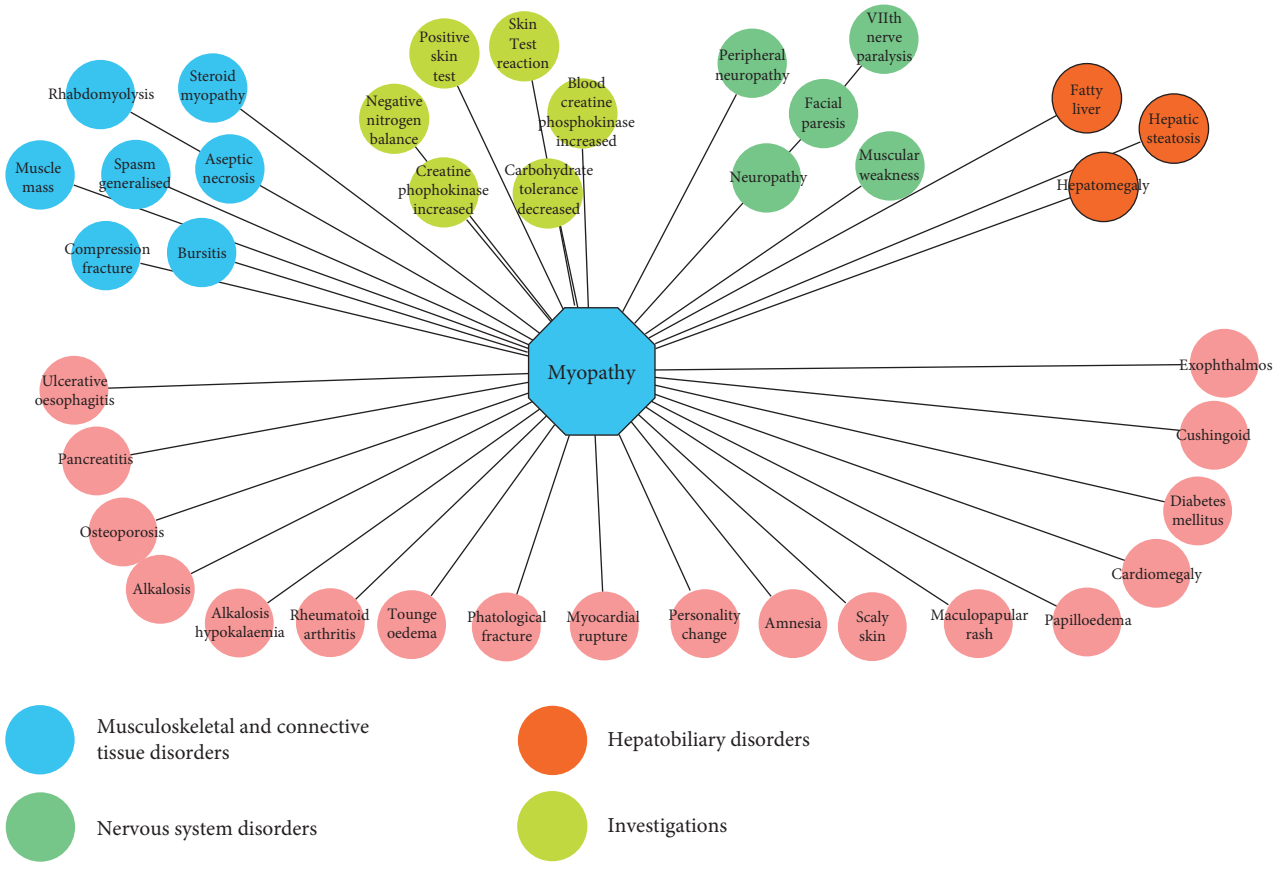



Figure 5: Network myopathy-related side effects. The myopathy and side effects were extracted based on SIDER database.

four SOCs (i.e., musculoskeletal and connective tissue disorders, investigations, nervous system disorders, and hepatobiliary disorders) with more side effects (Figure 5). It was indicated that myopathy had multiple organ association with other side effects, which partly explained the multiple organ toxicity of certain therapeutic drugs [19].

We further conducted a literature survey in PubMed by using "myopathy" and enriched side effect to investigate

Table 1: Selected examples of myopathy and associated side effects by literature survey.

| Myopathy-associated side effects | $p$ value by Fisher's exact test | Notes | References |
|---|---|---|---|
| Diabetes mellitus | 1.00$E$−11 | In any diabetic condition, a failure to maintain healthy muscle is often observed and is termed diabetic myopathy | PMID: 24391596 |
| Rhabdomyolysis | <1.00$E$−15 | Rhabdomyolysis is a severe form of myopathy | PMID: 25991405 |
| Creatine phosphokinase increased | 7.00$E$−11 | Increased dosage of cyclosporine induces myopathy with increased serum creatine kinase in an elderly patient on chronic statin therapy | PMID: 25512016 |
| Neuropathy | <1.00$E$−15 | Critically ill patients may develop muscle weakness or paralysis such as neuropathy during the course of sepsis and multiple organ failure | PMID: 15758592 |
| Alkalosis hypokalaemia | 6.11$E$−09 | A case report about hypokalemia-induced myopathy as the first manifestation of primary hyperaldosteronism due to unilateral adrenal hyperplasia | PMID: 19829865 |
| Rheumatoid arthritis | 9.69$E$−09 | Based on 86 patients with verified rheumatoid arthritis, 5.8% patients were found with peripheral myopathy | PMID: 13917616 |
| Osteoporosis | 3.90$E$−10 | Chronic use of glucocorticoids (GCs) is the most common cause of secondary osteoporosis. The glucocorticoids induce myopathy as well | PMID: 22870429 |
| Hepatic steatosis | 7.20$E$−10 | A case report showed that zidovudine treatment can induce mitochondrial multisystem disease, as revealed in our case by myopathy, liver steatosis, and lactic acidosis | PMID: 9927163 |
| Cardiomegaly | 1.60$E$−10 | Two unrelated 16-year-old boys had mental retardation, cardiomegaly, and proximal myopathy | PMID: 6450334 |

Table 2: Enriched KEGG pathways for 100 representative genes.

| Pathways | Number of hits | Involved genes | Adjusted $p$ value |
|---|---|---|---|
| Pathways in cancer | 9 | Jun, Runx1, MAX, Fas, FGF22, HSP90AA2, Cdk2, ETS1, Cdc42 | 0.005 |
| MAPK signaling pathway | 7 | Jun, MAX, Fas, FGF22, mapt, Il1r1, Cdc42 | 0.024 |
| GnRH signaling pathway | 4 | Jun, ITPR2, Gnas, Cdc42 | 0.049 |

whether the association generated could be verified by the independent studies (Table 1). For example, we found that a drug that could induce myopathy also tends to induce rheumatoid arthritis. Of the 86 patients with verified rheumatoid arthritis, 5.8% patients were found with peripheral myopathy. This observation suggests that there may be a common mechanism between myopathy and rheumatoid arthritis. Many independent verified studies in Table 1 were either case reports or control population studies, which is time-consuming and cost-intensive. Such results showed the applicability of our proposed approach on assessing the side effect relationship.

*3.3. Transcriptomic Analysis of DIM Drugs.* We mapped the 75 DIM to CMap version 02 and obtained 29 common DIM drugs. As mentioned in Section 2, the frequency analysis was conducted based on the merged signature for each DIM drug. Then, the top 100 genes were extracted as representative genes to carry out the functional analysis (Table S2). Table 2 shows the enriched KEGG pathways for the 100 representative genes for DIM. Three pathways including *pathways in cancer*, *MAPK signaling pathway*, and *GnRH signaling pathway* were enriched. It was reported that the transforming growth factor-beta (TGF-$\beta$) superfamily includes a variety of cytokines expressed in the skeletal muscle. Members of this superfamily that are of great importance in the skeletal muscle are TGF-$\beta$1, mitogen-activated protein kinases (MAPKs), and myostatin [20]. Therefore, this shows that the underlying mechanism of myopathy could be related to MAPK pathways.

## 4. Discussion

Although the current drug toxicity focus in clinical trials is liver and cardiovascular toxicity, the high incidence of other organ toxicities such as myopathy was also worth studying. In this study, we employed systematic approaches to provide a landscape of drug-induced myopathy, which aims to provide the better understanding of the underlying mechanism of DIM for further development of safer drugs. Specifically, the DIM drugs are assessed based on chemical structure space, phenotypic information, and transcriptomic profiling. Some key features in each space are enriched, which could be used to develop screening assay or in silico approaches to further facilitate the research in this field.

Due to the prevalence issue, some side effects have more case reports than others. It limits the understanding due to idiosyncratic natures. Here, we assess the side effect similarity

to enrich the myopathy-related side effect and further could derive the potential cause of myopathy. Furthermore, the multiple organ toxicity could be assessed by the proposed approach, which helps the understanding of toxicity from systems biology's point of views.

Some caveats are also important to mention. (i) In this study, we do not take into consideration of the host factor for drug-induced myopathy. However, it could play an important role. For instance, based on our analysis, we found that the CYP450 family proteins were enriched by DIM drugs, which could be quite different from individuals. However, considering the idiosyncratic nature of myopathy and causality relationship between drug and host, we only focus on drug sides. (ii) Transcriptomic data in this study was based on different cancer cell lines, which could be quite different with the muscle cells. Therefore, we do see the enriched pathways are mostly cancer related. (iii) Some other data types such as in vitro assay and high-content assay data could also provide some unique understanding for DIM. All these caveats will be accomplished in our future study.

## 5. Conclusions

In summary, the proposed analysis pipeline could provide systematic approach to understand DIM. Although the DIM is the current focus of the study, the proposed approaches could be applied to other toxicity endpoints.

## Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publications of this paper.

## References


[1] J. S. Le Quintrec and J. L. Le Quintrec, "Drug-induced myopathies," *Baillière's Clinical Rheumatology*, vol. 5, pp. 21–38, 1991.

[2] R. A. Zager, "Rhabdomyolysis and myohemoglobinuric acute renal failure," *Kidney International*, vol. 49, pp. 314–326, 1996.

[3] P. D. Thompson, P. Clarkson, and R. H. Karas, "Statin-associated myopathy," *Journal of the American Medical Association*, vol. 289, pp. 1681–1690, 2003.

[4] A. Bonifacio, G. M. Sanvee, J. Bouitbir, and S. Krähenbühl, "The AKT/mTOR signaling pathway plays a key role in statin-induced myotoxicity," *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, vol. 1853, pp. 1841–1849, 2015.

[5] L. Mo, J. He, Q. Yue, B. Dong, and X. Huang, "Increased dosage of cyclosporine induces myopathy with increased seru creatine kinase in an elderly patient on chronic statin therapy," *Journal of Clinical Pharmacy and Therapeutics*, vol. 40, pp. 245–248, 2015.

[6] E. Link, S. Parish, J. Armitage et al., "SLCO1B1 variants and statin-induced myopathy- a genomewide study," *The New England Journal of Medicine*, vol. 359, pp. 789–799, 2008.

[7] P. Vicart, A. Caron, P. Guicheney et al., "A missense mutation in the alpha B-crystallin chaperone gene causes a desmin-related myopathy," *Nature Genetics*, vol. 20, pp. 92–95, 1998.

[8] F. Cheng, W. Li, X. Wang et al., "Adverse drug events: database construction and in silico prediction," *Journal of Chemical Information and Modeling*, vol. 53, pp. 744–752, 2013.

[9] M. Kuhn, M. Campillos, I. Letunic, L. J. Jensen, and P. Bork, "A side effect resource to capture phenotypic effects of drugs," *Molecular Systems Biology*, vol. 6, 2010.

[10] W. D. Ihlenfeldt, E. E. Bolton, and S. H. Bryant, "The PubChem chemical structure sketcher," *Journal of Cheminformatics*, vol. 1, 2009.

[11] V. Law, C. Knox, Y. Djoumbou et al., "DrugBank 4.0: shedding new light on drug metabolism," *Nucleic Acids Research*, vol. 42, pp. D1091–D1097, 2014.

[12] J. Lamb, E. D. Crawford, D. Peck et al., "The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease," *Science*, vol. 313, pp. 1929–1935, 2006.

[13] F. Iorio, R. Bosotti, E. Scacheri et al., "Discovery of drug mode of action and drug repositioning from transcriptional responses," *Proceedings of the National Academy of Sciences*, vol. 107, pp. 14621–14626, 2010.

[14] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources," *Nature Protocols*, vol. 4, pp. 44–57, 2008.

[15] D. Szklarczyk, A. Franceschini, S. Wyder et al., "STRING v10: protein–protein interaction networks, integrated over the tree of life," *Nucleic Acids Research*, vol. 43, pp. D447–D452, 2015.

[16] L. Yang and P. Agarwal, "Systematic drug repositioning based on clinical side-effects," *PLoS One*, vol. 6, article e28025, 2011.

[17] C. M. Ballantyne, A. Corsini, M. H. Davidson et al., "Risk for myopathy with statin therapy in high-risk patients," *Archives of Internal Medicine*, vol. 163, pp. 553–564, 2003.

[18] T. T. Abd and T. A. Jacobson, "Statin-induced myopathy: a review and update," *Expert Opinion on Drug Safety*, vol. 10, pp. 373–387, 2011.

[19] A. Corsonello, A. Abbatecola, S. Fusco et al., "The impact of drug interactions and polypharmacy on antimicrobial therapy in the elderly," *Clinical Microbiology and Infection*, vol. 21, pp. 20–26, 2015.

[20] T. N. Burks and R. D. Cohn, "Role of TGF-beta signaling in inherited and acquired myopathies," *Skeletal Muscle*, vol. 1, 2011.