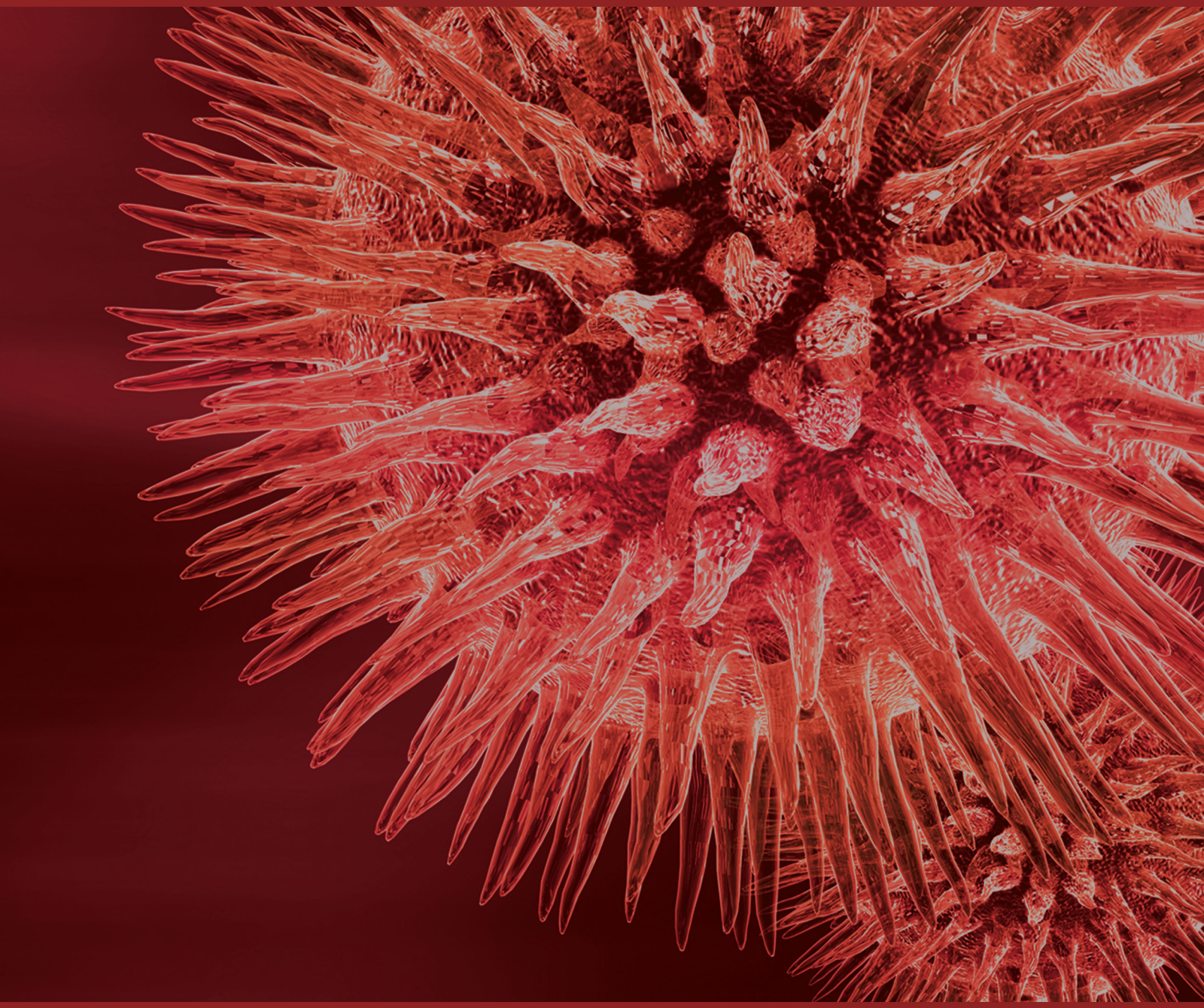
BioMed Research International

Advanced Computational Approaches for Medical Genetics and Genomics

Guest Editors: Zhi Wei, Xiao Chang, and Junwen Wang



Advanced Computational Approaches for Medical Genetics and Genomics

BioMed Research International

Advanced Computational Approaches for Medical Genetics and Genomics

Guest Editors: Zhi Wei, Xiao Chang, and Junwen Wang

Contents

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 705469, 2 pages
http://dx.doi.org/10.1155/2015/705469

Editorial

Advanced Computational Approaches for Medical Genetics and Genomics

Zhi Wei,[1] Xiao Chang,[2] and Junwen Wang[3]

[1]*Department of Computer Science, New Jersey Institute of Technology, Newark, NJ 07102, USA*
[2]*The Center for Applied Genomics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA*
[3]*Centre for Genomic Sciences and Department of Biochemistry, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong*

Correspondence should be addressed to Zhi Wei; zhi.wei@njit.edu

Received 9 June 2015; Accepted 9 June 2015

Remarkable advances have been made in genetics and genomics over the last decade due to the rapid technology innovation in microarray and sequencing. Thanks to biomedical discoveries, it is feasible now to improve the diagnosis and treatment by genetic and genomic tests, for paving the way for an era of personalized/precision medicine in health care. However, it remains challenging to identify causal mutations from massive amounts of genomic data. There is an unprecedented demand for novel computational methods and analytical strategies to improve the accuracy of variants identification and the power of association tests. This special issue aims to publish applications of innovative analysis pipelines and algorithms to find the better solutions of complex genetic and genomic problems in a time efficient manner. We look for original research findings and practical applications that contribute to the diagnosis and management of human disorders.

In this special issue, we selected ten papers from dozens of submissions after in-depth review. We summarize their key contributions and findings as follows.

In the field of *medical genetics*, we collected three research papers and one review paper. S. Perez-Alvarez et al. developed a statistical algorithm (FARMS) for variable selection aiming to identify variables with the optimal predication performance for a specific outcome. The authors further applied FARMS to a high-dimensional dataset of over 800 individuals and showed that the proposed method is more efficient than other approaches such as regression based method. Y. Wang et al., in another methodology paper, proposed a novel method (LRSDec) to identify gene modules and genetic interactions between them. It is based on regularized low-rank approximation and enjoys nearly optimal error bounds. We expect its wide applications in the field of genetic interaction data analysis, image processing, and so on. Using a statistical genomic approach T. Du et al. reported the discovery of the association between FSHR polymorphisms and polycystic ovary morphology in women with polycystic ovary syndrome. In a review article, R. de Vlaming and P. J. F. Groenen surveyed the use of ridge regression for prediction in quantitative genetics by genotyping data. They also performed a suite of simulations to estimate the effect of sample size, the number of SNPs, and trait heritability on the accuracy of the results.

In the field of *medical genomics*, we collected six research articles. Four of them employed network-based computational strategies to investigate the molecular mechanisms of complex diseases including cancer, obesity, and Type 2 Diabetes (T2D). By computing coexpression gene pairs in two types of lung cancers and the normal lung tissues, F. Long et al. identified molecular biomarkers that distinguish small-cell lung cancer and non-small-cell lung cancer. In another article, Q. Zou et al. proposed a network-based method to predict the association between microRNAs and diseases and further developed it into a web server (http://datamining.xmu.edu.cn/~jinjinli/MircoDAP.html.) for use by the community. In "Network-Based Association Study of Obesity and Type 2 Diabetes with Gene Expression Profiles," S. Zhang et al. integrated multiple omics data of obesity and T2D to construct a comprehensive biological network. Their novel strategy revealed the pathways associated with both obesity and T2D. Another article by S. Park et al. explored the impact of

inflammatory responses to the risk of colorectal cancer and Alzheimer's disease in the view of systems biology. Y. Sun and Q. Liu presented a computational framework for deciphering the correlation between breast tumor samples and cell lines. They proposed to integrate both copy-number changes and gene expression profiles. Their investigation can serve as a useful guide to bridge the gap between cell lines and tumors and help to select the most suitable cell line models for personalized cancer studies. The article by T. Liu et al. conducted a gene coexpression and evolutionary conservation analysis of the human preimplantation embryos. Their study demonstrated the patterns of evolutionary constraints that were imposed on different stages of human preimplantation embryos.

We believe the publications in this collection will attract and benefit readers from multiple disciplines, not only bioinformaticians and statistical geneticists for methodology development, but also researchers and clinicians in using the novel tools for their science discovery.

Acknowledgments

We would like to express our sincere gratitude and appreciation to the anonymous reviewers for their time and effort. We are also grateful to all authors for their excellent and fundamental contributions and their patience in communicating with us.

Zhi Wei
Xiao Chang
Junwen Wang

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 685303, 8 pages
http://dx.doi.org/10.1155/2015/685303

Research Article

Identification of Gene Biomarkers for Distinguishing Small-Cell Lung Cancer from Non-Small-Cell Lung Cancer Using a Network-Based Approach

Fei Long,[1] Jia-Hang Su,[2] Bin Liang,[1] Li-Li Su,[1] and Shu-Juan Jiang[1]

[1]*Department of Respiratory Medicine, Shandong Provincial Hospital Affiliated to Shandong University, Jinan, Shandong 250021, China*
[2]*Department of Pharmacy, Yantai Chinese Medicine Hospital, Yantai, Shandong 264100, China*

Correspondence should be addressed to Li-Li Su; sulili801@126.com and Shu-Juan Jiang; shujuan-jiang@163.com

Received 12 January 2015; Revised 10 March 2015; Accepted 17 March 2015

Academic Editor: Xiao Chang

Lung cancer consists of two main subtypes: small-cell lung cancer (SCLC) and non-small-cell lung cancer (NSCLC) that are classified according to their physiological phenotypes. In this study, we have developed a network-based approach to identify molecular biomarkers that can distinguish SCLC from NSCLC. By identifying positive and negative coexpression gene pairs in normal lung tissues, SCLC, or NSCLC samples and using functional association information from the STRING network, we first construct a lung cancer-specific gene association network. From the network, we obtain gene modules in which genes are highly functionally associated with each other and are either positively or negatively coexpressed in the three conditions. Then, we identify gene modules that not only are differentially expressed between cancer and normal samples, but also show distinctive expression patterns between SCLC and NSCLC. Finally, we select genes inside those modules with discriminating coexpression patterns between the two lung cancer subtypes and predict them as candidate biomarkers that are of diagnostic use.

1. Introduction

Lung cancer is the most commonly occurring type of cancers worldwide. In China, lung cancer has become the first cause of cancer death in China, with the fastest rising mortality in population [1]. According to the third nationwide Sampling Survey of Death Cause Review conducted by China's Ministry of Health in 2006, the mortality caused by lung cancer has increased by 75.77% since 1990s and has increased by 33.25% after excluding factors of age structure changes. Although it has been widely recognized that smoking is the most related risk factor for lung cancer [2], the complete understanding of pathogenesis, disease diagnostics, and the development of therapy of lung cancer are still under active research.

Lung cancer consists of two major histological types: small-cell lung cancer (SCLC) and non-small-cell lung cancer (NSCLC) [3]. SCLC is defined as "a malignant epithelial tumor consisting of small cells with scant cytoplasm, ill-defined cell borders, finely granular nuclear chromatin, and absent or inconspicuous nucleoli" [4]. Though SCLC is not very common in lung cancers (about 20%), it has a strong relationship with smoking and has a rapid growth rate, early metastases, and high initial response rates [5]. It usually arises from major bronchi centrally, with extensive mediastinal adenopathy [3]. NSCLC includes squamous-cell (epidermoid) carcinoma, adenocarcinoma, and large-cell carcinoma. Adenocarcinoma and large-cell carcinoma usually arise from the small bronchi, bronchioles, or alveoli (peripheral tumors) of the distant airway of the lung peripherally, whereas squamous-cell carcinoma usually arise centrally [6]. Squamous-cell carcinoma is characterized by lobar collapse, obstructive pneumonia, or hemoptysis and shows late development of distant metastases. Adenocarcinoma is associated with early development of metastases, with some

Table 1: Summary of datasets used in our study.

GEO ID	Sample type	Sample number	Sample information	Reference
GSE23546	Normal	1349	Nontumor lung tissues from patients with lung cancer	[13]
GSE41271	NSCLC	275	275 tumor specimens from ~1,700 non-small-cell lung cancer specimens collected at the MD Anderson Cancer Center over the years 1997 to 2005	[14]
GDS4794	SCLC	23	23 clinical small-cell lung cancer (SCLC) samples from patients undergoing pulmonary resection	[15]
GSE62021	SCLC	25	25 SCLC tumor tissues	Unpublished

of the primary tumors as a symptomless peripheral lesion. Large-cell carcinoma has large peripheral masses, sometimes with cavitation [3].

Lung cancer patients usually have no symptoms or have nonspecific symptoms such as shortness of breath, coughing, and weight loss at the early stage of cancer development, making early diagnosis of lung cancer extremely difficult. Most diagnosed lung cancer patients are at their late stage of cancer development, which is the main reason for the high mortality rate of lung cancer. It is therefore of dire need to develop sensitive and reliable tools for diagnosis of lung cancer. Usually, the diagnosis of lung cancer involves (i) the identification and complete classification of malignancy, (ii) immunohistochemistry to distinguish lung cancer subtypes, and (iii) molecular testing [7]. Given that different subtypes of lung cancer have significantly different responsiveness to treatment, for example, SCLC usually responds better to chemotherapy and radiotherapy, whereas NSCLC patients often have to be treated with surgery [8], in this study we focus on developing method for distinguishing different types of lung cancer that can be of diagnostic use.

During the development and progression of cancer, there are significant genetic and molecular alterations in cells, which can be permanent, irreversible, and dynamic and cause significant variations in gene expression [9]. Thus, detecting genes with signature expression change can provide diagnosis markers. For instance, Zhang et al. analyzed the genome-wide expression profiles of both mRNAs and miRNAs in three NSCLC cell lines and identified hundreds of genes and 10 miRNAs with significant expression changes that can be used as potential diagnosis markers [10]. Mishra et al. investigated the differential gene expression profile in different lung cancer cell line and discovered a number of genes whose expression is significantly correlated with patients' survival [11]. Here, we aim to identify genes with significant gene expression change between the two subtypes of lung cancer. This could be done by simply comparing the expression level of a gene between the two subtypes of lung cancer. However, mounting evidence has suggested that this is not a reliable approach. First, this approach is not robust and is significantly biased by expression noise [12]. Second, the top differentiated expressed genes may not be the causal genes but the downstream response genes. Third, in cell genes do not function by themselves but in a complicated network, and ignoring the relationships between genes will significantly reduce the power of finding the truly important genes. Thus, we attempt to tackle the problem by integrating gene expression profiles with protein-protein interaction information using a network-based approach. We first collect gene expression datasets from GEO, including normal lung tissue, SCLC, and NSCLC [13–15]. Then, after determining gene pairs with significant positive or negative correlation in gene expression in each of the three types of datasets, we map those gene pairs to the STRING network [16], a large-scale gene functional association network, and then construct a lung cancer-specific functional association network. We further partition this network into gene modules and identify the modules with significant association with either SCLC or NSCLC in terms of gene expression variation. These gene modules are considered to be potentially useful for distinguishing SCLC from NSCLC. Functional enrichment analysis has revealed that those gene modules are highly related to lung cancer development. Finally, from the gene modules we identify genes with specific association with either SCLC or NSCLC. These genes can be exploited as potential biomarkers of diagnostic use for distinguishing SCLC or NSCLC.

2. Methods and Materials

2.1. Dataset Collection and Processing. We use GEO query [17] to download gene expression datasets from GEO database [18] and prepare three datasets corresponding to normal lung tissue, SCLC, and NSCLC, respectively (Table 1). These three datasets are selected using the following rules: (1) the samples must be human tissue samples; (2) the sample size should be greater than 100; and (3) there is no special treatment. Since SCLC is not common, we reduce the sample size to be more than 20. We use R package limma [19] for gene expression normalization. Then, for each gene expression dataset we calculate Pearson Correlation Coefficient (PCC) for each pair of genes using their expression profile and select the top 1% and bottom 1% gene pairs as positively or negatively coexpressed genes, respectively. These positively or negatively coexpression gene pairs from the three datasets are combined together and are mapped to the STRING [16] network. These mapped gene pairs with a functional association score greater than 500 in the STRING network are retained and are used to construct a lung cancer-specific functional association network. This network is then partitioned into gene modules using a network partition algorithm called iNP [20]. For validation purpose, we also prepare a new dataset for NSCLC (GSE10245 [21]) whose sample size is 58 and repeat

the above procedures to construct a new lung cancer-specific functional association network.

2.2. Identification of Lung Cancer-Specific Gene Modules. For each of the partitioned gene modules, we test whether it has significant gene expression variation between normal, SCLC, and NSCLC samples. Since each module includes a group of genes that are significantly functionally related to each other, a module can be considered as a pathway. There are many ways to determine whether a pathway is significantly differentially expressed between cancer and normal samples. Here, we choose a simple method, the median expression value. We use the median expression value for all genes inside the module as the representative expression value for the module in a given sample. Then, we perform *t*.test to compare the module's expression values in either SCLC or NSCLS samples against those in normal samples and define the differentially expressed modules between lung cancer and normal samples as those with P value < 0.01 (adjusted by FDR) and the log (fold change) greater than 2. Next, from those differentially expressed modules we further quantitatively determine their specificity in distinguishing SCLC or NSCLC by plotting a ROC curve using the module's expression value in SCLC samples against the values in NSCLC samples. The AUC (area under curve) of the ROC curve is calculated and is used to indicate the specificity of the module to either SCLC or NSCLC. The AUC ranges from 0 to 1, with random association equaling to 0.5. AUC of 0 or 1 corresponds to perfect specific in distinguishing the two lung cancer subtypes. In addition, if the AUC is greater than 0.5, it indicates that the module tends to be upregulated in SCLC compared to that in NSCLC or vice versa. We use Cytoscape 3.1 to display module structure and use R to draw clustering figures.

2.3. Function Enrichment Analysis. Given a selected gene module, we perform function enrichment for the genes inside the module. GO annotation file is downloaded from [22] on Nov. 23rd, 2014. Pathway annotation from MSigDB is downloaded from GSEA [23]. Biological process GO terms and MSigDB pathways are tested for enrichment using Fisher's test() in R. The significance of threshold was set at 0.01. Since there are many enriched gene sets that are highly overlapping (sharing common genes), we use a cluster-and-filter strategy to reduce the enriched gene sets. In the cluster-step, we first calculate the relatedness between all gene sets (defined as the number of overlapped genes/the number of union genes) and form a gene set-based network. Then, we use iNP to partition the network into modules within which gene sets are highly overlapping. In the filter step, after the enrichment analysis, we map all enriched gene sets to gene set modules and then select the most significantly enriched gene set within each module as the representative gene sets. Finally, we collect all representative gene sets of each enriched module to form a reduced list of enriched gene sets.

2.4. Detection of Genes with Significantly Different Expression Pattern between the Stages and Subtypes of NSCLC. The NSCLC datasets provide the stage and subtype information for each sample. By grouping samples according to stages or subtypes, we perform ANOVA test to determine whether a given gene's expression value is significantly different in at least one stage or subtype of NSCLC. We set P value threshold as 0.05 with fdr adjustment.

3. Results

3.1. Construction of a Lung Cancer-Specific Functional Association Network. We prepare three gene expression datasets for normal lung tissues (1,349 samples), SCLC (90 samples), and NSCLC (275 samples), respectively, from GEO database (Table 1). For each dataset, we calculate Pearson Correlation Coefficient (PCC) for each pair of genes using their expression profiles in the datasets. PCC ranges from −1 to 1, with 1 indicating positive correlation and −1 indicating negative correlation. After sorting all gene pairs according to their PCC values, from each of the three datasets we select the top 1% and the bottom 1% gene pairs as the positively and negatively coexpressed gene pairs. Then, we combine all gene pairs selected from the three datasets and map them to the STRING network [16], a functional association network. By retaining those gene pairs that have a functional association score greater than 500 (the score indicates a strong functional association between the pair of genes), we obtain a lung cancer-specific gene functional association network, which is a binary network, consisting of 7,572 genes and 43,816 edges.

3.2. Identification of Differentially Expressed Gene Modules in Lung Cancer. Using a network partition algorithm called iNP [20], we partition the lung cancer-specific gene functional association network into 737 modules with a modularity of 0.53. Genes within each module are highly functionally associated with each other. To determine whether a gene module is differentially expressed between cancer and normal samples, we first determine the expression value of the module using the median expression value of genes inside the module in each sample. Then, similar to determining differentially expressed genes, we use *t*.test to identify modules that are differentially expressed between SCLC or NSCLC samples and normal lung tissues samples. We find a total number of 71 modules that are significantly differentiated expressed between lung cancer and normal tissues (Figure 1). Of these modules, 23 and 28 are significantly up- and downregulated in SCLC, respectively. As for NSCLC, there are 18 upregulated and 18 downregulated modules. In addition, there are 8 modules that are upregulated in both SCLC and NSCLC, and 8 modules that are downregulated in both SCLC and NSCLC.

3.3. Functional Analysis of Differentially Expressed Gene Modules in Lung Cancer. To obtain functional insights about the differentially expressed gene modules, we conduct pathway enrichment analysis for genes inside those modules.

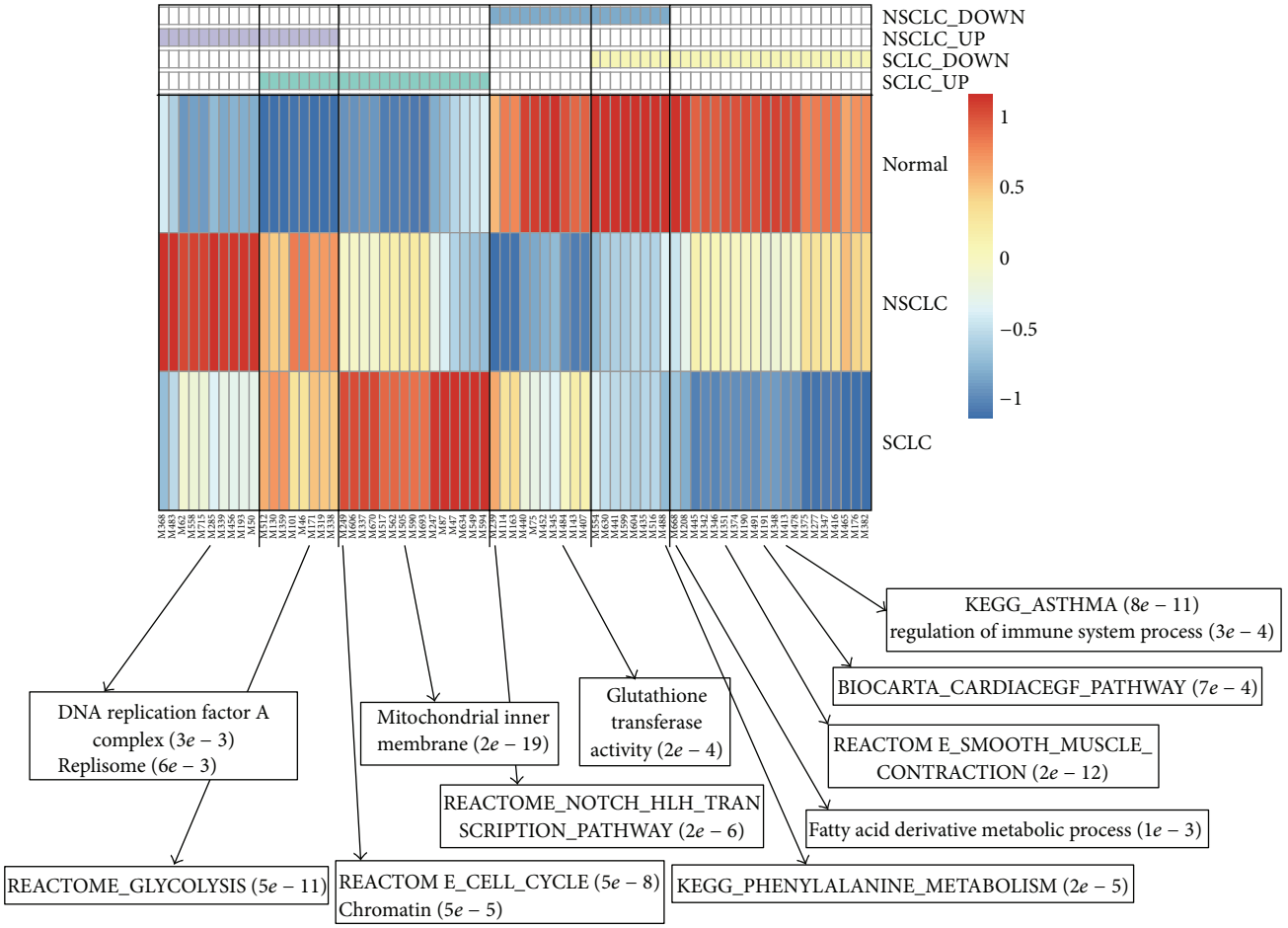


Figure 1: Heatmap of gene modules that have differential expression pattern between SCLC or NSCLC and normal lung tissue samples. Each column represents a module. The top four rows represent the types of differential expression pattern. For example, SCLC_up means the module is upregulated in SCLC, compared to normal tissues. The most significantly enriched functions of selected gene modules are shown with P values at the bottom of the heatmap.

The full list of the module ID, genes inside the module, and enriched functions can be found in Supplemental Table 1 in Supplementary Material available online at http://dx.doi.org/10.1155/2015/685303. Here, we only provide a reduced list of enriched functions (see Methods for details) in order to reduce the redundancy among enriched functions. The most significantly enriched functions of selected gene modules are shown in Figure 1 for illustrative purpose. For gene modules that are upregulated in both SCLC and NSCLC, some of them are enriched with glycolysis metabolic process, a common hallmark for many types of cancers [24]. For the gene modules that are only upregulated in SCLC, we find that their functions are involved in mitochondrial, cell cycle, and chromatin organization. This is consistent with the phenotypic description of SCLC, which says that cancer cells of SCLC have finely granular nuclear chromatin, suggesting that chromatin organization may be disrupted. It has also been proposed that a therapeutic strategy for SCLC can be to target the mitochondrial apoptosis pathway [5]. For gene modules that are downregulated in only SCLC, they are enriched with diverse functions, such as asthma, regulation of immune system process, muscle contraction, and fatty acid metabolism. Though asthma is rarely related to SCLC, some cases have been reported [25]. Fatty acids in erythrocyte are treated as potential biomarkers in the diagnosis of lung cancer [26]. For gene modules that are upregulated in only NSCLC, some are enriched with function involved in DNA replication, which is consistent with previous findings that genes DNA repair are closely related to the risk of NSCLC [27]. For gene modules downregulated only in NSCLC, some of them are enriched with Notch pathway and glutathione transferase activity, both of which have been reported to be strongly related to lung cancer [28, 29]. Thus, functional analysis of differentially expressed genes supports that these gene modules are strongly related to lung cancer development, making it possible for us to conduct further exploitation to identify gene biomarkers that can be of diagnosis use for distinguishing the two subtypes of lung cancer.

3.4. Identification of Potential Diagnosis Biomarkers that Distinguish SCLC and NSCLC. Given the 71 gene modules that are significantly differentially expressed between lung

cancer and normal lung tissues, we seek to find those that are specific to either SCLC or NSCLC lung cancer subtype. To do this, we use the module's expression value (the median expression value of all genes inside the module) to classify SCLC and NSCLC cancer samples and plot an ROC curve for each module. The AUC (area under curve) of an ROC ranges from 0 to 1, with 0.5 indicating randomness. AUC of 0 or 1 corresponds to perfect specific in distinguishing the two lung cancer subtypes. In addition, as we try to classify SCLC against NSCLC, an AUC close to 1.0 indicates that the module is upregulated in SCLC compared to that in NSCLC, and if it is close to 0.0, then the module is downregulated in SCLC. We find 10 modules with AUC score greater than 0.9, indicating that these modules are strongly upregulated in SCLC compared to NSCLC. There are also 16 modules with scores smaller than 0.1, suggesting that they are strongly downregulated in SCLC.

After identifying gene modules that have distinct expression pattern between SCLC and NSCLC, we further seek to find genes that are of diagnostic use for distinguishing SCLC and NSCLC. For this purpose, we develop a score function to measure the ability to distinguish the two subtypes of lung cancer for each gene. This score integrates both the cancer type specificity of the module and the coexpression value difference of the gene. The cancer type specificity of a module is simply defined by the absolute value of AUC-0.5. The coexpression value difference of a gene in a given module is determined by the average of the coexpression value difference of the interactions involving this gene in this module. The coexpression value difference for a given interaction is defined by the difference in coexpression pattern of the two interacting genes between SCLC and NSCLC. For example, for two interacting genes A and B, if they are positively coexpressed in SCLC while negatively coexpressed in NSCLC (see Section 2 for the definition of positively or negatively coexpressed gene pairs), then the coexpression value difference will be $1 - (-1) = 2$ and vice versa. However, if the two interacting genes are positively or negatively coexpressed in one cancer subtype while randomly coexpressed in another one, then the coexpression value will be $1 - 0 = 1$. If both genes are either positively or negatively coexpressed in both cancer types, then the coexpression value difference will be 0. By computing the coexpression value difference of all interactions in the module that involve the gene under test, we can compute the average and use it to represent the coexpression value difference for this gene. Finally, the coexpression value of the gene is multiplied by the cancer type specificity of the module to produce a score for the gene.

Following the above-described strategy, we have obtained 137 genes with significant power in distinguishing SCLC from NSCLC and show the top 10 upregulated and top 10 downregulated genes in SCLC as candidate genes of diagnostic use in Figure 2. Three upregulated genes in SCLC are from one module named M505 (Figure 3(a)). This module is enriched with mitochondrial related functions. The AUC score for this module is 0.957, indicating that this module is significantly upregulated in SCLC compared to NSCLC. One of the three genes is UQCRB that has the highest score. UQCRB interacts

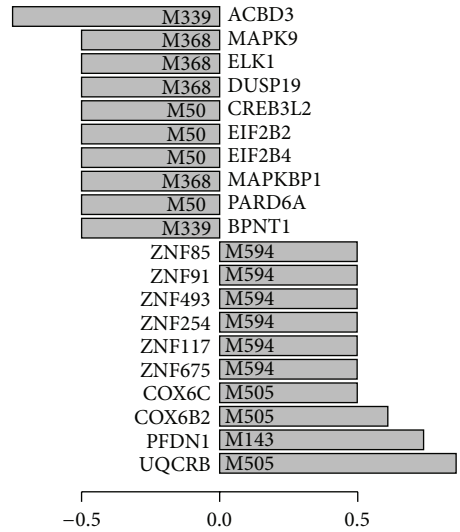


Figure 2: Bar plot for the top 10 upregulated and top 10 downregulated genes in SCLC, compared to that in NSCLC. The bar height of each gene represents the predicted score for the gene that distinguishes SCLC samples from NSCLC samples (see Section 2 for details). Genes with score greater than 0 are upregulated genes, while those with score smaller than 0 are downregulated genes. The IDs of the module where the predicted genes belong are shown inside the bar.

with 18 genes in this module. Among the 18 interactions involving UQCRB, 16 are negatively coexpressed in SCLC while 14 are positively coexpressed in NSCLC. UQCRB is ubiquinol-cytochrome c reductase binding protein and is strongly upregulated in SCLC. It is likely that this gene may be important for the disrupted mitochondrial functions in SCLC. Interestingly, we find that there is a SNP located on UQCRB, rs7827095, which is strongly related to lung disease according to a cohort study [30]. As for genes downregulated in SCLC, ACBD3 has the highest score. ACBD3 is present in a module named M339 that is strongly downregulated in SCLC with an AUC score of 0.000152 (Figure 3(b)). ACBD3 interacts with two genes in this module, which are BPNT1 and BLZF1. Interestingly, ACBD3 and BPNT1 are negatively coexpressed in SCLC, while ACBD3 and BLZF1 are negatively coexpressed in NSCLC. ACBD3 is acyl-CoA binding domain containing 3 protein and is involved in hormone regulation of steroid formation. This gene has been reported to show significant difference between responders (PR) and nonresponders (PD) to gefitinib in NSCLC treatment [31]. Here, our study suggests that ACBD3 is worthy of further exploitation to understand its involvement in the development of NSCLC. The full list of predicted genes is shown in Supplementary Table 1.

3.5. Further Validation of the Proposed Method. To further validate our method, we select a new NSCLC dataset (GSE10245) which has 58 samples (the original dataset includes more than 100 samples) and then repeat the analysis to predict genes that can distinguish SCLC from NSCLC. We obtain 101 genes in which 15 overlaps with the original

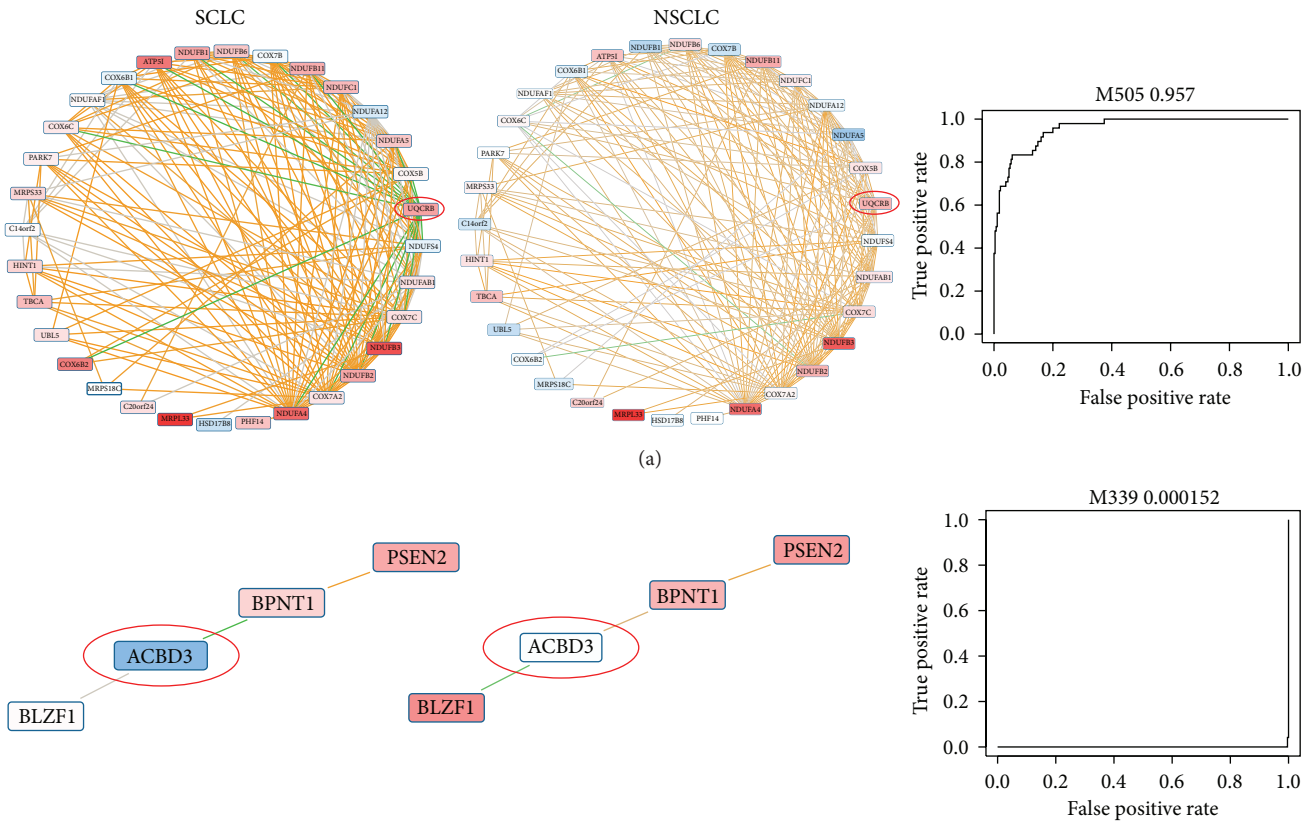


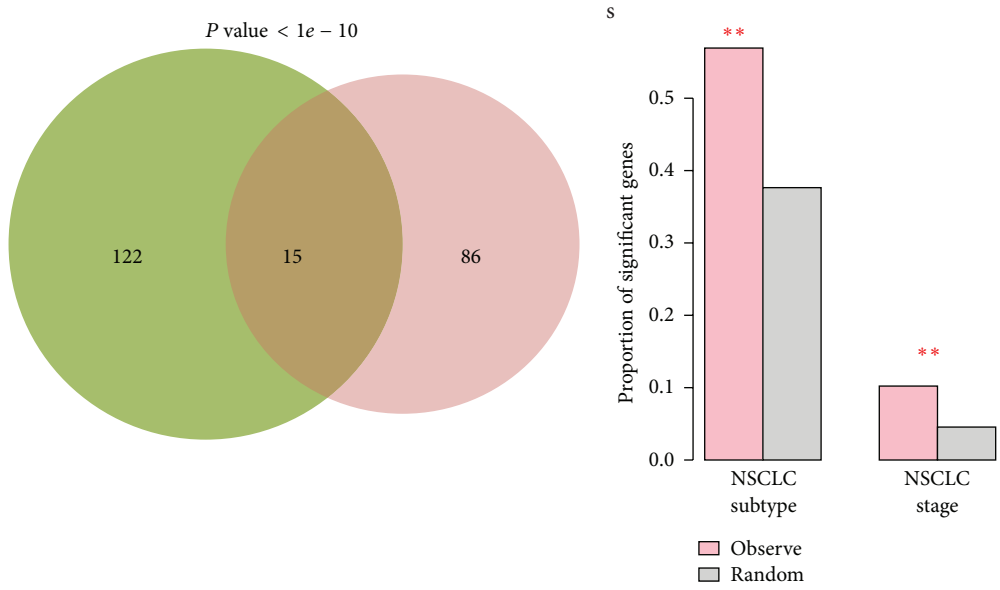
FIGURE 3: Examples of gene modules and genes that can distinguish SCLC from NSCLC. (a) shows a module that is upregulated in SCLC, while (b) shows a module with only four genes that is downregulated in SCLC. The ROC curve of the two modules is shown at the right of the figure. The color of each gene inside the module represents the expression pattern of this gene between a cancer subtype and the normal lung tissues, with red indicating that it is upregulated in cancer, while blue indicating that it is downregulated. The edge color of each interaction inside the module represents the coexpression information in the corresponding cancer subtypes, with orange indicating positive coexpression while green indicating negative coexpression.

predictions (Figure 4(a)). The overlap is significantly higher than random (P value $< 1e-10$), indicating that our method is robust to the choice of datasets.

In addition, we inspect whether the predicted genes based on the original datasets can be used to distinguish different stages of lung cancer or the subtypes within either SCLC or NSCLC. The NSCLC dataset actually includes not only stage information, but also the subtypes of NSCLC information. The stage information is 48 IA, 84 IB, 11 IIA, 39 IIB, 51 IIIA, 35 IIIB, and 7 other stages. In this dataset, there are 14 subtypes of NSCLC, including 183 adenocarcinomas, 80 squamous, and 12 other subtypes of NSCLC. Among the 137 predicted genes, we find 78 genes that can distinguish adenocarcinoma from squamous samples based on their expression levels and 14 genes that have significant gene expression variation in at least one stage of NSCLC (see Section 2 for details of the test procedure). The proportion of those genes among the predicted genes is significantly higher than that of randomly selected genes (both P values < 0.01) (Figure 4(b)). These results suggest that a significant proportion of our predicted genes may also carry information to distinguish between different stages and subtypes of NSCLC.

4. Discussion

Lung cancer is the leading cause for cancer related death worldwide. It can be classified into two main subtypes (SCLC and NSCLC) according to their physiological phenotypes. In this study, we have conducted a computational study to predict molecular biomarkers that are of potential diagnostic use for distinguishing the two types of lung cancer. By collecting gene pairs that are significantly positively or negatively coexpressed in normal lung tissues, SCLC samples, and NSCLC samples and considering their functional associations, we have constructed a lung cancer-specific gene functional association network. After partitioning the network into gene modules, we identify gene modules that have significant expression variation between lung cancer and normal lung tissues. We then compute the specificity of each of these modules in distinguishing SCLC from NSCLC and further identify candidate genes inside the module that



(a) (b)

FIGURE 4: (a) Venn diagram to show the overlap between the genes predicted by using the new NSCLC dataset and those predicted based on the original three datasets. (b) Proportion of genes among the predicted genes that are significant in distinguishing between different subtypes and stages of NSCLC and that are among randomly selected genes. "∗∗" indicate that the observed proportion is significantly higher than random.

have discriminating power for lung cancer subtypes. There are several interesting findings that resulted from this study. First of all, we find that gene modules that are upregulated in both lung cancer subtypes are significantly enriched with glycolysis metabolism, while those downregulated gene modules shared for the two lung cancer subtypes are enriched with phenylalanine metabolism, suggesting that there are common altered metabolisms in the development of both SCLC and NSCLC. Secondly, we find gene modules that are differentially expressed only in one subtype of lung cancer and are enriched with specific functions. For example, gene modules upregulated only in SCLC are enriched with functions in chromatin modeling, while those upregulated in NSCLC are enriched with DNA replication process, suggesting that there may be unique mechanism for the development of specific lung cancer subtypes. Thirdly, we have obtained a list of genes that have significant discriminating power in distinguishing lung cancer subtypes. These genes are identified by integrating both cancer-specificity information and coexpression information and provide novel hypothesis for the development of specific subtype of lung cancer. Finally, although the study is designed for identifying molecular biomarker for distinguishing lung cancer subtypes, the methodology developed here can be readily applied for distinguishing the subtypes of other types of diseases.

Conflict of Interests

The authors declare that they have no conflict of interests.

References

[1] W. Chen, S. Zhang, and X. Zou, "Estimation and projection of lung cancer incidence and mortality in China," *Zhongguo Fei Ai Za Zhi*, vol. 13, no. 5, pp. 488–493, 2010.

[2] B. L. P. Boyle, *World Cancer Report 2008*, IARC Press, Paris, France, 2008.

[3] P. C. Hoffman, A. M. Mauer, and E. E. Vokes, "Lung cancer," *The Lancet*, vol. 355, no. 9202, pp. 479–485, 2000.

[4] W. D. Travis, E. Brambilla, H. K. Müller-Hermelink, and C. C. Harris, Eds., *World Health Organization Classification of Tumours: Pathology and Genetics: Tumours of the Lung, Pleura, Thymus and Heart*, vol. 10, IARC Press, Lyon, France, 2004.

[5] J. P. van Meerbeeck, D. A. Fennell, and D. K. M. De Ruysscher, "Small-cell lung cancer," *The Lancet*, vol. 378, no. 9804, pp. 1741–1755, 2011.

[6] I. I. Wistuba, "Genetics of preneoplasia: lessons from lung cancer," *Current Molecular Medicine*, vol. 7, no. 1, pp. 3–14, 2007.

[7] E. Thunnissen, K. M. Kerr, F. J. F. Herth et al., "The challenge of NSCLC diagnosis and predictive analysis on small samples. Practical approach of a working group," *Lung Cancer*, vol. 76, no. 1, pp. 1–18, 2012.

[8] S. L. Edwards, C. Roberts, M. E. McKean, J. S. Cockburn, R. R. Jeffrey, and K. M. Kerr, "Preoperative histological classification of primary lung cancer: accuracy of diagnosis and use of the non-small cell category," *Journal of Clinical Pathology*, vol. 53, no. 7, pp. 537–540, 2000.

[9] S. K. Arya and S. Bhansali, "Lung cancer and its early detection using biomarker-based biosensors," *Chemical Reviews*, vol. 111, no. 11, pp. 6783–6809, 2011.

[10] H.-H. Zhang, Z.-Y. Zhang, C.-L. Che, Y.-F. Mei, and Y.-Z. Shi, "Array analysis for potential biomarker of gemcitabine identification in non-small cell lung cancer cell lines," *International Journal of Clinical and Experimental Pathology*, vol. 6, no. 9, pp. 1734–1746, 2013.

[11] D. K. Mishra, C. J. Creighton, Y. Zhang, D. L. Gibbons, J. M. Kurie, and M. P. Kim, "Gene expression profile of A549 cells from tissue of 4D model predicts poor prognosis in lung cancer patients," *International Journal of Cancer*, vol. 134, no. 4, pp. 789–798, 2014.

[12] S. Karni, H. Soreq, and R. Sharan, "A network-based method for predicting disease-causing genes," *Journal of Computational Biology*, vol. 16, no. 2, pp. 181–189, 2009.

[13] Y. Bossé, D. S. Postma, D. D. Sin et al., "Molecular signature of smoking in human lung tissues," *Cancer Research*, vol. 72, no. 15, pp. 3753–3763, 2012.

[14] M. Sato, J. E. Larsen, W. Lee et al., "Human lung epithelial cells progressed to malignancy through specific oncogenic manipulations," *Molecular Cancer Research*, vol. 11, no. 6, pp. 638–650, 2013.

[15] T. Sato, A. Kaneda, S. Tsuji et al., "PRC2 overexpression and PRC2-target gene repression relating to poorer prognosis in small cell lung cancer," *Scientific Reports*, vol. 3, article 1911, 2013.

[16] D. Szklarczyk, A. Franceschini, M. Kuhn et al., "The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored," *Nucleic Acids Research*, vol. 39, supplement 1, pp. D561–D568, 2011.

[17] S. Davis and P. S. Meltzer, "GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor," *Bioinformatics*, vol. 23, no. 14, pp. 1846–1847, 2007.

[18] T. Barrett, D. B. Troup, S. E. Wilhite et al., "NCBI GEO: mining tens of millions of expression profiles—database and tools update," *Nucleic Acids Research*, vol. 35, supplement 1, pp. D760–D765, 2007.

[19] G. K. Smyth, "Limma: linear models for microarray data," in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, pp. 397–420, Springer, New York, NY, USA, 2005.

[20] S. Sun, X. Dong, Y. Fu, and W. Tian, "An iterative network partition algorithm for accurate identification of dense network modules," *Nucleic Acids Research*, vol. 40, no. 3, p. e18, 2011.

[21] R. Kuner, T. Muley, M. Meister et al., "Global gene expression analysis reveals specific patterns of cell junctions in non-small cell lung cancer subtypes," *Lung Cancer*, vol. 63, no. 1, pp. 32–38, 2009.

[22] M. Ashburner, C. A. Ball, J. A. Blake et al., "Gene ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[23] A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, and J. P. Mesirov, "Molecular signatures database (MSigDB) 3.0," *Bioinformatics*, vol. 27, no. 12, pp. 1739–1740, 2011.

[24] R. A. Gatenby and R. J. Gillies, "Why do cancers have high aerobic glycolysis?" *Nature Reviews Cancer*, vol. 4, no. 11, pp. 891–899, 2004.

[25] M. S. Dionisi and S. Rubino, "Asthma associated with small-cell lung cancer," *Supportive Care in Cancer*, vol. 3, no. 5, pp. 317–318, 1995.

[26] J. de Castro, M. C. Rodríguez, V. S. Martínez-Zorzano, P. Sánchez-Rodríguez, and J. Sánchez-Yagüe, "Erythrocyte Fatty acids as potential biomarkers in the diagnosis of advanced lung adenocarcinoma, lung squamous cell carcinoma, and small cell lung cancer," *American Journal of Clinical Pathology*, vol. 142, no. 1, pp. 111–120, 2014.

[27] S. Zienolddiny, D. Campa, H. Lind et al., "Polymorphisms of DNA repair genes and risk of non-small cell lung cancer," *Carcinogenesis*, vol. 27, no. 3, pp. 560–567, 2006.

[28] P. Yang, W. R. Bamlet, J. O. Ebbert, W. R. Taylor, and M. de Andrade, "Glutathione pathway genes and lung cancer risk in young and old populations," *Carcinogenesis*, vol. 25, no. 10, pp. 1935–1944, 2004.

[29] P. Galluzzo and M. Bocchetta, "Notch signaling in lung cancer," *Expert Review of Anticancer Therapy*, vol. 11, no. 4, pp. 533–540, 2011.

[30] B. A. Ong, J. Li, J. M. McDonough et al., "Gene network analysis in a pediatric cohort identifies novel lung function genes," *PLoS ONE*, vol. 8, no. 9, Article ID e72899, 2013.

[31] J. Fan, J. Liu, M. Culty, and V. Papadopoulos, "Acyl-coenzyme A binding domain containing 3 (ACBD3; PAP7; GCP60): an emerging signaling molecule," *Progress in Lipid Research*, vol. 49, no. 3, pp. 218–234, 2010.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 619730, 9 pages
http://dx.doi.org/10.1155/2015/619730

Research Article

Network-Based Association Study of Obesity and Type 2 Diabetes with Gene Expression Profiles

Siyi Zhang,[1] Bo Wang,[1] Jingsong Shi,[2] and Jing Li[1,3]

[1] *Department of Bioinformatics & Biostatistics, School of Life Science and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China*
[2] *National Clinical Research Center of Kidney Disease, Jinling Hospital, Nanjing University School of Medicine, Nanjing 210016, China*
[3] *Shanghai Center for Bioinformation Technology, Shanghai 201203, China*

Correspondence should be addressed to Jing Li; jing.li@sjtu.edu.cn

Received 29 December 2014; Accepted 5 February 2015

Academic Editor: Xiao Chang

The increased prevalence of obesity and type 2 diabetes (T2D) has become an important factor affecting the health of the human. Obesity is commonly considered as a major risk factor for the development of T2D. However, the molecular mechanisms of the disease relations are not well discovered yet. In this study, the combination of multiple differential expression profiles and a comprehensive biological network of obesity and T2D allowed us to identify and compare the disease-responsive active modules and subclusters. The results demonstrated that the connection between obesity and T2D mainly relied on several pathways involved in the digestive metabolism, immunization, and signal transduction, such as adipocytokine, chemokine signaling pathway, T cell receptor signaling pathway, and MAPK signaling pathways. The relationships of almost all of these pathways with obesity and T2D have been verified by the previous reports individually. We also found that the different parts in the same pathway are activated in obesity and T2D. The association of cancer, obesity, and T2D was identified too here. As a conclusion, our network-based method not only gives better support for the close connection between obesity and T2D, but also provides a systemic view in understanding the molecular functions underneath the links. It should be helpful in the development of new therapies for obesity, T2D, and the associated diseases.

1. Introduction

A sedentary life-style coupled with calorie-dense dietary behavior of contemporary human causes the accumulation of body fat. In the past decades, the prevalence of obesity increased rapidly in industrialized societies with its undesirable consequences such as type 2 diabetes (T2D), high blood pressure, and heart diseases [1]. The latest National Health and Nutrition Examination Survey (NHANES) program estimated that the prevalence of obesity (defined as a body mass index greater than 30) in adults has reached 36% in the United State [2], while the global incidence of diabetes mellitus is expected to increase to 366 million cases by the year 2030 [3].

Obesity is commonly considered as a major risk factor for the development of T2D. It has been reported that the altered glucose and lipid metabolism in liver, skeletal muscles, and adipose tissues with the disorganized insulin signals lead to the systemic and chronic inflammation [4, 5]. They also recognized that the obesity-caused metabolic inflammation could connect obesity to the insulin resistance (IR), which is associated with T2D [6, 7]. And very few disease genes have been reported in both obesity and diabetes, such as PPARG [8, 9] and UCP3 [10]. However, the molecular mechanism in the association between obesity and diabetes is still far from being fully understood.

Recently biological network and high-throughput gene expression data are emerging as useful resources in revealing the molecular mechanisms of complex disease [11–14]. In this study, using genome-scale gene differential expression profiles and an integrated biological network of obesity and T2D, which contained the information in protein-protein interactions, transcriptional regulation, and metabolic pathways, we identified and compared the gene network and the active subnetworks in pathology between obesity and T2D in

order to provide novel insight to understand the molecular association between them.

2. Materials and Methods

2.1. Disease Genes and Gene Links. 26 obesity genes and 34 T2D genes were collected from the Online Mendelian Inheritance in Man (OMIM) [15] as the seed genes. Three seed genes were common between obesity and T2D. The experimentally validated protein-protein interactions and transcriptional regulation of these seed genes and their neighbors were extracted from the human protein interaction database HPRD [16] (Release 9) and TRANSFAC database [17] (Release 2013.2), as well as from 29 KEGG pathways [18] (Release 71.1, September 1, 2014) enriched by the known obesity and T2D genes. In the interaction file downloaded from KEGG, only the PPrel (protein-protein interaction) and GErel (gene expression interaction) were extracted and added to this study.

2.2. Gene Expression Profiles and Processing. We collected twelve microarray datasets totally in case-control design from the NCBI Gene Expression Omnibus (GEO) [19] for obesity (GSE10946, GSE15653, GSE29718, GSE48964, GSE9624) and T2D (GSE18732, GSE13760, GSE20966, GSE23343, GSE25724, GSE38396, and GSE38642). All of these datasets were curated and reported in the GEO Datasets (GDS). Each dataset was required to have at least three samples for both case and control groups. And the samples from these patients who suffered both obesity and diabetes were excluded.

The preprocessing of microarray data was conducted by the RMA [20–22] integrative method, and the statistical analysis of gene differential expression was computed by the linear models and empirical Bayes methods [23]. And then the P values of each gene were obtained.

2.3. Identification of Active Modules and Subclusters. From the gene network of obesity and T2D, we used the jActiveModules [24] and multiple gene expression profiles to find the active gene modules showing significant changes in expression in disease/normal conditions. The jActiveModules (Version 1.8) is a widely used method for identifying active modules integrating multiple gene differential expression datasets. In the algorithm of jActiveModules [24], the P values of each gene in a subnetwork in a single condition are transformed into one standard normal z-score by the binomial order statistic. The highest score obtained in multiple experiments is recorded as the final score for a subnetwork. Higher z-score represents more significant expression changes. Here the top 5 scoring modules of obesity and T2D were enumerated separately by jActiveModules with default parameters in Cytoscape [25].

In order to further identify the subclusters with tight topology structures, we decomposed the active modules and the disease seed genes into several subclusters. As a result, ten and seven subclusters were identified by the MCODE method [26] for obesity and T2D, respectively. The workflow was illustrated in Figure 1.

3. Results and Discussion

3.1. The NOT2D Network. Through collecting the protein interactions and transcriptional regulation data of the known genes of human obesity and T2D and their interacting neighbors from HPRD, TRANSFAC, and KEGG pathways, we compiled a multi-level biological network of human obesity and T2D called NOT2D (gene network of obesity and type 2 diabetes) (Figure 1). As shown in Figure 2, the majority of links in the obesity network were obtained from KEGG database while HPRD and KEGG databases contributed almost equally to the T2D network. Very few interactions are reported by two or more data sources. Finally, there are 606 nodes and 2907 edges in the obesity network and 1211 nodes and 4089 edges in the network of T2D. Among 7170 unique edges in the NOT2D network, there are 6229 protein-protein interactions and 941 gene regulatory links. 374 out of 1443 nodes in the NOT2D network are shared by obesity and T2D. Surprisingly the average degree of the shared genes is 19.2, which is more than twice the average value of 9.5 in the whole NOT2D network.

The interaction data of the NOT2D network can be downloaded at http://lilab.life.sjtu.edu.cn:8080/NOT2D/.

3.2. Topology and Function of the Active Subclusters. By combing the differential expression profiles from multiple datasets and the NOT2D network, the top 5 scoring active modules were identified in the obesity samples as well as in T2D. And then the top 5 active modules and the seed genes were merged into an active network for both obesity and T2D.

To better understand the biological processes or molecular function underneath the active gene network of obesity or T2D, we decomposed the active networks into 10 obesity clusters and 7 T2D functional clusters by the MCODE method, of which from 3 to 43 genes were contained. The topology structures of these clusters were displayed in Figure 3.

As an example, a seed gene IL6 in the obesity cluster 1 is regulated by JUN and FOS, which is secreted by M1 macrophages, and often takes effect in promoting obesity-associated inflammation which aggravates the progression of metabolic complications, such as cardiovascular disease and insulin resistance [27]. Specially, as a member of lipid-sensing peroxisome proliferator-activated receptor (PPAR) family, PPAR-γ in obesity cluster 2 is a common disease gene for both obesity and T2D, which is a master regulator in adipocyte differentiation and whole-body insulin sensitivity [28, 29], while in T2D cluster 1, the insulin receptor (INSR) and insulin receptor substrate-1/2 (IRS1/2) were identified many years ago as key factors for insulin pathways to keep the carbohydrate homeostasis [30–32]. In addition, the proopiomelanocortin (POMC) and the agouti related protein homolog (AGRP) in T2D cluster 7 play a vital role in the balance of food intake and energy expenditure, through the generated neuronal and hormonal signals [33–35].

It can be found from Figure 3 that some active clusters did not contain any seed genes of obesity or/and T2D, such as obesity clusters 3, 5, 7, and 9 and T2D clusters 2, 4, 5, and 6. We inferred that most of these active clusters would be involved in the important processes related to obesity or

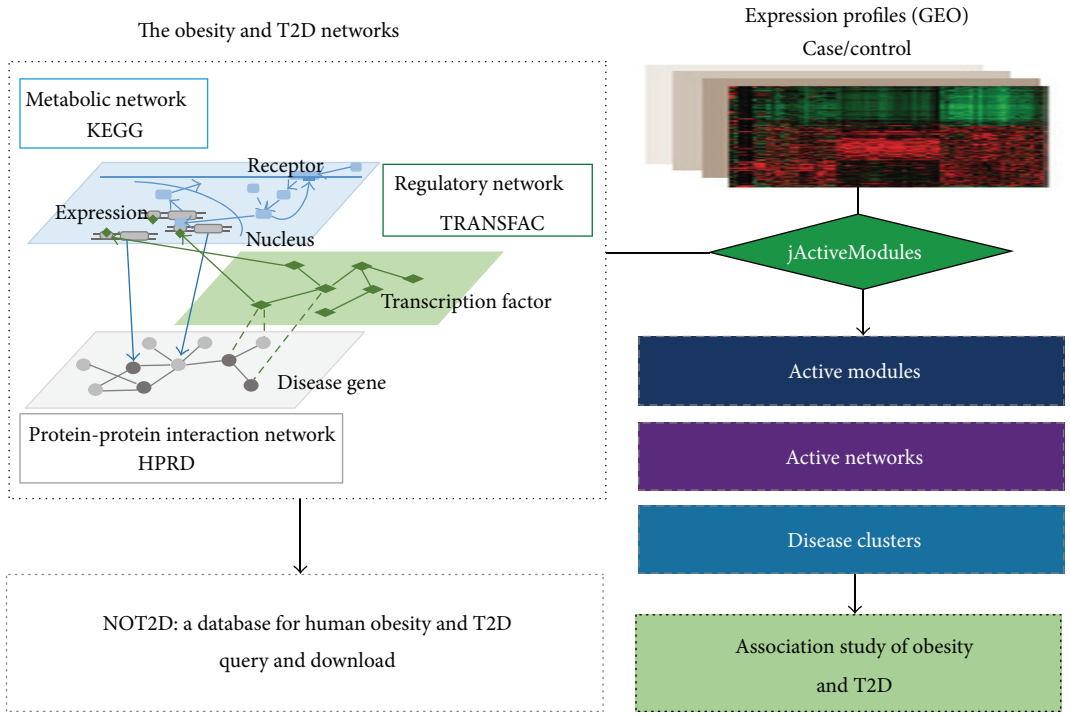


Figure 1: The identification of active modules and subclusters in human obesity and T2D. In the workflow, the disease genes of obesity and T2D were obtained from OMIM, and the interacting neighbors were collected from HPRD, KEGG, and TRANSFAC to construct a gene network of obesity and T2D (NOT2D). Multiple differential expression datasets in case/control design for obesity or T2D were integrated with the NOT2D network using jActiveModules method in order to identify the active modules and clusters. And finally a network association study of obesity and T2D was performed based on these active modules and clusters.

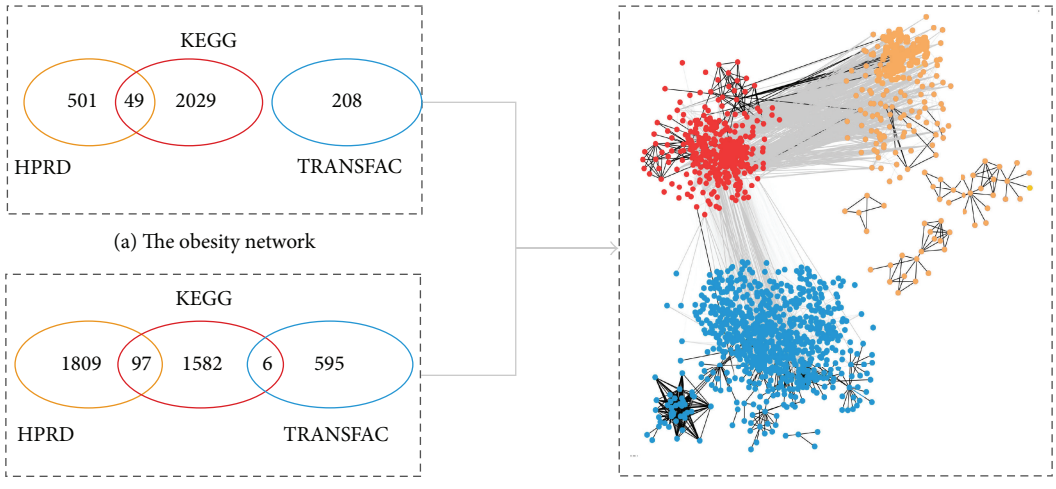


Figure 2: The edges and nodes distributions in the NOT2D network. (a) and (b) show the numbers of edges derived from the HPRD, KEGG, and TRANSFAC in the obesity and T2D networks. (c) displays the shared genes (red) and the specific genes of the obesity (orange) and T2D (blue) network.

T2D. In order to verify our point, the KEGG enrichment analysis was performed to all of the active subnetworks by the WebGestalt [36] (BH adjusted P value < 0.05).

As a result, there are 16 and 12 pathways significantly enriched in these obesity and T2D active clusters with or without the seed genes (Figure 4). It verified our conjecture very well as almost all these enriched pathways have been reported for their connection with the development of obesity and/or T2D, such as PPAR signaling pathway [9], insulin signaling pathway [37], and MAPK signaling pathway [38].

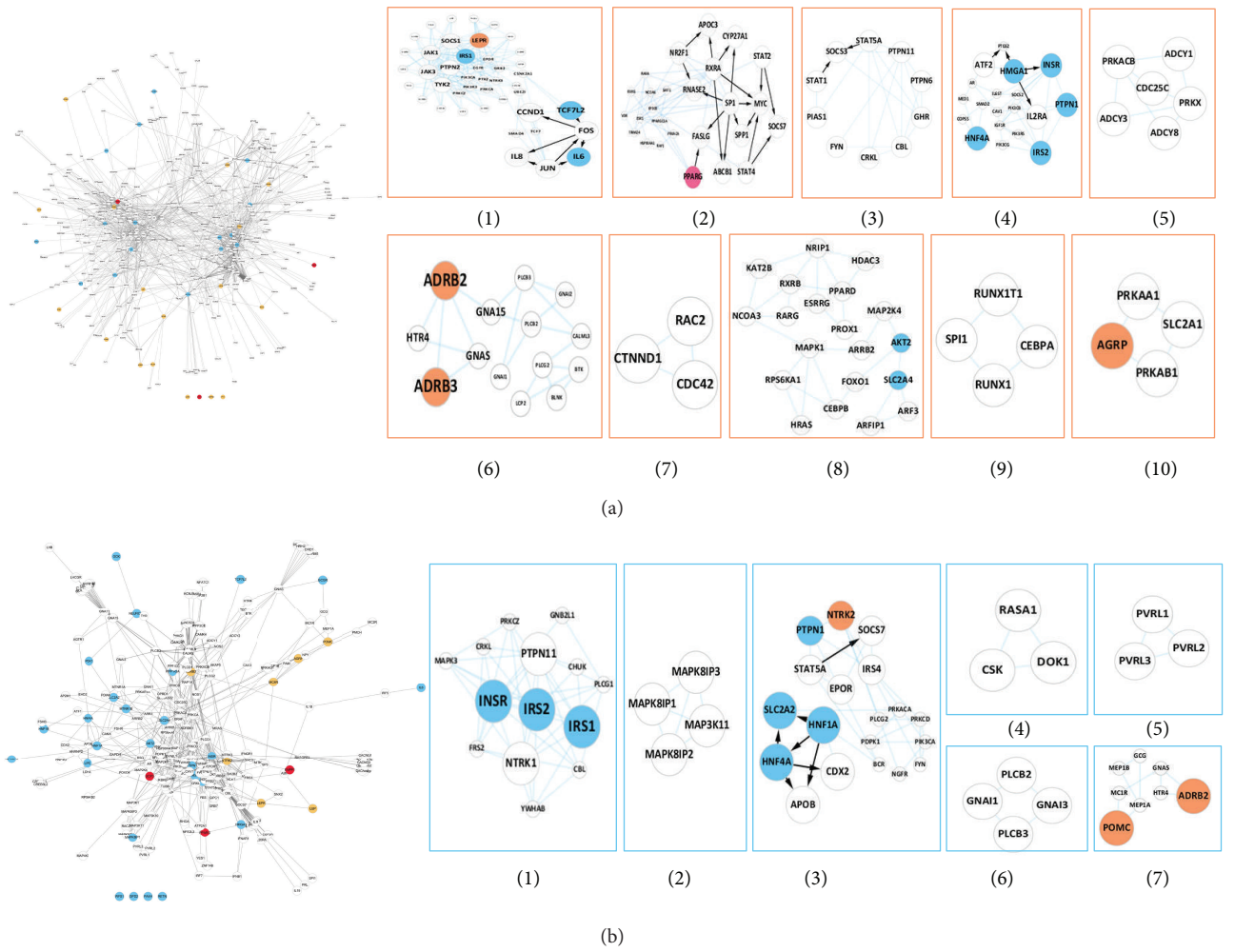


FIGURE 3: The obesity or T2D active networks and subclusters. (a) The active networks and ten subclusters of obesity. (b) The active networks and seven subclusters of T2D. The known seed genes of obesity (orange) and T2D (blue) are highlighted, as well as the shared genes (pink). The important genes in subclusters are also shown by bigger node sizes. The black arrows indicate the regulatory interactions while the blue links are protein interactions.

For instance, the PPAR signaling pathway is enriched in active cluster 2 of obesity, which has a vital function in adipocyte proliferation and differentiation in liver, muscle and adipose tissues [9]. The PPARs not only regulate lipid, carbohydrate, and amino acid metabolism, but also play an important role in systemic insulin sensitization through the combined effects of the production of adiponectin and reduction of lipotoxicity [9].

The insulin signaling pathway was enriched in obesity cluster 3 containing no seed gene [37]. Three genes (CRKL, CBL, and SOCS3) in this cluster paly roles in the Insulin signaling pathway, and SOCS3 gene has been reported for its inhibition of insulin signals of the adipose tissues [37]. The insulin signals have marked function of blood sugar reduction and improving sugar tolerance, which result in the development of obesity and T2D [38]. The obesity-associated insulin resistance is a major risk factor for type 2 diabetes and cardiovascular disease [39]. As a second example, all of the genes in T2D cluster 2 are the important members of MAPKs family which participates in the enriched MAPK signaling pathway. MAPK signaling pathway, which can be activated by insulin, is required for an array of metabolic events. The excessive activation of MAPKs is associated with detrimental effects on obesity and diabetes that contribute to disease progression [40]. These genes detected in the active clusters might be new candidate disease genes or biomarkers for obesity and T2D.

3.3. The Network Association between Obesity and T2D. There are lots of evidences demonstrating the strong connection between obesity and T2D. But so far only three genes (ENPP1, PPARG, and UCP3) are common in 26 obesity genes and 34 T2D genes annotated in the OMIM database. In order to explore the molecular association of obesity and T2D at the level of biological network, we constructed and compared the disease networks of obesity and T2D instead of the individual genes. As shown in Figure 2(b), the percentage of the shared nodes in the obesity and T2D networks is

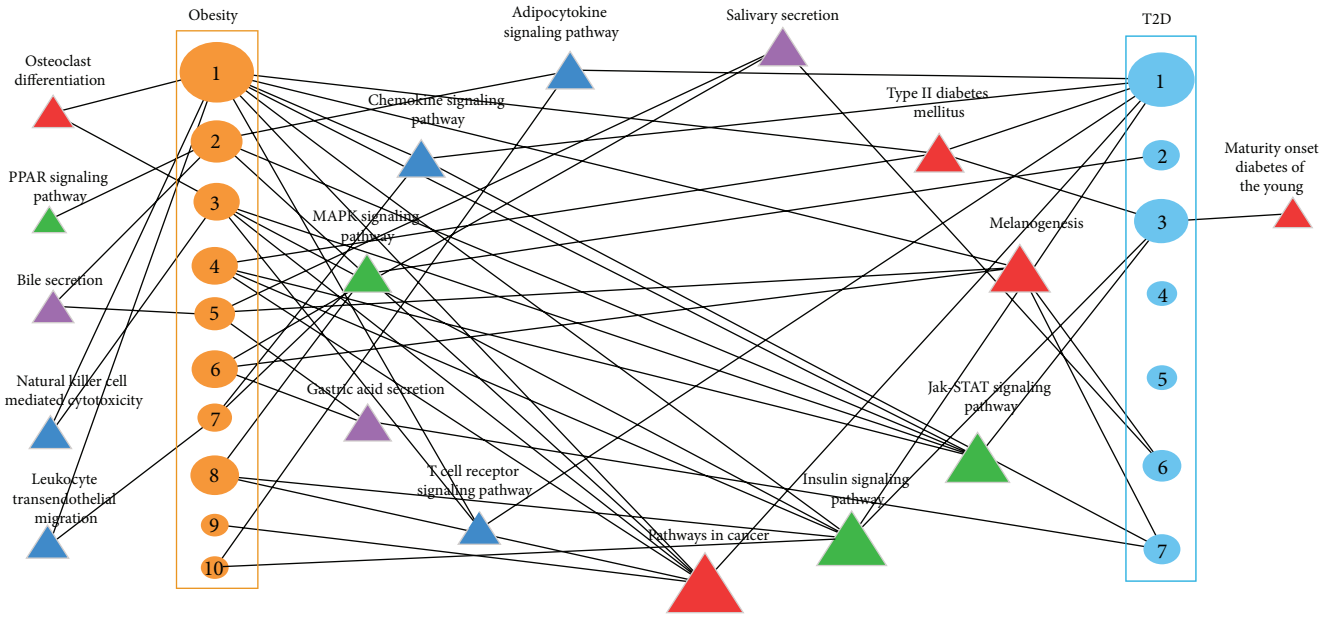


FIGURE 4: The KEGG pathways enriched in obesity and T2D clusters. The enriched pathways of the obesity (orange ellipse) and T2D (blue ellipse) clusters are classified into three regulatory groups (metabolic, immune response, and signaling) and one disease-related group, which were highlighted by the colored triangles (purple, blue, red, and grass green). The size of ellipse represents the number of genes in the subclusters and the triangle size is proportional to the number of the links with obesity and T2D subclusters.

increased dramatically to 26% while only 5% genes are shared at individual gene level. Additionally, we found in Figure 2(c) that the hub genes play critical roles in linking obesity and T2D since the average degree of the shared genes is significantly higher than the remains in the NOT2D network.

Whereafter, by applying the differential gene expression data in case/control design into the network, we identified the active gene networks and the subclusters of obesity and T2D. In the results, the node overlap of the active subclusters in obesity and T2D is very rare. However, the following functional analysis revealed that most of the pathways activated in obesity and T2D are the same (Figure 4). Eleven of the twelve activated pathways identified in obesity were also reported in T2D. Given an example, pathways in cancer, insulin signaling pathway, and other three pathways are enriched not only in obesity cluster 1 but also in T2D cluster 1, even though the gene overlap of these two clusters is very few (Figure 5). When we looked at more in insulin signaling pathway that regulating the whole glucose and lipid metabolism, and found that the activated parts of this pathway in obesity and T2D are different distinctly. Seven genes in the T2D cluster 1 and six genes in obesity cluster 1 are involved in insulin signaling pathway, but only two genes are the same. This result suggests that the association between obesity and T2D depends on the coactivated gene clusters or pathways rather than a few individual disease genes.

In general, the activated pathways connecting obesity and T2D mainly fall into four categories: digestive metabolism system, immune system, signaling transduction, and disease related pathways. Two digestive metabolism pathways, gastric acid secretion [41] and salivary secretion [42], are essential for digestion and absorption of protein, fats, and fat-soluble vitamins in the small intestine. In addition, the dysregulation of the energy metabolism may induce the accumulation of fats that lead to the obesity finally [43].

The responses of immune system also held a very important part in linking obesity and T2D, such as chemokine and T cell receptor signaling pathways. Previous studies reviewed that the nutrient and energy overload can induce the accumulation of adipose tissues and triggered the inflammatory cytokine expression; also the chronicity metabolic inflammation of fat cells could certainly change the energy intake and expenditure and insulin sensitivity states [7, 44]. Some chemokines are considered proinflammatory and can be induced during an immune response to recruit cells of the immune system to a site of infection [45]. The proinflammatory cytokine TNF alpha has been implicated as a link between obesity and insulin resistance [46].

Adipocytokines have been recently defined as soluble mediators derived mainly from adipocytes, in the interaction between adipose tissue, inflammation, and immunity. Thereby adipose tissue has been redefined as a key component not only of the endocrine system, but also of the immune system [47].

Some signal transduction pathways are also involved in the connection between obesity and T2D. The Jak-STAT signaling pathway is the principal signaling mechanism for a wide array of cytokines and growth factors, especially the critical role in regulating leukocyte maturation and activity [48], while MAPK signaling pathway is involved in cell proliferation, differentiation, and migration [49]. Obesity pathogenesis can be caused by mutations in the MC4R gene in MAPK signaling pathway [50].

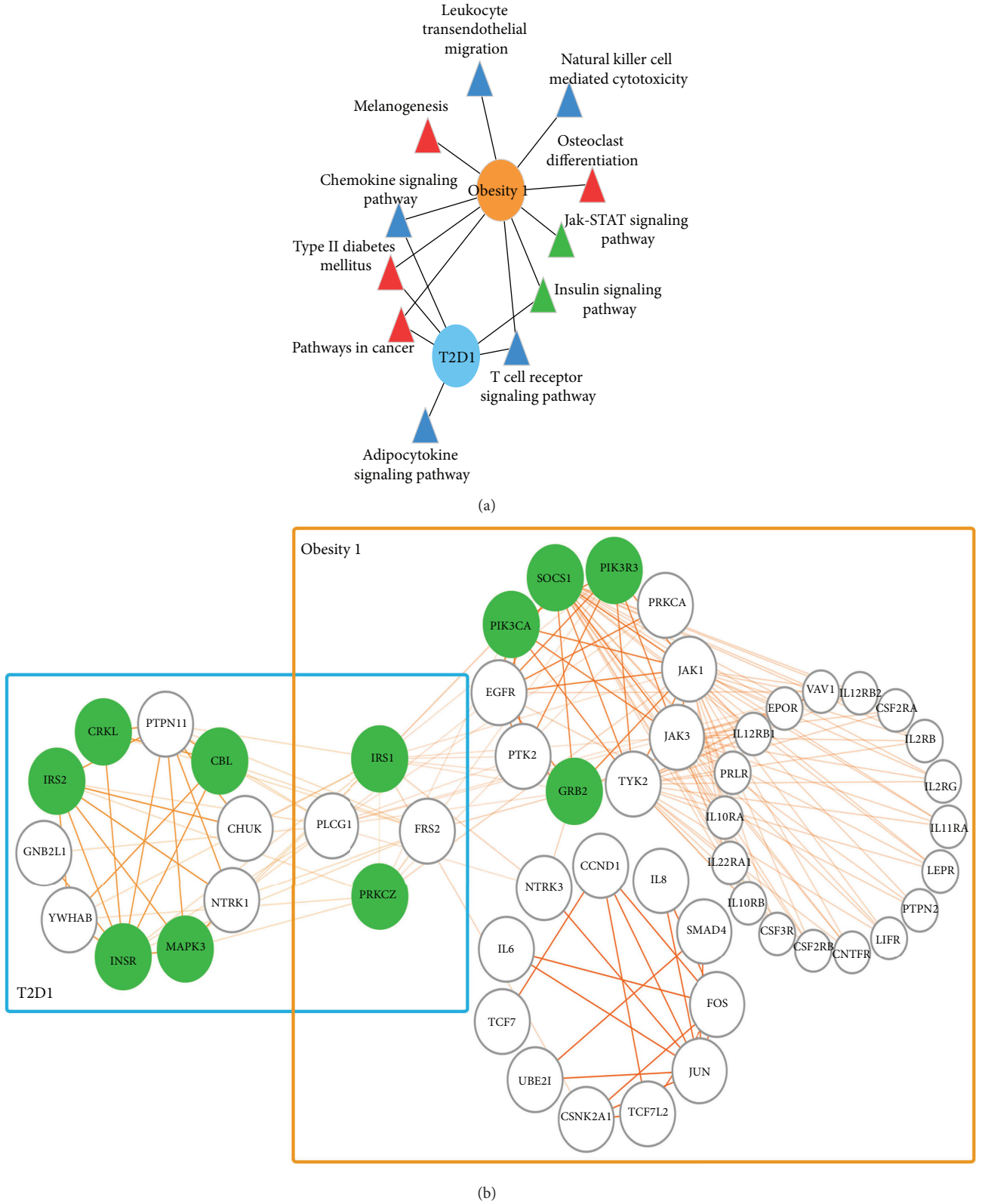


Figure 5: Functional connections brought by the active clusters between T2D and obesity. (a) The connecting pathways in the obesity (orange ellipse) and T2D (blue ellipse) cluster 1. The enriched pathways of the immune response, signaling, and disease-related groups were highlighted by the colored triangles (blue, red, and grass green). (b) The graph of T2D cluster 1 and obesity cluster 1. The genes involved in insulin signaling pathway were highlighted in green.

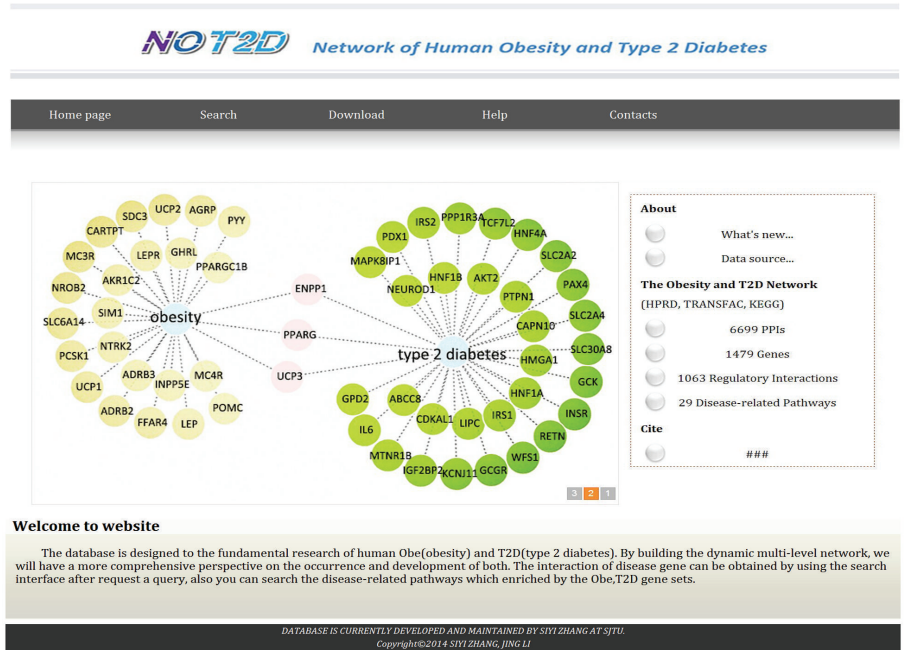


Figure 6: The snapshot of the NOT2D database.

Epidemiologic studies have indicated that diabetes and obesity are linked to an increased risk of certain cancers in association with higher levels of insulin and insulin-like growth factor 1 [51]. Newer therapies targeting the insulin and IGF1 systems are being developed for use in cancer therapy [52].

Based on our study and the previous reports, it is suggested that the association of obesity and T2D can be described as a "step-by-step" process in some way. At the initial stage, the abnormal accumulation of food intake and adipose tissues triggers the inflammatory cytokine expression and persistent immune reactions of fat cells. And then obesity-induced inflammation responses change the insulin sensitivity state and lead to the occurrence of insulin resistance consequently, which marks the transformation from obesity to obesity-related T2D.

But how these pathways interplay underlies the pathophysiology of obesity and T2D is still a big challenge. And the risk of false negative in the biological network also exists. Even with these challenges, network-based systems biology is increasingly attracting much attention from communities of both experimental and computational biologists and is expected to revolutionize our understanding of complicated disease as a whole. The methodologies and techniques of systems biology have been applied to analyzing the molecular mechanisms of complex diseases and provided new solutions for preventing and curing the diseases [53]. With vast amounts of omics data generated, the method still provides a new perspective for the future disease association studies.

3.4. NOT2D, a Database for Human Obesity and T2D. We constructed the NOT2D (network of human obesity and T2D) database to store the known disease genes and the interaction or regulatory links that are related to obesity and T2D (Figure 6). The obesity or T2D related genes, pathways, and networks can be accessed and downloaded from the website http://lilab.life.sjtu.edu.cn:8080/NOT2D/.

4. Conclusion

We studied the association between obesity and T2D by combing the gene expression profiles and the comprehensive biological network including both protein-protein interactions, metabolic and regulatory links. This study revealed that the connection of obesity and T2D mainly relied on several pathways involved in the digestive metabolism, immunization, and signal transduction. Our network-based association analysis provided better support and systematic explanation for the close connection between obesity and T2D than in view of individual gene.

Conflict of Interests

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

Acknowledgments

The work was supported by grants from the National Key Basic Research Program (2011CB910204), the National Natural Science Foundation of China (31271416), and the National High-Tech R&D Program (863) (2012AA020201, 2012AA101601).

References

[1] A. Guilherme, J. V. Virbasius, V. Puri, and M. P. Czech, "Adipocyte dysfunctions linking obesity to insulin resistance and type 2 diabetes," *Nature Reviews Molecular Cell Biology*, vol. 9, no. 5, pp. 367–377, 2008.

[2] K. M. Flegal, D. Carroll, B. K. Kit, and C. L. Ogden, "Prevalence of obesity and trends in the distribution of body mass index among US adults, 1999–2010," *The Journal of the American Medical Association*, vol. 307, no. 5, pp. 491–497, 2012.

[3] D. K. Patel, R. Kumar, D. Laloo, and S. Hemalatha, "Diabetes mellitus: an overview on its pharmacological aspects and reported medicinal plants having antidiabetic activity," *Asian Pacific Journal of Tropical Biomedicine*, vol. 2, no. 5, pp. 411–420, 2012.

[4] J. Ahmad, F. Ahmed, M. A. Siddiqui, B. Hameed, and I. Ahmad, "Inflammation, insulin resistance and carotid IMT in first degree relatives of north Indian type 2 diabetic subjects," *Diabetes Research and Clinical Practice*, vol. 73, no. 2, pp. 205–210, 2006.

[5] J. C. Pickup and M. A. Crook, "Is Type II diabetes mellitus a disease of the innate immune system?" *Diabetologia*, vol. 41, no. 10, pp. 1241–1248, 1998.

[6] G. M. Reaven, "Insulin resistance and human disease: a short history," *Journal of Basic and Clinical Physiology and Pharmacology*, vol. 9, no. 2–4, pp. 387–406, 1998.

[7] M. F. Gregor and G. S. Hotamisligil, "Inflammatory mechanisms in obesity," *Annual Review of Immunology*, vol. 29, pp. 415–445, 2011.

[8] T. O. Kilpeläinen, T. A. Lakka, D. E. Laaksonen et al., "SNPs in PPARG associate with type 2 diabetes and interact with physical activity," *Medicine and Science in Sports and Exercise*, vol. 40, no. 1, pp. 25–33, 2008.

[9] R. M. Evans, G. D. Barish, and Y.-X. Wang, "PPARs and the complex journey to obesity," *Nature Medicine*, vol. 10, no. 4, pp. 355–361, 2004.

[10] J.-J. Jia, X. Zhang, C.-R. Ge, and M. Jois, "The polymorphisms of UCP2 and UCP3 genes associated with fat metabolism, obesity and diabetes: etiology and pathophysiology," *Obesity Reviews*, vol. 10, no. 5, pp. 519–526, 2009.

[11] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabási, "The human disease network," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 21, pp. 8685–8690, 2007.

[12] U. Stelzl, U. Worm, M. Lalowski et al., "A human protein-protein interaction network: a resource for annotating the proteome," *Cell*, vol. 122, no. 6, pp. 957–968, 2005.

[13] C. T. Harbison, D. B. Gordon, T. I. Lee et al., "Transcriptional regulatory code of a eukaryotic genome," *Nature*, vol. 431, no. 7004, pp. 99–104, 2004.

[14] T. Ravasi, H. Suzuki, C. V. Cannistraci et al., "An atlas of combinatorial transcriptional regulation in mouse and man," *Cell*, vol. 140, no. 5, pp. 744–752, 2010.

[15] A. Hamosh, A. F. Scott, J. Amberger, C. Bocchini, D. Valle, and V. A. McKusick, "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders," *Nucleic Acids Research*, vol. 30, no. 1, pp. 52–55, 2002.

[16] T. S. K. Prasad, R. Goel, K. Kandasamy et al., "Human protein reference database—2009 update," *Nucleic Acids Research*, vol. 37, no. 1, pp. D767–D772, 2009.

[17] V. Matys, E. Fricke, R. Geffers et al., "TRANSFAC: transcriptional regulation, from patterns to profiles," *Nucleic Acids Research*, vol. 31, no. 1, pp. 374–378, 2003.

[18] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori, "The KEGG resource for deciphering the genome," *Nucleic Acids Research*, vol. 32, pp. D277–D280, 2004.

[19] T. Barrett, D. B. Troup, S. E. Wilhite et al., "NCBI GEO: archive for functional genomics data sets-10 years on," *Nucleic Acids Research*, vol. 39, no. 1, pp. D1005–D1010, 2011.

[20] R. A. Irizarry, B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed, "Summaries of affymetrix GeneChip probe level data," *Nucleic acids research*, vol. 31, no. 4, 2003.

[21] B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed, "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," *Bioinformatics*, vol. 19, no. 2, pp. 185–193, 2003.

[22] R. A. Irizarry, B. Hobbs, F. Collin et al., "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics*, vol. 4, no. 2, pp. 249–264, 2003.

[23] G. K. Smyth, "Linear models and empirical Bayes methods for assessing differential expression in microarray experiments," *Statistical Applications in Genetics and Molecular Biology*, vol. 3, no. 1, pp. 1–25, 2004.

[24] T. Ideker, O. Ozier, B. Schwikowski, and A. F. Siegel, "Discovering regulatory and signalling circuits in molecular interaction networks," *Bioinformatics*, vol. 18, supplement 1, pp. S233–S240, 2002.

[25] P. Shannon, A. Markiel, O. Ozier et al., "Cytoscape: a software Environment for integrated models of biomolecular interaction networks," *Genome Research*, vol. 13, no. 11, pp. 2498–2504, 2003.

[26] G. D. Bader and C. W. V. Hogue, "An automated method for finding molecular complexes in large protein interaction networks," *BMC Bioinformatics*, vol. 4, article 2, 2003.

[27] A. Chawla, K. D. Nguyen, and Y. P. S. Goh, "Macrophage-mediated inflammation in metabolic disease," *Nature Reviews Immunology*, vol. 11, no. 11, pp. 738–749, 2011.

[28] S. A. Kliewer, B. M. Forman, B. Blumberg et al., "Differential expression and activation of a family of murine peroxisome proliferator-activated receptors," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 91, no. 15, pp. 7355–7359, 1994.

[29] E. D. Rosen, C. J. Walkey, P. Puigserver, and B. M. Spiegelman, "Transcriptional regulation of adipogenesis," *Genes & Development*, vol. 14, no. 11, pp. 1293–1307, 2000.

[30] X. Lin, A. Taguchi, S. Park et al., "Dysregulation of insulin receptor substrate 2 in beta cells and brain causes obesity and diabetes," *Journal of Clinical Investigation*, vol. 114, no. 7, pp. 908–916, 2004.

[31] A. Brunetti, G. Manfioletti, E. Chiefari, I. D. Goldfine, and D. Foti, "Transcriptional regulation of human insulin receptor gene by the high-mobility group protein HMGI(Y)," *The FASEB Journal*, vol. 15, no. 2, pp. 492–500, 2001.

[32] J. C. Bruning, D. Gautam, D. J. Burks et al., "Role of brain insulin receptor in control of body weight and reproduction," *Science*, vol. 289, no. 5487, pp. 2122–2125, 2000.

[33] C. T. M. van Rossum, H. Pijl, R. A. H. Adan, B. Hoebee, and J. C. Seidell, "Polymorphisms in the NPY and AGRP genes and body fatness in Dutch adults," *International Journal of Obesity*, vol. 30, no. 10, pp. 1522–1528, 2006.

[34] H. Krude, H. Biebermann, W. Luck, R. Horn, G. Brabant, and A. Grüters, "Severe early-onset obesity, adrenal insufficiency and red hair pigmentation caused by POMC mutations in humans," *Nature Genetics*, vol. 19, no. 2, pp. 155–157, 1998.

[35] J. Delplanque, M. Barat-Houari, C. Dina et al., "Linkage and association studies between the proopiomelanocortin (POMC) gene and obesity in Caucasian families," *Diabetologia*, vol. 43, no. 12, pp. 1554–1557, 2000.

[36] J. Wang, D. Duncan, Z. Shi, and B. Zhang, "WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013," *Nucleic Acids Research*, vol. 41, no. W1, pp. W77–W83, 2013.

[37] S. Fröjdö, H. Vidal, and L. Pirola, "Alterations of insulin signaling in type 2 diabetes: a review of the current evidence from humans," *Biochimica et Biophysica Acta—Molecular Basis of Disease*, vol. 1792, no. 2, pp. 83–92, 2009.

[38] F. Bost, M. Aouadi, L. Caron, and B. Binétruy, "The role of MAPKs in adipocyte differentiation and obesity," *Biochimie*, vol. 87, no. 1, pp. 51–56, 2005.

[39] M. Qatanani and M. A. Lazar, "Mechanisms of obesity-associated insulin resistance: many choices on the menu," *Genes and Development*, vol. 21, no. 12, pp. 1443–1455, 2007.

[40] H. Gehart, S. Kumpf, A. Ittner, and R. Ricci, "MAPK signalling in cellular metabolism: stress or wellness?" *EMBO Reports*, vol. 11, no. 11, pp. 834–840, 2010.

[41] X. Yao and J. G. Forte, "Cell Biology of Acid Secretion by the Parietal Cell," *Annual Review of Physiology*, vol. 65, pp. 103–131, 2003.

[42] R. J. Turner and H. Sugiya, "Understanding salivary fluid and protein secretion," *Oral Diseases*, vol. 8, no. 1, pp. 3–11, 2002.

[43] M. I. Goran and M. I. Goran, "Energy metabolism and obesity," *Medical Clinics of North America*, vol. 84, no. 2, pp. 347–362, 2000.

[44] T.-D. Kanneganti and V. D. Dixit, "Immunological complications of obesity," *Nature Immunology*, vol. 13, no. 8, pp. 707–712, 2012.

[45] T. H. Mogensen, "Pathogen recognition and inflammatory signaling in innate immune defenses," *Clinical Microbiology Reviews*, vol. 22, no. 2, pp. 240–273, 2009.

[46] Y. Li, L. Ding, W. Hassan, D. Abdelkader, and J. Shang, "Adipokines and hepatic insulin resistance," *Journal of Diabetes Research*, vol. 2013, Article ID 170532, 8 pages, 2013.

[47] H. Tilg and A. R. Moschen, "Adipocytokines: mediators linking adipose tissue, inflammation and immunity," *Nature Reviews Immunology*, vol. 6, no. 10, pp. 772–783, 2006.

[48] C. W. Schindler, "Series introduction. JAK-STAT signaling in human disease," *Journal of Clinical Investigation*, vol. 109, no. 9, pp. 1133–1137, 2002.

[49] Z. Chen, T. B. Gibson, F. Robinson et al., "MAP kinases," *Chemical Reviews*, vol. 101, no. 8, pp. 2449–2476, 2001.

[50] S. He and Y.-X. Tao, "Defect in MAPK signaling as a cause for monogenic obesity caused by inactivating mutations in the melanocortin-4 receptor gene," *International Journal of Biological Sciences*, vol. 10, no. 10, pp. 1128–1137, 2014.

[51] E. J. Gallagher, Y. Fierz, R. D. Ferguson, and D. LeRoith, "The pathway from diabetes and obesity to cancer, on the route to targeted therapy," *Endocrine Practice*, vol. 16, no. 5, pp. 864–873, 2010.

[52] D. H. Cohen and D. LeRoith, "Obesity, type 2 diabetes, and cancer: the insulin and IGF connection," *Endocrine-Related Cancer*, vol. 19, no. 5, pp. F27–F45, 2012.

[53] P. Villoslada, L. Steinman, and S. E. Baranzini, "Systems biology and its application to the understanding of neurological diseases," *Annals of Neurology*, vol. 65, no. 2, pp. 124–139, 2009.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 316735, 11 pages
http://dx.doi.org/10.1155/2015/316735

Research Article

Gene Coexpression and Evolutionary Conservation Analysis of the Human Preimplantation Embryos

Tiancheng Liu,[1] Lin Yu,[2] Guohui Ding,[1,3] Zhen Wang,[1] Lei Liu,[1] Hong Li,[1] and Yixue Li[1,3]

[1]*Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 320 Yueyang Road, Shanghai 200031, China*
[2]*Key Laboratory of Contraceptive Drugs and Devices of National Population and Family Planning Commission of China, Shanghai Institute of Planned Parenthood Research, 2140 Xietu Road, Shanghai 200032, China*
[3]*Shanghai Center for Bioinformation Technology, 1278 Keyuan Road, Shanghai 201203, China*

Correspondence should be addressed to Hong Li; honglibio@gmail.com and Yixue Li; yxli@sibs.ac.cn

Received 7 December 2014; Accepted 27 January 2015

Academic Editor: Zhi Wei

Evolutionary developmental biology (EVO-DEVO) tries to decode evolutionary constraints on the stages of embryonic development. Two models—the "funnel-like" model and the "hourglass" model—have been proposed by investigators to illustrate the fluctuation of selective pressure on these stages. However, selective indices of stages corresponding to mammalian preimplantation embryonic development (PED) were undetected in previous studies. Based on single cell RNA sequencing of stages during human PED, we used coexpression method to identify gene modules activated in each of these stages. Through measuring the evolutionary indices of gene modules belonging to each stage, we observed change pattern of selective constraints on PED for the first time. The selective pressure decreases from the zygote stage to the 4-cell stage and increases at the 8-cell stage and then decreases again from 8-cell stage to the late blastocyst stages. Previous EVO-DEVO studies concerning the whole embryo development neglected the fluctuation of selective pressure in these earlier stages, and the fluctuation was potentially correlated with events of earlier stages, such as zygote genome activation (ZGA). Such oscillation in an earlier stage would further affect models of the evolutionary constraints on whole embryo development. Therefore, these earlier stages should be measured intensively in future EVO-DEVO studies.

1. Introduction

Evolutionary developmental biology (EVO-DEVO) studies how the dynamics of development affect the phenotypic variation arising from genetic variation and its correlation with phenotypic evolution. In this subject there is a central issue: which is the most conserved period or the crucial section during the entire developmental process of an organism. While it is unarguable that the later stages of embryogenesis are not conserved among species, two major models have been proposed: the "funnel-like" model, in which the earliest embryo shows the most conserved pattern, and the "hourglass" model, in which the middle point of development is imposed with the most evolutionary constraints [1]. The "hourglass" model, which assumes the midembryonic stage (phylotypic), which shows developmental constraints and functional importance, was originally proposed due to the expression of Hox genes in middle point of vertebrate development [2] and has been preferred in comparative transcriptomic studies nowadays [3, 4]. In addition to the transcriptomic similarity of phylotypic stages between different species, transcriptome age index (TAI) based methods, which address the total evolutionary ages of expressed genes in each developmental stage, show convergent evolution matching an hourglass pattern of embryogenesis in animals and plants [5, 6].

Mammalian preimplantation embryonic development (PED) starts from fertilization and ends at implantation of the embryo in the endometrial lining of the uterus [7]. After fertilization, the major genetic substances in the transcriptome of the zygote are the maternally deposited transcripts. After 2-3 rounds of cell divisions, maternally inherited transcripts

are degraded gradually and new transcripts of zygote are produced by the new diploid nucleus. This process is termed zygote genome activation (ZGA) [8]. These changes are not easily captured by traditional gene expression microarray techniques, as the sensitivity of microarray technology is limited when detecting low expressed genes or expression in a single cell [5, 9]. With the development of single cell RNA sequencing technology [10], we were able to identify precisely gene expression changes during the embryo developmental process which are unapparent in the microarray analysis [5]. In order to illustrate the earliest developmental gene expression fluctuation of PED which contain the crucial ZGA process and may further affect the later developmental stages, it is meaningful to look into these PED stages and identify the genetic modules regulating in each period of PED [11].

From an EVO-DEVO viewpoint, inspection of the selective constrains in PED is interesting because the trend in this period would further influence the tendency of evolutionary constrains in the middle stages of embryo development such as the phylotypic stage. Despite the importance of the expression profile of the PED stage, previous comparative transcriptome research has yet to characterize it [3, 5]. The lack of understanding of these PED stages has led past researchers to conclude that selective constraints during the earlier developmental stages increase continuously in the "funnel-like" model while they decrease in the "hourglass" model. Analysis of the selective constrains of genes in each PED stage could aid in distinguishing between the formation mechanism of the "hourglass" model or the "funnel-like" model and also consummate the whole pattern of selective constrains that act on embryonic development.

Based on the single cell RNA sequencing results of human preimplantation embryos from the oocyte stage to late blastocyst stage [12]. Applying weighted gene coexpression network analysis (WGCNA) [13], we were able to identify representative genes in each stage and summarized selective pressure on these genes to clarify the selective trend in earlier developmental stages. We found certain patterns of the evolutionary constraints that imposed on different stages of human preimplantation embryos; therefore we illustrated selective constraints on PED stages, which also presented fluctuation properties, considering that these earlier stages should be included for studying the constraints on the whole embryo development.

2. Results

2.1. Coexpression Modules for Stages in Human Preimplantation Embryos. In the course of evolution, most biodiversity is due to alterations in gene regulation relationships rather than the sequence mutations on genes [14]. Coexpression gene modules tend to evolve together so as to share evolutionary patterns [15]. Therefore, we used gene coexpression analysis rather than differential expression to identify genes that may have close regulation relationships [16].

In order to study selective constrains in preimplantation embryonic development, we analyzed the transcriptome profiles of human preimplantation embryos (including oocyte, zygote, 2-cell, 4-cell, 8-cell, morulae, and late blastocyst stages) that were obtained by single cell RNA sequencing. Stage-specific coexpression modules were selected by the gene coexpression network analysis (WGCNA) (Figure 1) which is an unsupervised clustering method to group genes which have coexpression patterns into distinct modules [13]. This is a reliable gene coexpression analysis tool and is wildly adopted by many investigators [17–20]. After merging correlated modules with a stringent threshold, we assigned 27 out of 41 modules into a specific preimplantation developmental stage according to the correlation of eigengene of every module with each stage indicator ($r > 0.6$, $P < 0.001$). Some modules might correlate with two adjacent developmental stages because of the similarity of these two adjacent developmental stages. To remove the bias of stage comparison, we assigned this kind of modules to stage in which they had the highest correlation coefficient. After that, each module was classified into a specific developmental stage and most of the genes in each module showed consistent overexpressed behavior in corresponding developmental stage (Figure 1).

Genes in multiple modules of the same stage were merged together. In total, we obtained 2 coexpression modules for the oocyte stage, 1 module for the zygote stage, 1 module for the 2-cell stage, 5 modules for the 4-cell stage, 4 modules for the 8-cell stage, 5 modules for the morulae stage, and 9 modules for the late blastocyst stage (Figure 2). We obtained 1409, 583, 481, 1494, 1731, 1720, and 3132 specific genes for each stage, respectively. The large number of genes in the oocyte showed a complicated regulation mechanism that involved the expression of maternal genes. The number of coexpression modules and genes gradually increased with the progress of zygote development, which implied the formation of embryo complexity and modularity.

2.2. Validation of the Biological Function for Modules. We further investigated the biological functions of genes in each specific stage by using DAVID software [21]. Gene ontology biological process (GOBP) enrichment analysis showed that genes from each stage were enriched in the relevant functions of corresponding developmental process. We also verified the function of genes in each stage by comparing them with the known function categories that were identified by Xue et al. on a different dataset of human preimplantation embryos [11]. And we compared them with the functional term identified by different methods on same datasets [12] (Table 1). The zygote gene activation (ZGA) process, which is the principal transformation of the pre-implantation period, was endorsed by significant overrepresentation of genes involved in transcription and transcription regulation process from 4-cell stage to morulae stage. In the late blastocyst stage, genes were significantly enriched in protein translation and function-associated pathways such as protein localization, transport, and phosphorylation.

2.3. Various Selective Pressures on Gene Sequence. The nonsynonymous to synonymous substitution ratio (dN/dS) is a widely used method to measure gene sequence conservation [22]. We used dN/dS ratio for genes in each module to quantify the selective pressure on the corresponding developmental stage. The dN/dS ratios were calculated between

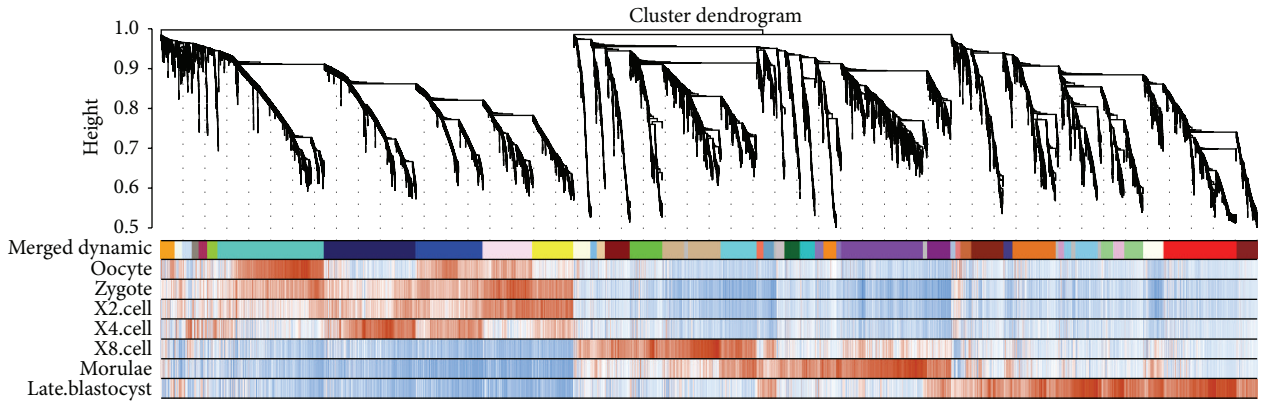


Figure 1: Coexpression gene modules in human preimplantation embryonic development. Hierarchical cluster tree shows the coexpression modules identified by WGCNA. The panels from top to bottom are merged dynamic modules labeled with different colors and genes correlation with indicators of each stage. The red means positive correlation while blue means negative correlation, and correlation coefficients are in direct proportion to the color depth.

mouse and human, as we intended to measure the pressure acting on sequence of genes in the mammalian species. Then their distributions were illustrated in Figure 3(a). Next we randomly sampled same number genes within each stage and calculated the median of the dN/dS distribution for the random dataset, which stood for the background. Figure 3(b) shows the median of dN/dS for genes in coexpression modules and the median of dN/dS for randomly selected genes.

The dN/dS ratios of stage-specific genes gradually decreased until the 4-cell stage, which may be caused by the consumption of the maternal genes and the expression of new genes of the zygote itself as shown by previous studies of preimplantation embryos [11, 12, 23–26]. Oocytes and zygotes had a higher median of dN/dS ratios relative to the median dN/dS for all genes whereas the dN/dS ratios of genes belonging to the 4-cell stage were significantly lower than the median dN/dS for all genes. From the zygote stage to the turning point 4-cell stage, the decreasing trend of dN/dS ratios was parallel with the process that the maternal genes expended and zygote genes emerged. The genes of the 4-cell stage were more inclined to be expressed by zygotes and had low dN/dS ratios. At the same time, genes regulated in the zygote or oocyte stages were left by maternal source and these genes had high dN/dS ratios. So the decreasing trend from maternity to zygote might suggest more striking selective pressure acting on the genes produced by zygote than selective pressure effecting genes inherited from maternity [27]. After the 4-cell stage we detected a pattern of increasing dN/dS ratios, which shows these stages expressed genes with selective pressure not as strong as 4-cell stage.

2.4. Stage-Specific Genes Were Born in Different Ancient Roots. The ages of stage-specific genes have been used as indices of evolutionary constraint [5]. We traced the root of every gene expressed in human preimplantation embryo in the phylogeny and used the ancient level of the root to represent the conservation of the gene. Based on the phylogenetic taxonomy of their roots, genes were separated into four groups: (1) Opisthokonta-Bilateria, (2) Sarcopterygii-Amniota, (3) Chordata-Euteleostomi, and (4) Mammalia-Eutheria. For each gene set, the number of genes in each of the 4 groups was calculated to represent the age distribution. Next we marked every preimplantation developmental stage with a specific age distribution and used the age distribution of all genes as background. To detect the difference of gene age during different development stages, the age distribution of genes in each stage was compared with the background distribution of all genes (Figure 4).

We detected a clear changing trend for the Opisthokonta-Bilateria genes, with their proportion decreasing from the zygote to the 8-cell stage and then increasing until the late blastocyst. In particular, the genes belonging to the 8-cell stage were significantly depleted in Opisthokonta-Bilateria and overrepresented in Mammalia-Eutheria, which implied most genes expressed in this stage are recently born in the Mammalia-Eutheria lineage compared to other stages. In other words, these new genes, which were expressed and regulated as modules in 8-cell stage, were products of developmental evolution in the Mammalia-Eutheria lineage. This suggested that genes expressed in the 8-cell stage had a crucial function for the ZGA process of organism in Mammalia-Eutheria lineage [11, 28, 29].

As Figure 3 shows, after the 8-cell stage there was an opposite trend of increasing Opisthokonta-Bilateria genes from depletion to overrepresentation and decreasing Mammalia-Eutheria genes from overrepresentation to depletion. Finally the late blastocyst stage showed the opposite pattern—the stage was significantly depleted genes belonging to Mammalia-Eutheria and Chordata-Euteleostomi groups and it was overrepresented of genes in Opisthokonta-Bilateria group. This sort of opposite pattern illustrated that the late blastocyst stage was conserved as it tended to express the oldest genes.

2.5. Genes in Each Stages Present Diverse Duplicated States. Gene duplication state is also an indicator of selective pressure [30]. In order to evaluate the conservation of genes more widely, we chose the zebra fish, an evolutionary distant

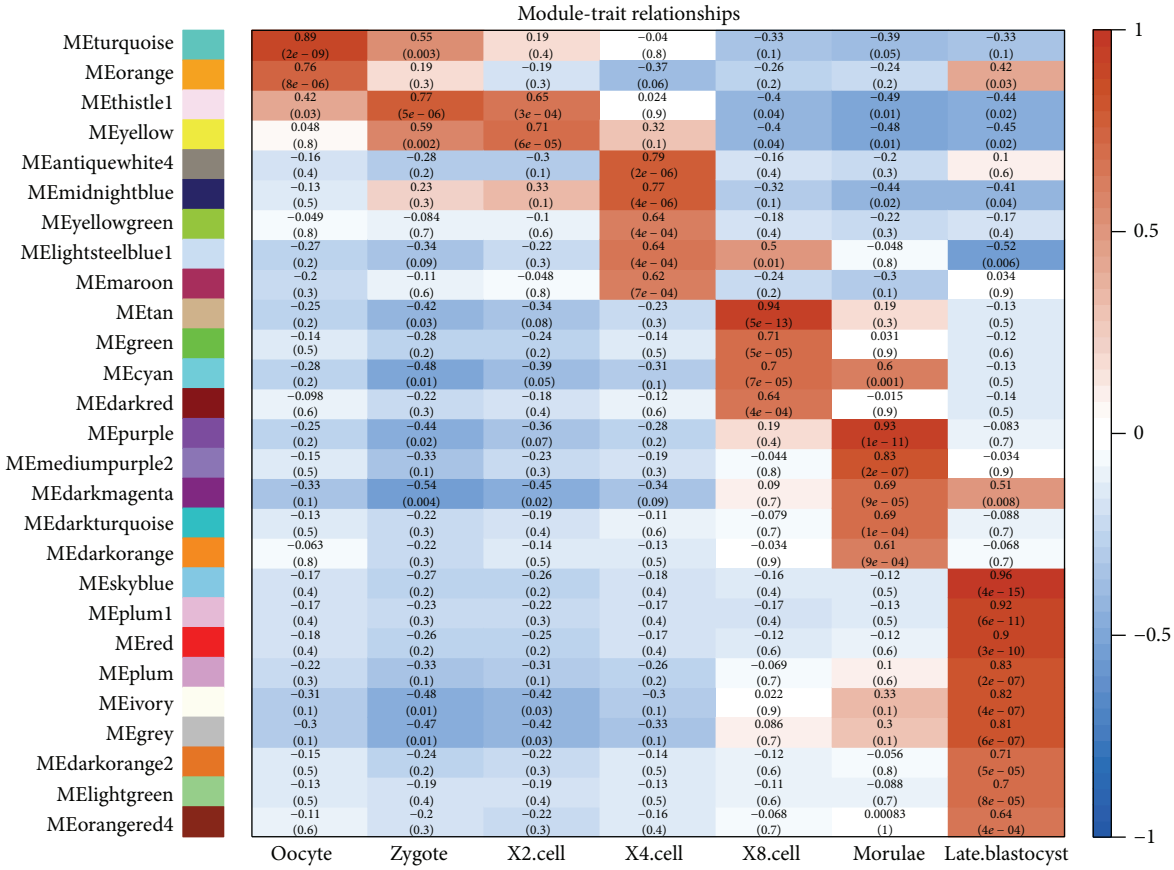


FIGURE 2: Modules are distributed to each stage according to the correlation between eigengenes and stage indicators. The red means positive correlation while blue means negative correlation, and correlation coefficients are in direct proportion to the color depth as shown in right side. Each element of the matrix denotes the correlation coefficients between module eigengenes (row) and stage (column); then the significant level of correlation is marked below the correlation coefficients.

species, as reference to check the gene-duplicated situation of human genes in each development stage. Genes were separated into four groups based on the gene duplication states: (1) one-to-many, (2) one-to-one, (3) many-to-many, and (4) new gene (no ortholog in the zebra fish genome). We removed the many-to-many gene pairs because it is difficult to evaluate their conservation. As stated above, we compared the observed distribution of genes in each stage with the expected distribution that was recorded by distributing all genes into these 3 groups (Figure 5).

Genes falling in one-to-many orthologs revealed that they were single copy in human and their orthologs had duplications in zebra fish. Knowing that constrained developmental stages should display less change in gene family size [31], the genes, which duplicate in other species but keep singleton in human developmental stages, should be considered to be conserved specifically in *Homo sapiens*. Otherwise, the one-to-one orthologs retain the functions of ancestral gene since the last shared common ancestor, and left no duplication in the human or zebra fish lineage. Therefore genes of one-to-one orthologs also should be subject to functional constraints. Just as in the above age analysis of genes, the new genes, which were new products during the evolutionary process of *Homo sapiens*, were considered to be under less constraint. At last, many-to-many orthologs showed duplication events in both species and their conservation patterns were complicated; thus we ignored many-to-many orthologs in the further analysis.

As Figure 5(b) shows, the one-to-one (single-copy) orthologs and new genes exhibited opposite trends in the preimplantation period which implied the transformation of evolutionary constraints on different stages during the developmental process. In particular, the 4-cell stage showed significant depletion of the one-to-one genes but overrepresentation of the one-to-many genes that signify genes of this stage is under strong functional constraints on their sequence in *Homo sapiens*. It accorded with the dN/dS result showing that genes expressed in the 4-cell stage had significantly lower human-mouse dN/dS ratios and lends further evidence to the hypothesis of conservation of genes belonging to the 4-cell stage. Moreover, the 8-cell stage showed overrepresentation of the newborn genes and depletion of the one-to-one genes, which is consistent with the gene age analysis. The large number of new genes in the 8-cell stage offers further evidence for human-specific embryonic development occurring in this stage [11, 28, 29].

Table 1: Enriched biological process terms for stage-specific genes in preimplantation embryonic development. Similar function annotation clustering was presented by one typical function term. The last column stands for the validation of corresponding term in other studies of human preimplantation embryonic development.

Stage	Functional term	Bonferroni P value	Consistence with previous report
Oocyte	DNA repair	$3.35E-4$	[11, 12]
	Cell cycle	$2.92E-3$	
	DNA metabolic process	$7.30E-3$	
Zygote	Cell cycle	$3.33E-4$	[12]
2-cell	Posttranscriptional regulation of gene expression	$3.56E-2$	
4-cell	Transcription	$1.76E-4$	[11, 12]
	Regulation of transcription	$6.34E-3$	[11, 12]
	Small GTPase mediated signal transduction	$3.14E-2$	[11]
8-cell	DNA packaging	$8.09E-6$	
	RNA processing	$9.27E-5$	[11]
	Transcription	$1.02E-4$	[11, 12]
	Regulation of transcription	$3.08E-4$	[11, 12]
	Translational elongation		
	Ribonucleoprotein complex biogenesis	$5.89E-4$	[12]
	RNA elongation	$2.09E-3$	
Morulae	RNA processing	$2.38E-8$	[11]
	Transcription	$7.33E-8$	[11, 12]
	Regulation of transcription	$5.38E-7$	[11, 12]
	Ribonucleoprotein complex biogenesis	$9.81E-6$	[12]
Late blastocyst	Generation of precursor metabolites and energy	$7.78E-18$	[12]
	Translation	$4.11E-12$	[12]
	Oxidative phosphorylation	$9.24E-12$	[12]
	Cellular respiration	$1.69E-8$	[12]
	Membrane organization	$9.28E-6$	
	Protein localization	$2.74E-5$	
	Cofactor metabolic process	$3.64E-5$	
	Protein transport	$6.17E-5$	
	Monosaccharide metabolic process	$3.67E-3$	

As with the above gene age analysis, we also detected conserved convergence from the 8-cell stage to late blastocyst stage reflected by the transition from an overrepresented state to a depleted state of the new genes and by the transition from depleted state to overrepresented state of the one-to-one genes. Finally, the late blastocyst stage reached a conserved state, which was significantly depleted of new genes and overrepresented for one-to-one genes.

2.6. Evolvability of Regulatory Regions in Upstream of Stage-Specific Genes. Conservation of *cis*-regulatory sequences is also a critical standard for measuring the selective pressure on genes [14, 32], and highly conserved noncoding elements (HCNEs) are often considered to be associated with developmental regulatory genes or transcription factors (TFs) [23, 33]. Therefore, we determined the transcriptional importance of stage-specific genes by analyzing their potential to become TFs and the distribution of HCNEs in their promoter regions (Figure 6).

We found that promoter regions of genes in the 2-cell stage were significantly enriched for HCNEs, and there are more TFs in 2-cell stage than expected. These enriched transcriptional factors and transcriptional regulatory elements may promote effective gene transcripts in the 2-cell embryo and launch the progress of zygote genome activation (ZGA). TFs were significantly enriched in 4-cell, 8-cell, and morulae stages, which indicated that the gene expression and regulation network became more sophisticated during the zygote gene activation (ZGA) process. Our finding of a relatively desolate transcriptional scenario in the late blastocyst stage accords with the findings of Piasecka et al. [31], who proposed that the cleavage/blastula modules of zebra fish development are not enriched with transcriptional devices. Finally, the gathering of these transcriptional elements during the ZGA

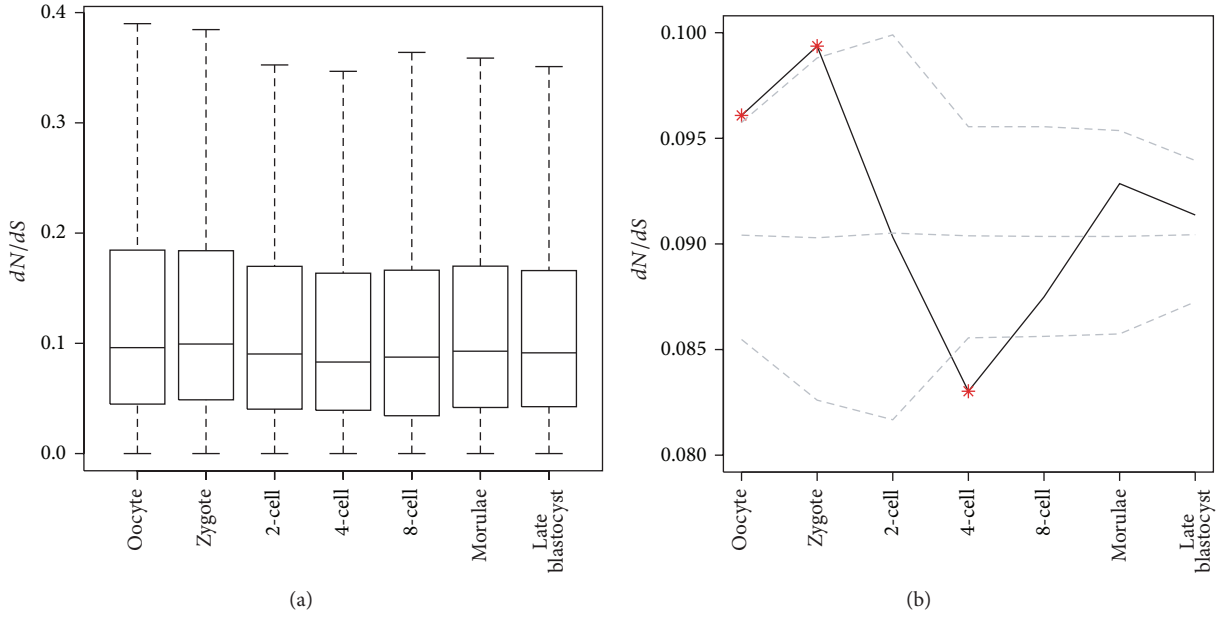


FIGURE 3: (a) Box plot of conserved index during different development stages. Conservation is measured by the nonsynonymous to synonymous substitution ratios (dN/dS). (b) The median of dN/dS ratio in each stage is represented by the solid line. The dash lines denote the median and the 95% confidence interval for the dN/dS ratio of randomly selected genes, which have same number of genes in corresponding stage. The asterisks denote that dN/dS are significantly different ($P < 0.05$) from the median of background.

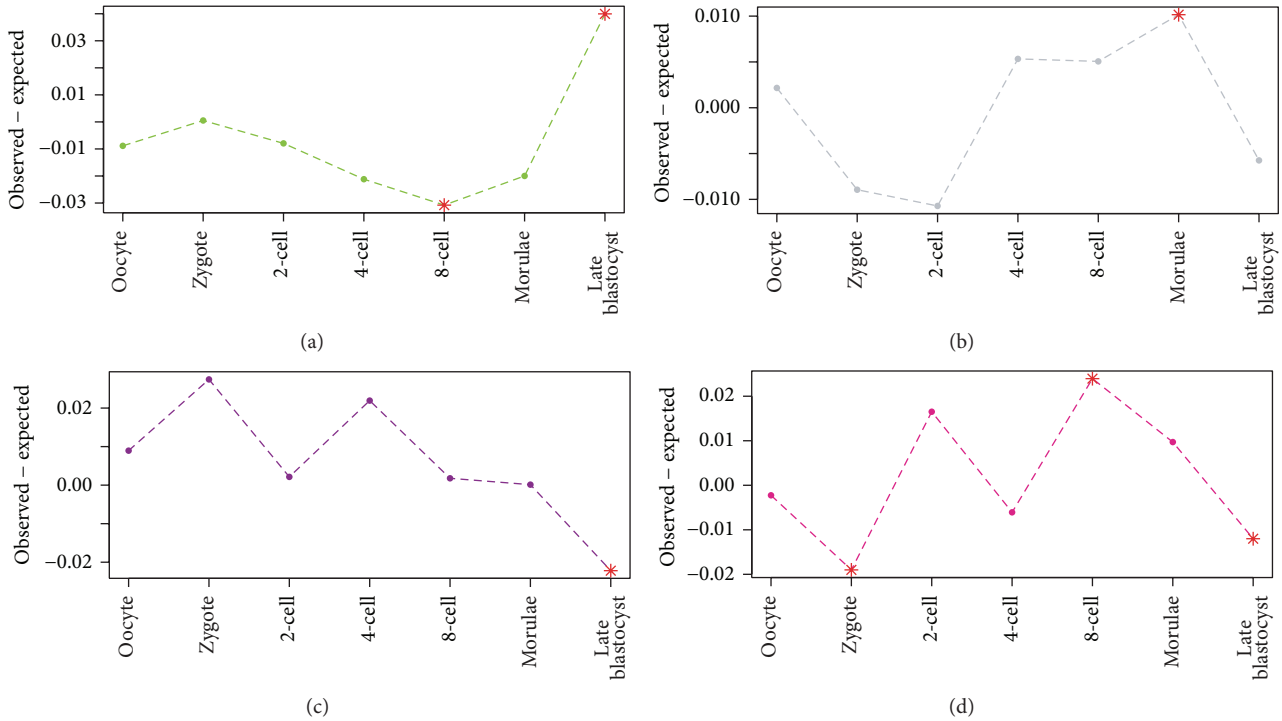


FIGURE 4: Distributions of gene ages. Genes were classified into four groups based on their first appearance in the phylogeny: (a) Opisthokonta-Bilateria, (b) Sarcopterygii-Amniota, (c) Chordata-Euteleostomi, and (d) Mammalia-Eutheria. For each stage, the vertical axis shows the observed frequencies minus expected frequencies of gene ages. The asterisks denote significant enrichment ($P < 0.01$) in a specific gene group for each stage.

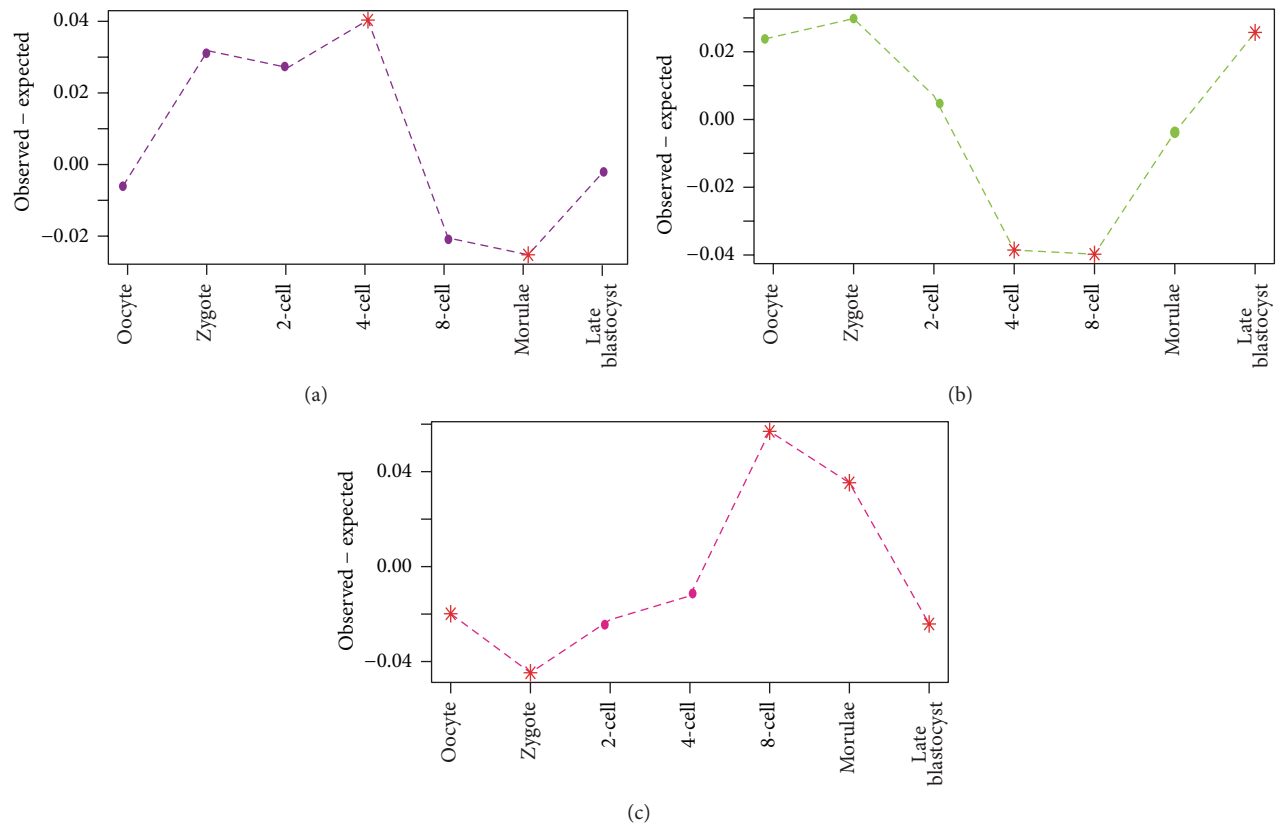


FIGURE 5: Distributions of different gene ortholog types comparing human with zebra fish: (a) one-to-many ortholog, (b) one-to-one ortholog, and (c) new genes. For each stage, the vertical axis shows the observed proportions minus expected proportions of genes in each ortholog type. The asterisks denote significant enrichment ($P < 0.01$) in a specific ortholog scenario for each stage.

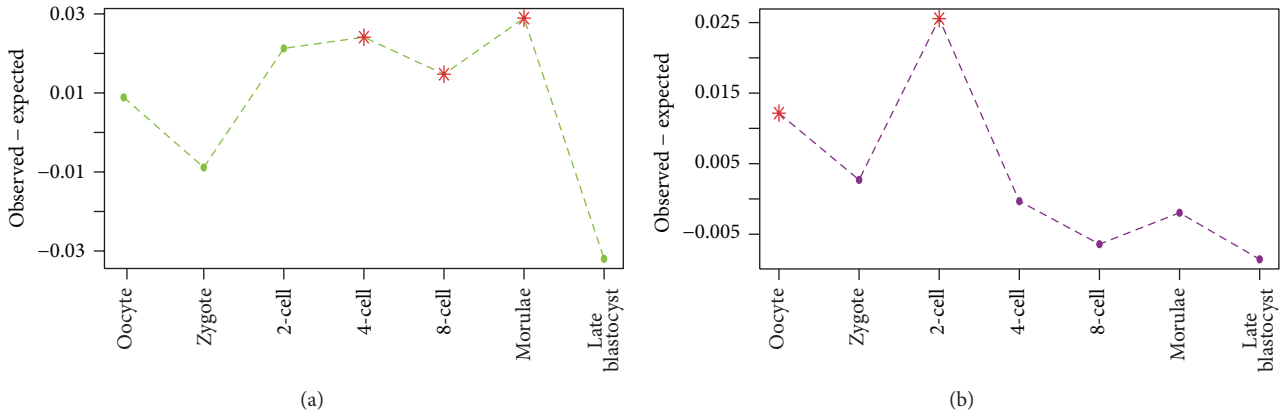


FIGURE 6: (a) Distributions of transcription factors (TFs) in stage-specific modules. (b) Distributions of genes with highly conserved noncoding elements (HCNEs) in their *cis*-regulatory regions. For each stage, the vertical axis shows the observed proportions minus expected proportions of TF and genes with HCNEs, respectively. The asterisks denote significant enrichment ($P < 0.01$).

process could not be disregarded and this might further influence the evolutionary model or regulation mechanism of the whole developmental schedule.

2.7. Patterns of Evolutionary Constraints in Preimplantation Embryonic Development. Based on WGCNA, we clustered the genes of human preimplantation embryonic development into modules and linked these modules to specific stages of this developmental process. Next, we checked four conservation properties for stage-specific genes, including gene sequences, gene ages, gene orthologs, and regulatory elements. All of these indices implied several features during

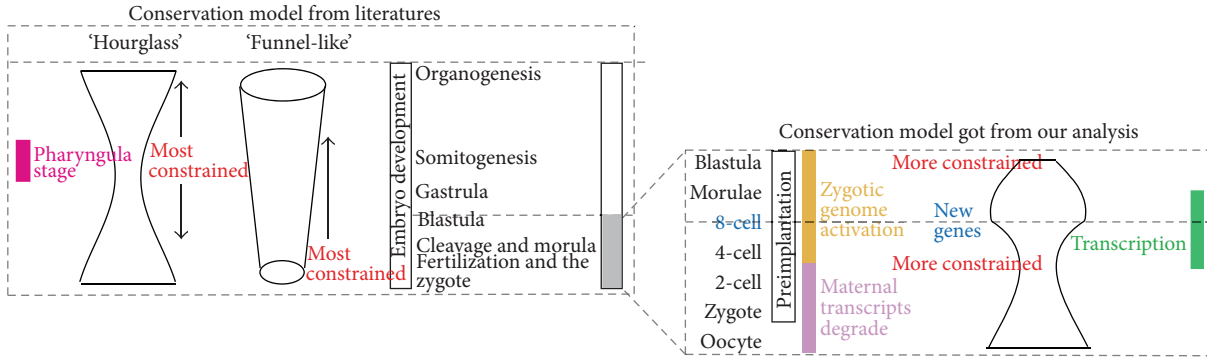


FIGURE 7: (a) Two models—the "funnel-like" model and the "hourglass" model, revealed by the evolutionary constraint on the whole developmental process of organism. Corresponding stages are labeled on the right side and earlier stages to late stages are in bottom to up sequence. Pharyngula stage (phylotypic stage) is displayed by pink rectangle on the left side. (b) The fluctuation of evolutionary constraints on the stages of human preimplantation embryonic development. In the left side, preimplantation stages are ranked from bottom to up. Maternal transcripts degrade and ZGA are labeled with purple rectangle and yellow rectangle, respectively.

the process of human preimplantation embryonic development.

First, we observed that maternal genes were under less selective constraints while there were strong selective constraints effecting on earlier zygote-activated genes. This was verified via the reduction of the dN/dS ratio accompanied by the consumption of maternal mRNAs and ZAG expressing from zygote to 4-cell stage (Figure 3), and the overrepresentation of the conserved one-to-many orthologs in the 4-cell stage (Figure 5). Secondly, we discovered a switch of the evolutionary constraints at 8-cell stage in which the embryo tended to express new genes. This trend is reflected by the fact that dN/dS begin elevating after the 8-cell stage (Figure 3). Meanwhile, genes in 8-cell stage present depletion of oldest Opisthokonta-Bilateria genes and show overrepresented of newest Mammalia-Eutheria genes (Figure 4). The burst of new genes in the 8-cell stage was further demonstrated by the depletion of one-to-one zebra fish orthologs genes and overrepresentation of human specific genes in this stage (Figure 5). Lastly, the selective pressure on late blastocyst stage tended to increase again. The late blastocyst stage was overrepresented in the oldest Opisthokonta-Bilateria genes (Figure 4) and one-to-one orthologs of zebra fish (Figure 5). The phylotypic stage in middle development was specifically enriched for transcriptional elements so that the transcriptional factory was subtly working in this stage [31]. Our work revealed that ZGA in early stages also showed the enrichment of transcriptional elements (Figure 6), which indicated the ZGA process is under precise regulation as phylotypic stage.

In summary, we found that, in the earlier developmental stages of the human embryo, the conservation indices presented the sequence of increasing—decreasing—increasing (Figure 7), rather than increasing or decreasing monotonically. And this trend is potentially correlated with the maternal transcripts degrade and ZGA. As part of the developmental process, earlier embryo development turns out to be a complicated process which also involves the fluctuation of selective pressure.

3. Discussion

Mammalian developments comprise three important processes: zygote genome activation (ZGA) at earlier stages [11], expression of Hox genes at middle stages [2], and morphological formation at late stages [34]. The evolutionary conservation of these three stages has been debated at length. Previous EVO-DEVO studies concerning development [1, 4, 5, 31] focused on conservation during the whole development process, while changes of selective pressure during earlier development were neglected. These studies typically used one stage (such as the zygote stage) to represent the earlier embryonic stages [5]. Most stages of the earlier embryo (such as 2-cell to 8-cell stages) were discarded as they are relatively short compared to the long time interval of stages in middle and late embryo. By monitoring these transient stages of earlier development, the exquisite regulation mechanism of the ZGA process could be revealed [11, 12]. Rather than monitoring once after certain time intervals [4], the detecting time points should be chosen according to developmental events and time interval of different stages should be specially selected. Therefore it is meaningful to set more observing points during the earlier stages [35]. Here we analyze eight stages in preimplantation embryonic development. Our results show that the conservation scenario of the earlier embryo has a degree of fluctuation different from the direct increase or decrease previously reported [1, 3, 5, 6, 31]. The fluctuation in PED stages was probably associated with the events occurring during these stages. For instance, more selective pressure was on early zygotic genes than maternal genes [27]; therefore the measure of conservation increased during the maternal transcripts degrading process. Then the embryo was building the infrastructure so that it expressed highly conserved genes during the 2-cell stage to 4-cell stage. After this, the embryo expressed some species specific genes to determine its fate inclination [36]. Finally, the genes expressed near middle embryonic stages presented to be conserved, which was accordance with the hourglass

model. Our work illustrates that the dynamics of evolutionary indices during these short-time early stages should also be taken into consideration in discussions of the "hourglass" model or the "funnel-like" model of embryo development. We believe a precise exploration of the evolutionary indices of earlier developmental stages will lead to the creation of a more sophisticated model of selective pressure on the whole development process.

4. Methods

4.1. Transcriptional Profiling of Preimplantation Embryos. The gene expression profilings of human preimplantation embryos were downloaded from NCBI's Gene Expression Omnibus [37] (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE36552). It contained the whole transcriptomic RNA expression levels of oocyte stage, zygote stage, 2-cell stage, 4-cell stage, 8-cell stage, morulae stage, and late blastocyst stage, which were measured by RPKM (Reads Per Kilobase of transcript per Million mapped reads) via single cell RNA sequencing [38]. Each stage was composed of 4 biologically replicated samples except for oocyte and zygote stages, which had 3 samples. To eliminate the bias of genes which have zero or extremely low expression levels in many stages, genes with low expression in all stages (average RPKM < 0.5) were removed. Each gene symbol of the whole profile was mapped to its corresponding Ensembl gene ID and the gene symbols that have no corresponding Ensembl gene ID were discarded to reduce the potential noise. At last the expression profiles in each sample were processed by quantile normalization that accounts for different amounts of RNA present throughout embryo earlier development.

4.2. Weighted Gene Coexpression Network Analysis. The final expression matrix was proceed by the step by step WGCNA [13] method. First, we built a matrix which includes pairwise correlation coefficients between all pairs of genes [39]. Next, with the power of 12 which is the default value, the adjacency matrix was constructed. Depending on the resulting adjacency matrix, we calculated the topological overlap matrix, which measures the interconnectedness of the coexpression network [40]. And then this topological overlap matrix was used to perform hierarchical clustering in which genes with coexpression relationships were grouped together and formed a gene clustering tree. The primary modules were identified by Dynamic Hybrid Tree Cut algorithm [41] to cut the hierarchal clustering tree with coefficients deepSplit = 4. At last, we calculated the correlation coefficients of each pair of module eigengenes that stand for the first principal component of the module and merged highly similar modules by a stringent threshold (correlation > 0.9).

After identifying the coexpression modules, we associated these modules with specific embryo developmental stage and picked hub genes for each module. This process was based on correlating each module eigengene which represented each module with the stage indicator genes and all genes on the matrix. For genes with high correlation coefficients (correlation above 0.9 and P value < 0.01) with specific module, we treated them as the hub genes of this module. To associate these modules with developmental stage, we used a threshold (correlation coefficient > 0.7 and P value < 0.01) to pick up modules which belong to a certain stage. Modules that correlated with two stages were only kept in the stage with the highest correlation coefficients.

4.3. Gene Ontology Analysis. Functional annotation was performed with the DAVID Bioinformatics Resources. To correct multiple testing, the Bonferroni correction was applied. And the enriched GO biological process categories were picked up by the corrected P value (<0.01). Then we checked whether these enriched GO categories were also presented in the same stages of similar studies [11, 12].

4.4. dN/dS Analysis. We downloaded dN and dS values of all human genes using BioMart [42], which was calculated by the ortholog genes between human and mouse. After removing genes that were not presented on the expression matrix, we got 12865 dN/dS value.

We calculated the median dN/dS ratio of genes in each stage and evaluated if it was significantly higher (lower) than randomly selected genes. For each stage which has k genes, we generated 10000 sets of k randomly chosen genes from the background 12865 genes and calculated the median dN/dS ratio for each random set. P value was calculated as the tail probability of real dN/dS ratio in the distribution of randomly generated dN/dS.

4.5. Gene Age Analysis. Genes of *Homo sapiens* originated in different taxonomic root of the phylogeny so that genes have different age index. We could label every gene with an age index by its first appearance in the phylogeny. For each gene in our expression matrix the oldest node of its gene tree was retrieved from Ensembl release 75 [43] by Ensembl comparative genomics API. After that, each gene was marked with a unique age index from oldest Opisthokonta node to the latest human node.

In order to make subsequent test convincible we removed some genes falling in age interval from Eutheria to human because the number of genes in these interval is very rare (less than 5) which will obscure the statistical test. And the rest of the genes were merged into one of the following age intervals: Opisthokonta-Bilateria, Opisthokonta-Bilateria, Sarcopterygii-Amniota, and Mammalia-Eutheria. That made each category have sufficient number of genes to perform the statistical test.

Next, for every module we collected all the age indices of its k genes and counted the number of genes falling to each age interval. Then the age index distribution of expected background was estimated by classifying all genes (11919 genes) presented on the expression matrix into these categories. The number of genes in each age category was transformed to the proportion by dividing the number of all expressed genes in this stage. Then we plotted the "observed minus expected" proportions of each stage and performed Fisher's exact test to compare the observed and expected numbers of age indices in each stage. We picked up stages in which genes with certain age index were overrepresented

or underrepresented (P value < 0.01) and highlighted these stages on the plot.

4.6. Human-Zebra Fish Orthologous Genes. Based on an evolutional distant species zebra fish, homology information between human and zebra fish genes was retrieved from Ensembl release 75 [43] by BioMart [42]. 10919 of the 12865 genes presented in the expression matrix have human-zebra fish paired orthologs, including 7265 one-to-one orthologs, 2995 one-to-many orthologs, and 659 many-to-many orthologs. Then the remaining 1946 human genes which do not have ortholog relationship with zebra fish genome were labeled as new gene in human.

We calculated the observed number of stage-specific genes that were in the three kinds of ortholog types (one-to-one, one-to-many, and no orthology), and constructed the expected background distribution from all genes. For each category we plotted the "observed minus expected" proportions of each stage and performed Fisher's exact test to compare the observed and expected numbers.

4.7. Gene Transcriptional Region Analysis. Gene transcription analysis was based on the number of transcription factors (TFs) and highly conserved noncoding elements (HCNEs) in the promoter regions of genes. Genes with GO category annotation (GO: 0006355, regulation of transcription, DNA-dependent) were defined as TFs. Location data of HCNE between human and mouse with identity above 90% was downloaded from Ancora (http://ancora.genereg.net/downloads/hg19/vs_mouse/) [33]. For each of the 12865 genes considered in our analysis, we checked if there were HCNEs located in 500 base-pairs upstream from the transcription start site. Totally we annotated 1438 and 848 genes as TFs and HCNEs, respectively. For every stage we performed the hypergeometric test to assess if genes in this stage were significantly enriched in TFs and HCNEs.

Conflict of Interests

The authors declare they have no competing interests.

Authors' Contribution

Tiancheng Liu carried out the coexpression analysis and evolutionary analysis and then drafted the paper. Lin Yu provided the instruction of developmental biology and collected the datasets. Guohui Ding and Zhen Wang participated in understanding biological question and phylogenetic analysis. Hong Li revised the paper. Hong Li and Yixue Li conceived and designed the study. All authors read and approved the final paper. Tiancheng Liu and Lin Yu contributed equally to the study.

Acknowledgments

The authors appreciate revisions of paper from O'Sullivan Nathan. This work was supported by National Basic Scientific Research Fund (2009FY120100), State Key Basic Research Program (973) (2011CBA00801), SIBS Knowledge Innovation Program (2014KIP215), and SA-SIBS Scholarship Program.

References

[1] N. Irie and S. Kuratani, "Comparative transcriptome analysis reveals vertebrate phylotypic period during organogenesis," *Nature Communications*, vol. 2, article 248, 2011.

[2] D. Duboule, "Temporal colinearity and the phylotypic progression: a basis for the stability of a vertebrate Bauplan and the evolution of morphologies through heterochrony," *Development*, supplement, pp. 135–142, 1994.

[3] Z. Wang, J. Pascual-Anaya, A. Zadissa et al., "The draft genomes of soft-shell turtle and green sea turtle yield insights into the development and evolution of the turtle-specific body plan," *Nature Genetics*, vol. 45, no. 6, pp. 701–706, 2013.

[4] A. T. Kalinka, K. M. Varga, D. T. Gerrard et al., "Gene expression divergence recapitulates the developmental hourglass model," *Nature*, vol. 468, no. 7325, pp. 811–816, 2010.

[5] T. Domazet-Lošo and D. Tautz, "A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns," *Nature*, vol. 468, no. 7325, pp. 815–819, 2010.

[6] M. Quint, H.-G. Drost, A. Gabel, K. K. Ullrich, M. Bönn, and I. Grosse, "A transcriptomic hourglass in plant embryogenesis," *Nature*, vol. 490, no. 7418, pp. 98–101, 2012.

[7] M. S. H. Ko, "Embryogenomics of pre-implantation mammalian development: current status," *Reproduction, Fertility and Development*, vol. 16, no. 1-2, pp. 79–85, 2004.

[8] J. Kaňka, "Gene expression and chromatin structure in the preimplantation embryo," *Theriogenology*, vol. 59, no. 1, pp. 3–19, 2003.

[9] R. Vassena, S. Boué, E. González-Roca et al., "Waves of early transcriptional activation and pluripotency program initiation during human preimplantation development," *Development*, vol. 138, no. 17, pp. 3699–3709, 2011.

[10] A.-E. Saliba, A. J. Westermann, S. A. Gorski, and J. Vogel, "Single-cell RNA-seq: advances and future challenges," *Nucleic Acids Research*, vol. 42, no. 14, pp. 8845–8860, 2014.

[11] Z. Xue, K. Huang, C. Cai et al., "Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing," *Nature*, vol. 500, no. 7464, pp. 593–597, 2013.

[12] L. Yan, M. Yang, H. Guo et al., "Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells," *Nature Structural and Molecular Biology*, vol. 20, no. 9, pp. 1131–1139, 2013.

[13] P. Langfelder and S. Horvath, "WGCNA: an R package for weighted correlation network analysis," *BMC Bioinformatics*, vol. 9, article 559, 2008.

[14] S. B. Carroll, "Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution," *Cell*, vol. 134, no. 1, pp. 25–36, 2008.

[15] M. Sémon and L. Duret, "Evolutionary origin and maintenance of coexpressed gene clusters in mammals," *Molecular Biology and Evolution*, vol. 23, no. 9, pp. 1715–1723, 2006.

[16] A. de la Fuente, "From 'differential expression' to 'differential networking'—identification of dysfunctional regulatory networks in diseases," *Trends in Genetics*, vol. 26, no. 7, pp. 326–333, 2010.

[17] X. A. Chang, S. A. Liu, Y.-T. Yu, Y.-X. Li, and Y.-Y. Li, "Identifying modules of coexpressed transcript units and their

organization of *Saccharopolyspora erythraea* from time series gene expression profiles," *PLoS ONE*, vol. 5, no. 8, Article ID e12126, 2010.

[18] X. Chang, L. L. Shi, F. Gao et al., "Genomic and transcriptome analysis revealing an oncogenic functional module in meningiomas," *Neurosurgical Focus*, vol. 35, no. 6, article E3, 2013.

[19] C. R. Farber, "Identification of a gene module associated with BMD through the integration of network analysis and genome-wide association data," *Journal of Bone and Mineral Research*, vol. 25, no. 11, pp. 2359–2367, 2010.

[20] C. G. J. Saris, S. Horvath, P. W. J. van Vught et al., "Weighted gene co-expression network analysis of the peripheral blood from amyotrophic lateral sclerosis patients," *BMC Genomics*, vol. 10, p. 405, 2009.

[21] 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, no. 1, pp. 44–57, 2009.

[22] H. J. M. Aarts, E. H. M. Jacobs, G. van Willigen, N. H. Lubsen, and J. G. G. Schoenmakers, "Different evolution rates within the lens-specific β-crystallin gene family," *Journal of Molecular Evolution*, vol. 28, no. 4, pp. 313–321, 1989.

[23] R. Sanges, Y. Hadzhiev, M. Gueroult-Bellone et al., "Highly conserved elements discovered in vertebrates are present in non-syntenic loci of tunicates, act as enhancers and can be transcribed during development," *Nucleic Acids Research*, vol. 41, no. 6, pp. 3600–3618, 2013.

[24] E. M. Thompson, "Chromatin structure and gene expression in the preimplantation mammalian embryo," *Reproduction Nutrition Development*, vol. 36, no. 6, pp. 619–635, 1996.

[25] L. Li, X. Lu, and J. Dean, "The maternal to zygotic transition in mammals," *Molecular Aspects of Medicine*, vol. 34, no. 5, pp. 919–938, 2013.

[26] T. Hamatani, M. S. Ko, M. Yamada et al., "Global gene expression profiling of preimplantation embryos," *Human Cell*, vol. 19, no. 3, pp. 98–117, 2006.

[27] J. Mensch, F. Serra, N. J. Lavagnino, H. Dopazo, and E. Hasson, "Positive selection in nucleoporins challenges constraints on early expressed genes in *Drosophila* development," *Genome Biology and Evolution*, vol. 5, no. 11, pp. 2231–2241, 2013.

[28] A. Galán, D. Montaner, M. E. Póo et al., "Functional genomics of 5- to 8-cell stage human embryos by blastomere single-cell cDNA analysis," *PLoS ONE*, vol. 5, no. 10, Article ID e13615, 2010.

[29] Z. Jiang, J. Sun, H. Dong et al., "Transcriptional profiles of bovine in vivo pre-implantation development," *BMC Genomics*, vol. 15, no. 1, article 756, 2014.

[30] J. P. Demuth and M. W. Hahn, "The life and death of gene families," *BioEssays*, vol. 31, no. 1, pp. 29–39, 2009.

[31] B. Piasecka, P. Lichocki, S. Moretti, S. Bergmann, and M. Robinson-Rechavi, "The hourglass and the early conservation models—co-existing patterns of developmental constraints in vertebrates," *PLoS Genetics*, vol. 9, no. 4, Article ID e1003476, 2013.

[32] G. A. Wray, "The evolutionary significance of cis-regulatory mutations," *Nature Reviews Genetics*, vol. 8, no. 3, pp. 206–216, 2007.

[33] P. G. Engström, D. Fredman, and B. Lenhard, "Ancora: a web resource for exploring highly conserved noncoding elements and their association with developmental regulatory genes," *Genome Biology*, vol. 9, no. 2, article R34, 2008.

[34] A. C. Burke, C. E. Nelson, B. A. Morgan, and C. Tabin, "Hox genes and the evolution of vertebrate axial morphology," *Development*, vol. 121, no. 2, pp. 333–346, 1995.

[35] Q. Deng, D. Ramsköld, B. Reinius, and R. Sandberg, "Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells," *Science*, vol. 343, no. 6167, pp. 193–196, 2014.

[36] F. H. Biase, X. Y. Cao, and S. Zhong, "Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell RNA sequencing," *Genome Research*, vol. 24, no. 11, pp. 1787–1796, 2014.

[37] R. Edgar, M. Domrachev, and A. E. Lash, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository," *Nucleic Acids Research*, vol. 30, no. 1, pp. 207–210, 2002.

[38] G. P. Wagner, K. Kin, and V. J. Lynch, "Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples," *Theory in Biosciences*, vol. 131, no. 4, pp. 281–285, 2012.

[39] B. Zhang and S. Horvath, "A general framework for weighted gene co-expression network analysis," *Statistical Applications in Genetics and Molecular Biology*, vol. 4, article 17, 2005.

[40] A. M. Yip and S. Horvath, "Gene network interconnectedness and the generalized topological overlap measure," *BMC Bioinformatics*, vol. 8, article 22, 2007.

[41] P. Langfelder, B. Zhang, and S. Horvath, "Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R," *Bioinformatics*, vol. 24, no. 5, pp. 719–720, 2008.

[42] D. Smedley, S. Haider, B. Ballester et al., "BioMart—biological queries made easy," *BMC Genomics*, vol. 10, article 22, 2009.

[43] P. Flicek, M. R. Amode, D. Barrell et al., "Ensembl 2014," *Nucleic Acids Research*, vol. 42, no. 1, pp. D749–D755, 2014.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 483726, 7 pages
http://dx.doi.org/10.1155/2015/483726

Research Article

Statistical Genomic Approach Identifies Association between FSHR Polymorphisms and Polycystic Ovary Morphology in Women with Polycystic Ovary Syndrome

Tao Du,[1] Yu Duan,[1] Kaiwen Li,[2] Xiaomiao Zhao,[1] Renmin Ni,[1] Yu Li,[1] and Dongzi Yang[1]

[1]*Department of Gynecology & Obstetrics, Sun Yat-sen Memorial Hospital, Sun Yat-sen University, 107 West Yanjiang Road, Guangzhou 510120, China*
[2]*Department of Urology, Sun Yat-sen Memorial Hospital, Sun Yat-sen University, Guangzhou, China*

Correspondence should be addressed to Dongzi Yang; yangdz@mail.sysu.edu.cn

Received 12 January 2015; Accepted 13 February 2015

Academic Editor: Xiao Chang

Background. Single-nucleotide polymorphisms (SNPs) in the follicle stimulating hormone receptor (FSHR) gene are associated with PCOS. However, their relationship to the polycystic ovary (PCO) morphology remains unknown. This study aimed to investigate whether PCOS related SNPs in the FSHR gene are associated with PCO in women with PCOS. *Methods*. Patients were grouped into PCO ($n = 384$) and non-PCO ($n = 63$) groups. Genomic genotypes were profiled using Affymetrix human genome SNP chip 6. Two polymorphisms (rs2268361 and rs2349415) of FSHR were analyzed using a statistical approach. *Results*. Significant differences were found in the allele distributions of the GG genotype of rs2268361 between the PCO and non-PCO groups (27.6% GG, 53.4% GA, and 19.0% AA versus 33.3% GG, 36.5% GA, and 30.2% AA), while no significant differences were found in the allele distributions of the GG genotype of rs2349415. When rs2268361 was considered, there were statistically significant differences of serum follicle stimulating hormone, estradiol, and sex hormone binding globulin between genotypes in the PCO group. In case of the rs2349415 SNP, only serum sex hormone binding globulin was statistically different between genotypes in the PCO group. *Conclusions*. Functional variants in FSHR gene may contribute to PCO susceptibility in women with PCOS.

1. Introduction

Polycystic ovary syndrome (PCOS) affects about 4%–12% women within reproductive age [1]. It is characterized by hyperandrogenism, menstrual irregularity, infertility, and polycystic ovarian (PCO) morphology [2, 3]. Although the term PCO is used, not all patients have polycystic ovaries [4, 5]. Ultrasonography is the only way to detect PCO, the results of which may be related to the experience of doctors, the definition of devices, and the detected time of the menstruation period.

The pathogenesis of PCO and PCOS is still not completely understood. Recent studies report that PCOS may be associated with single-nucleotide polymorphisms (SNP). Our previous GWAS identified two SNPs (rs2268361 and rs2349415) in the follicle stimulating hormone receptor (FSHR) gene associated with PCOS in the Han Chinese population [6]. FSHR is specifically expressed by the granulosa cells of the ovary and bound by its ligand serum follicle stimulating hormone (FSH), which mediates the functions of FSH and plays important roles in the folliculogenesis by affecting the maturation and development of follicles [7, 8]. If FSHR is null, females would be sterile, with small ovaries and blocked follicular development [9].

While rs2268361 and rs2349415 SNPs are confirmed to be associated with PCOS, their relationship to the PCO morphology remains unknown. The present study performs a case-control study in the Han Chinese women with PCOS in South China and investigates the association between the two FSHR gene SNPs and the PCO morphology.

2. Methods

2.1. Subjects. Patients with PCOS were recruited by Reproductive Medicine Center, Sun Yat-sen Memorial Hospital of Sun Yat-sen University from January 2010 to August 2011. All subjects were unrelated Han Chinese from south China. The study was approved by the local Medical Ethical Committee and all participants have written consented for the purpose of the study.

The patients were selected based on the 2003 Rotterdam Diagnostic Criteria [2], requiring the presence of any two of the following three criteria: oligoovulation and/or anovulation; clinical and/or biochemical signs of hyperandrogenism; and polycystic ovary morphology (PCO) and exclusion of other etiologies. Patients were excluded if they: (a) used hormonal medication, including oral contraceptives, for at least 3 months prior to starting the study; (b) had other endocrine disorders or known causes of hyperandrogenism, such as thyroid abnormalities, congenital adrenal hyperplasia, Cushing syndrome, and androgen-secreting tumors; (c) had previous ovary-related surgery, chemotherapy, or radiotherapy; and (d) had pregnancy. Patients were classified into PCO or non-PCO groups according to the ovary appearance on ultrasound. A finding of PCO was determined when 12 or more follicles measuring 2–9 mm in diameter were scanned in either ovary or the ovarian volume was greater than 10 mL [2].

2.2. Clinical Measurements. Detailed medical information of included patients was collected, including the parameters age, body mass index (BMI), waist-hip ratio (WHR), transvaginal ultrasound results, modified Ferriman-Gallwey score (mF-G score) [10], and acne score [11]. Transvaginal ultrasonography was performed on cycle days 3–5 (or between days 3 and 5 after a progestin-induced withdrawal bleeding for the amenorrhea patients).

2.3. Biochemical Measurements. On cycle days 3–5 (or between days 3 and 5 after a progestin-induced withdrawal bleeding for the amenorrhea patients), venous blood samples were taken between 8 a.m. and 9 a.m. after a 12 h overnight fast. The serum hormone levels, including follicle stimulating hormone (FSH), luteinizing hormone (LH), prolactin (PRL), estradiol (E2), total testosterone (TT), androstenedione (A2), free testosterone (FT), dehydroepiandrosterone sulfate (DHEA-S), and estrone (E1), were measured by the chemiluminescence immunoassays Access 2 (Beckman, Fullerton, CA). Sex hormone binding globulin (SHBG) and 17-hydroxyprogesterone (17-OHP) were measured using the Access 2 ELISA kit (Beckman, Fullerton, CA). Serum cholesterol (CHOL), triglycerides (TG), high-density lipoprotein-cholesterol (HDL-C), and low-density lipoprotein (LDL-C) levels were measured using an enzymatic calorimetric method (Hitachi 7600). Plasma glucose was detected by the glucose oxidase method (Hitachi 7600), and insulin was detected by chemiluminescence immunometric assay and commercial kit (Immulite2000 Analyzer; CPC). The homeostatic model assessment for insulin resistance (HOMA-IR), insulin sensitivity (HOMA-IS), and β (HOMA-β) was calculated as follows: (fasting insulin [mIU/L] × fasting glucose [mmol/L])/22.5, 1/(fasting insulin [mIU/L] × fasting glucose [mmol/L]), 20 × fasting insulin [mIU/L]/(fasting glucose [mmol/L] − 3.5) (%), respectively.

2.4. Genotyping. The detailed procedures of genotyping were referred to in the GWAS study [6]. EDTA anticoagulated venous blood samples were collected from all patients. Genomic DNA was extracted from peripheral blood lymphocytes by using Flexi Gene DNA kits (Qiagen, Germany) and was diluted to working concentrations of 50 ng/μL for genome-wide genotyping and 15–20 ng/μL for the validation study. Genotyping was performed using Affymetrix human genome SNP chip 6. The two polymorphisms (rs2268361 and rs2349415) of FSHR were selected according to the previous criteria [6, 12].

2.5. Statistical Analysis. Statistical analysis was performed using Stata/SE 12.0 (StataCorp LP, TX, USA). Continuous variables were compared using Student's t-test or ANOVA for parametric data and Mann-Whitney test or Kruskal-Wallis test for nonparametric data. Categorical data were compared using Chi-squared test. Genotype frequencies were tested for Hardy-Weinberg equilibrium. Linkage disequilibrium (LD) was estimated with D' and r^2 on SHEsis (http://analysis.bio-x.cn/myAnalysis.php). All tests were considered statistically significant at $P < 0.05$.

3. Results

A total of 447 women with PCOS met the eligible criteria for the study. The mean age was 27.13 ± 3.88. Among them, 384 women had PCO, but 63 did not. The baseline characteristics of the subjects were shown in Table 1. There was no statistically significant difference in age, BMI, WHR, acne score, mF-G score, HOMA-IR, HOMA-ISI, HOMA-β%, serum CHOL, TG, HDL-C, FSH, LH, PRL, E2, TT, A2, DHEA-S, E1, and 17-OHP levels between the PCO and non-PCO groups. However, serum LDL-C and FT were significantly higher in PCO group compared with non-PCO group, while serum SHBG level was significantly lower (P = 0.02, 0.04, and 0.04, resp.) (Table 1).

The allele and genotype frequencies for rs2268361 and rs2349415 SNPs in PCOS women with or without PCO were shown in Table 2. The genotype frequencies of the two SNPs were in accordance with the Hardy-Weinberg equilibrium (P = 0.5763 and 0.6452, resp.). PCOS women with genotype G/A had significantly higher risk of suffering from PCO (P = 0.03). Therefore, polymorphisms of rs2268361 were associated with PCO morphology. However, there were no statistical difference in the allele frequencies of rs2268361 and in the allele or genotype frequencies of rs2349415 between PCO group and non-PCO group. The extent of linkage disequilibrium in pairwise combinations of the two SNPs was calculated. Among all the PCOS women, linkage disequilibrium (D') between the rs2268361 and rs2349415 SNPs of FSHR gene was 0.610 (r^2 = 0.113), indicating a low linkage of the two polymorphisms.

Table 1: Demographic characteristics and endocrine profiles of PCOS women with or without polycystic ovary morphology.

	Non-PCO	PCO	P value*
Case number	63	384	—
Age, year	26.86 ± 3.65	27.18 ± 3.92	0.55
BMI, kg/m^2	23.28 ± 9.47	22.83 ± 4.70	0.74
WHR	0.84 ± 0.21	0.82 ± 0.07	0.67
Acne score	1 (0–4)	0 (0–5)	**0.049**
mF-G score	6 (0–22)	4 (0–35)	0.35
CHOL, mmol/L	4.73 ± 0.75	4.97 ± 0.87	0.08
TG, mmol/L	1.26 ± 0.73	1.32 ± 0.84	0.67
HDL-C, mmol/L	1.62 ± 0.48	1.57 ± 0.44	0.51
LDL-C, mmol/L	2.62 ± 0.67	2.92 ± 0.82	**0.02**
HOMA-IR	2.19 ± 1.72	2.34 ± 1.98	0.59
HOMA-ISI	0.03 ± 0.02	0.03 ± 0.03	0.71
HOMA-β%	135.96 ± 115.58	131.87 ± 130.32	0.82
FSH, IU/L	6.68 ± 2.34	6.46 ± 2.33	0.51
LH, IU/L	10.24 ± 7.80	10.88 ± 6.99	0.53
PRL, ng/mL	16.00 ± 10.47	16.63 ± 11.20	0.70
E2, ng/L	84.62 ± 93.91	71.45 ± 75.47	0.34
TT, nmol/L	2.41 ± 1.02	2.40 ± 1.31	0.99
A2, ng/mL	4.80 ± 2.22	5.22 ± 2.46	0.37
FT, pg/mL	3.65 ± 3.01	4.67 ± 5.68	**0.04**
DHEA-S, ng/mL	2245.25 ± 1009.03	2159.46 ± 1007.76	0.54
E1, pg/mL	101.77 ± 37.95	117.18 ± 83.34	0.23
SHBG, nmol/L	85.10 ± 72.80	64.20 ± 55.802	**0.04**
17-OHP, ng/mL	1.17 ± 1.00	1.27 ± 0.87	0.43

*Significant difference was in bold.
PCO: polycystic ovary; BMI: body mass index; WHR: waist-hip ratio; mF-G score: modified Ferriman-Gallwey score; CHOL: cholesterol; TG: triglycerides; HDL-C: high-density lipoprotein-cholesterol; LDL-C: low-density lipoprotein-cholesterol; HOMA-IR/HOMA-ISI/HOMA-β%: homeostatic model assessment for insulin resistance/insulin sensitivity/β; FSH: follicle stimulating hormone; LH: luteinizing hormone; PRL: prolactin; E2: estradiol; TT: total testosterone; A2: androstenedione; FT: free testosterone; DHEA-S: dehydroepiandrosterone sulfate; E1: estrone; SHBG: sex hormone binding globulin; 17-OHP: 17-hydroxyprogesterone.

Table 2: Genotype distribution and allele frequency of FSHR polymorphisms in PCO women with or without polycystic ovary.

	rs2268361		P value*	H-W test		rs2349415		P value	H-W test
	PCO	Non-PCO				PCO	Non-PCO		
GG	106	21	**0.03**	0.5763	CC	202	39	0.38	0.6452
GA	205	23			CT	156	21		
AA	73	19			TT	26	3		
G	417	65	0.63		C	560	99	0.18	
A	351	61			T	208	27		

*Significant difference was in bold.
PCO: polycystic ovary; H-W test: Hardy-Weinberg equilibrium test.

Further analyses of phenotypes were performed for the rs2268361 and rs2349415 polymorphisms in both the PCO group and non-PCO group (Tables 3 and 4). When the rs2268361 SNP was considered, there were statistically significant differences in the levels of serum FSH ($P = 0.02$), E2 ($P = 0.03$), and SHBG ($P = 0.03$) between genotypes in the PCO women. However, no serum hormones were significantly different between genotypes in the non-PCO group. In case of the rs2349415 SNP, only the level of serum SHBG ($P = 0.02$) was statistically different between genotypes in the PCO group.

4. Discussion

Polycystic ovary syndrome is a complex hereditary endocrine disorder with complicated clinical manifestations. Higher FT and lower SHBG in serum were detected in PCO group compared with non-PCO group, indicating that more actions

Table 3: FSHR polymorphisms versus endocrine parameters in PCO patients.

	rs2268361			*P* value*	rs2349415			*P* value*
	GG	GA	AA		CC	CT	TT	
Case number	106	205	73	—	202	156	26	—
Age, year	26.82 ± 3.95	27.45 ± 3.67	26.92 ± 4.52	0.33	27.19 ± 4. 16	26.96 ± 3.70	28.42 ± 3.15	0.21
BMI, kg/m^2	22.67 ± 4.75	22.90 ± 4.91	22.89 ± 3.98	0.92	22.81 ± 4.32	22.83 ± 5.12	23.00 ± 5.02	0.98
WHR	0.82 ± 0.08	0.83 ± 0.08	0.82 ± 0.07	0.82	0.82 ± 0.08	0.82 ± 0.07	0.82 ± 0.06	0.99
Acne score	1 (0–4)	0 (0–5)	0 (0–5)	0.07	0 (0–4)	0 (0–5)	1 (0–4)	0.63
mF-G score	5 (0–27)	4 (0–35)	3.5 (0–22)	0.12	4 (0–35)	4 (0–29)	5.5 (0–16)	0.33
CHOL, mmol/L	4.89 ± 0.82	4.99 ± 0.88	5.04 ± 0.90	0.54	4.93 ± 0.89	5.01 ± 0.88	5.03 ± 0.69	0.69
TG, mmol/L	1.30 ± 0.82	1.37 ± 0.91	1.22 ± 0.62	0.44	1.27 ± 0.89	1.27 ± 0.89	1.39 ± 0.76	0.45
HDL-C, mmol/L	1.60 ± 0.42	1.58 ± 0.46	1.51 ± 0.42	0.42	1.53 ± 0.40	1.62 ± 0.49	1.56 ± 0.41	0.21
LDL-C, mmol/L	2.81 ± 0.76	2.93 ± 0.84	3.04 ± 0.84	0.21	2.94 ± 0.84	2.86 ± 0.83	3.02 ± 0.59	0.58
HOMA-IR	2.09 ± 1.78	2.45 ± 2.10	2.39 ± 1.91	0.30	2.29 ± 1.75	2.44 ± 2.32	2.12 ± 1.38	0.66
HOMA-ISI	0.03 ± 0.02	0.03 ± 0.03	0.03 ± 0.02	0.82	0.03 ± 0.03	0.03 ± 0.02	0.03 ± 0.03	0.97
HOMA-β%	120.98 ± 156.82	135.70 ± 121.63	136.74 ± 111.78	0.61	125.02 ± 125.61	140.06 ± 137.64	135.56 ± 122.83	0.56
FSH, IU/L	6.58 ± 2.16	6.65 ± 2.50	5.72 ± 1.90	**0.02**	6.31 ± 2.12	6.58 ± 2.65	6.82 ± 1.63	0.41
LH, IU/L	11.65 ± 7.47	10.62 ± 6.73	10.42 ± 7.02	0.41	10.40 ± 6.62	11.26 ± 7.17	12.11 ± 8.48	0.36
PRL, ng/mL	16.05 ± 10.02	17.46 ± 12.46	15.04 ± 8.60	0.28	16.11 ± 10.44	16.88 ± 11.78	19.19 ± 13.29	0.45
E2, ng/L	72.88 ± 67.09	63.21 ± 63.91	92.84 ± 106.84	**0.03**	71.16 ± 76.83	71.96 ± 64.84	70.74 ± 115.86	0.99
TT, nmol/L	2.43 ± 1.24	2.41 ± 1.46	2.35 ± 0.94	0.92	2.36 ± 1.07	2.50 ± 1.61	2.09 ± 0.98	0.29
A2, ng/mL	5.03 ± 2.20	5.37 ± 2.62	5.07 ± 2.42	0.71	5.18 ± 2.29	5.43 ± 2.80	4.34 ± 1.50	0.39
FT, pg/mL	4.36 ± 5.62	5.04 ± 6.34	4.11 ± 3.49	0.41	4.76 ± 5.81	4.78 ± 5.81	2.92 ± 2.04	0.37
DHEA-S, ng/mL	2124.35 ± 1049.61	2178.40 ± 989.35	2158.74 ± 1009.01	0.91	2165.36 ± 924.47	2206.65 ± 1103.63	1774.35 ± 1013.97	0.18
E1, pg/mL	124.49 ± 71.96	126.03 ± 98.92	87.97 ± 57.90	0.29	119.74 ± 73.46	116.73 ± 102.42	105.78 ± 72.28	0.91
SHBG, nmol/L	76.09 ± 72.20	61.64 ± 51.50	53.97 ± 33.05	**0.03**	56.46 ± 35.97	73.13 ± 73.71	71.07 ± 48.71	**0.02**
17-OHP, ng/mL	1.25 ± 1.04	1.26 ± 0.81	1.33 ± 0.76	0.82	1.33 ± 0.99	1.22 ± 0.74	1.13 ± 0.50	0.44

*Significant difference was in bold.

PCO: polycystic ovary; BMI: body mass index; WHR: waist-hip ratio; mF-G score: modified Ferriman-Gallwey score; CHOL: cholesterol; TG: triglycerides; HDL-C: high-density lipoprotein-cholesterol; LDL-C: low-density lipoprotein-cholesterol; HOMA-IR/HOMA-ISI/HOMA-β%: homeostatic model assessment for insulin resistance/insulin sensitivity/β; FSH: follicle stimulating hormone; LH: luteinizing hormone; PRL: prolactin; E2: estradiol; TT: total testosterone; A2: androstenedione; FT: free testosterone; DHEA-S: dehydroepiandrosterone sulfate; E1: estrone; SHBG: sex hormone binding globulin; 17-OHP: 17-hydroxyprogesterone.

Table 4: FSHR polymorphisms versus endocrine parameters in non-PCO patients*.

	rs2268361			P value
	GG	GA	AA	
Case number	21	23	19	—
Age, year	27.05 ± 3.57	26.69 ± 4.76	26.84 ± 2.00	0.95
BMI, kg/m^2	21.98 ± 4.13	24.89 ± 14.51	23.73 ± 4.93	0.64
WHR	0.81 ± 0.05	0.88 ± 0.34	0.81 ± 0.07	0.56
Acne score	1 (0–3)	1 (0–3)	1 (0–4)	0.63
mF-G score	7 (0–22)	5 (0–21)	4 (0–16)	0.55
CHOL, mmol/L	4.52 ± 0.77	4.68 ± 0.62	5.05 ± 0.85	0.18
TG, mmol/L	1.42 ± 1.09	1.22 ± 0.47	1.10 ± 0.35	0.50
HDL-C, mmol/L	1.76 ± 0.73	1.54 ± 0.24	1.54 ± 0.30	0.34
LDL-C, mmol/L	2.43 ± 0.77	2.54 ± 0.53	3.00 ± 0.63	0.06
HOMA-IR	2.66 ± 1.92	1.96 ± 1.94	1.96 ± 0.94	0.36
HOMA-ISI	0.03 ± 0.02	0.04 ± 0.03	0.03 ± 0.03	0.40
HOMA-β%	172.07 ± 121.06	120.98 ± 134.90	113.68 ± 65.57	0.25
FSH, IU/L	6.75 ± 1.91	7.25 ± 2.66	5.82 ± 2.17	0.18
LH, IU/L	13.73 ± 10.21	8.87 ± 6.39	8.29 ± 5.21	0.07
PRL, ng/mL	12.61 ± 5.76	19.79 ± 12.53	13.64 ± 9.45	0.06
E2, ng/L	83.42 ± 77.05	85.15 ± 74.46	84.97 ± 132.92	1.00
TT, nmol/L	2.14 ± 1.07	2.75 ± 0.73	2.26 ± 1.19	0.11
A2, ng/mL	4.08 ± 2.16	5.44 ± 2.39	4.56 ± 1.89	0.33
FT, pg/mL	3.32 ± 2.79	3.85 ± 3.22	3.80 ± 3.14	0.82
DHEA-S, ng/mL	2207.05 ± 1064.78	2523.01 ± 1031.51	1916.65 ± 842.80	0.17
E1, pg/mL	100.18 ± 45.10	101.58 ± 40.50	103.79 ± 31.81	0.99
SHBG, nmol/L	89.91 ± 81.37	83.32 ± 75.70	80.99 ± 58.88	0.93
17-OHP, ng/mL	1.23 ± 1.27	1.26 ± 1.00	0.94 ± 0.41	0.65

*No analysis was performed for rs2349415 SNP because the size of genotype pattern TT was too small.
PCO: polycystic ovary; BMI: body mass index; WHR: waist-hip ratio; mF-G score: modified Ferriman-Gallwey score; CHOL: cholesterol; TG: triglycerides; HDL-C: high-density lipoprotein-cholesterol; LDL-C: low-density lipoprotein-cholesterol; HOMA-IR/HOMA-ISI/HOMA-β%: homeostatic model assessment for insulin resistance/insulin sensitivity/β; FSH: follicle stimulating hormone; LH: luteinizing hormone; PRL: prolactin; E2: estradiol; TT: total testosterone; A2: androstenedione; FT: free testosterone; DHEA-S: dehydroepiandrosterone sulfate; E1: estrone; SHBG: sex hormone binding globulin; 17-OHP: 17-hydroxyprogesterone.

of androgen might be associated with folliculogenesis. Polymorphisms (rs2268361, but not rs2349415) in FSHR gene were found to be associated with the presentation of PCO.

FSHR, with its ligand FSH, is one of the main factors affecting PCOS. It is reported that FSHR is highly expressed in granulosa and ovarian surface epithelium cells from stimulated follicles of women with PCOS [13, 14]. It is a member of G protein-coupled receptors family including three domains, which are extracellular ligand-binding, transmembrane, and intracellular domains [12, 15]. FSHR gene is located on chromosome 2p21 and comprises 10 exons and 9 introns [16]. Overexpression of FSHR may lead to enhanced proliferation in ovarian surface epithelial cells, therefore generating more premature follicles [17]. In addition, the level of FSH is also controlled by FSHR and aberrant FSHR may affect ovarian folliculogenesis.

As several PCOS susceptibility genes including INS VNTR, TCF7L2, and PPARG [18–20], polymorphisms or SNPs for the FSHR gene may also be associated with PCOS. Many studies have focused on investigating the most controversial SNPs in FSHR gene, which are Thr307Ala and Asn680Ser polymorphisms in exon 10 [21]. The latest systematic review reveals that there exist no associations between the two SNPs and PCOS [21]. However, the sample size of this meta-analysis is relatively small, which may not have abundant statistical power in assessing the role of FSHR polymorphisms in the development of PCOS [21]. Our previous study analyzed genome-wide association data from two cohorts of Han Chinese in large sizes with 744 PCOS cases and 895 controls in GWAS 1 and 1,510 PCOS cases and 2,016 controls in GWAS 2. The study also followed up significantly associated signals identified in the combined results of GWAS 1 and 2 in a total of 8,226 cases and 7,578 controls. Among the top ten significant signals, two SNP signals (rs2268361 and rs2349415) located in the FSHR gene were identified to be associated with PCOS [6].

Women with PCOS can be diagnosed without PCO morphology according to the 2003 Rotterdam Criteria [2]. PCO appearance under ultrasound is only detected in 62.7% to 91% PCOS women [5, 22]. There are essentially two anomalies affecting folliculogenesis [23]: excessive small follicular growth and the defective selection of a dominant

follicle. Androgens are primarily responsible for promoting the growth of small follicles [24, 25], and adequate action of FSH is essential for selecting a dominant follicle. When the secretion or function of FSH is inhibited, PCO will be presented [26]. However, it is a common belief that women with PCOS have serum FSH levels at the same range as normal women in early follicular phase [27]. Therefore, the function of FSH via combining to FSHR may be the most important. In the present study, significantly higher FT and lower SHBG in PCO than non-PCO patients might initially promote the growth of small follicles. One of two analyzed SNP (rs2268361) in the FSHR gene was significantly associated with PCO. Patients with G/A at rs2268361 had high risk of presenting PCO. SNP at rs2268361 in the FSHR gene might result in changing the reaction to FSH stimulation, while FSH was similar between PCO and non-PCO groups. The function of FSHR in PCO-group women with G/A at rs2268361 might be defected comparing with the other two patterns, while serum estradiol was also significantly lower in G/A pattern. No association between SNP (rs2349415) in the FSHR gene and PCO morphology was detected.

The main limitation of this study is the relative small sample size. Further studies in larger populations are needed to come up with more supporting data. Furthermore, because of differences in ethnicity and sample size, effects of the SNPs in the FSHR genes could account for some conflicting results. However, to our knowledge, this study is the first to identify the association between SNPs in the FSHR gene and PCO morphology in PCOS women. It can bring a new insight into the cause and diagnosis of PCO.

5. Conclusions

This is the first study to identify the association between SNP (rs2268361) in the FSHR gene and PCO morphology in PCOS women. PCOS women with G/A at rs2268361 SNP will have higher risk of appearing PCO. There is no association between SNP (rs2349415) in the FSHR gene and PCO appearance. Further studies with populations in other ethnicities and larger size are needed to help clarify the PCO susceptible SNPs in the FSHR gene.

Abbreviations

A2:	Androstenedione
BMI:	Body mass index
CHOL:	Cholesterol
DHEA-S:	Dehydroepiandrosterone sulfate
E1:	Estrone
E2:	Estradiol
FSH:	Follicle stimulating hormone
FSHR:	Follicle stimulating hormone receptor
FT:	Free testosterone
HDL-C:	High-density lipoprotein-cholesterol
HOMA-IR/HOMA-ISI/HOMA-β%:	Homeostatic model assessment for insulin resistance/insulin sensitivity/β
LDL-C:	Low-density lipoprotein-cholesterol
LH:	Luteinizing hormone
mF-G score:	Modified Ferriman-Gallwey score
PCO:	Polycystic ovary
PCOS:	Polycystic ovary syndrome
PRL:	Prolactin
SHBG:	Sex hormone binding globulin
SNP:	Single nucleotide polymorphism
TG:	Triglycerides
TT:	Total testosterone
WHR:	Waist-hip ratio
17-OHP:	17-Hydroxyprogesterone.

Conflict of Interests

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

Authors' Contribution

Dongzi Yang participated in the design, recruitment, and evaluation of subjects and helped in data analysis and revising the paper critically for important intellectual content. Tao Du and Yu Duan helped Dongzi Yang in recruiting and evaluating subjects, read, revised, and approved the final paper. Xiaomiao Zhao, Yu Li and Renmin Ni participated in the laboratory aspects of the study and took part in data analysis, interpretation, revision, and approval of the paper. Kaiwen Li made substantial contributions to the conception, design, analysis, and interpretation of the data and wrote the paper. All authors read and approved the final paper. Tao Du, Yu Duan, and Kaiwen Li contributed equally to this work as co-first authors.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (81370680, 81320004, 81472382), the specialized research fund for the doctoral program of Chinese Ministry of Education (20130171130009), the Natural Science Foundation of Key Research Project of Guangdong Province (2013020012660). The funders had no role in the study design, in the collection, in analysis and interpretation of data, in the writing of the paper, and in the decision to submit the paper for publication. No others made contributions to this work.

References

[1] R. Azziz, K. S. Woods, R. Reyna, T. J. Key, E. S. Knochenhauer, and B. O. Yildiz, "The prevalence and features of the polycystic ovary syndrome in an unselected population," *The Journal of Clinical Endocrinology and Metabolism*, vol. 89, no. 6, pp. 2745–2749, 2004.

[2] The Rotterdam ESHRE/ASRM-Sponsored PCOS Consensus Workshop Group, "Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary

syndrome (PCOS)," *Human Reproduction*, vol. 19, no. 1, pp. 41–47, 2004.

[3] M. O. Goodarzi, D. A. Dumesic, G. Chazenbalk, and R. Azziz, "Polycystic ovary syndrome: etiology, pathogenesis and diagnosis," *Nature Reviews Endocrinology*, vol. 7, no. 4, pp. 219–231, 2011.

[4] J. J. Kim, K. R. Hwang, Y. M. Choi et al., "Complete phenotypic and metabolic profiles of a large consecutive cohort of untreated Korean women with polycystic ovary syndrome," *Fertility and Sterility*, vol. 101, no. 5, pp. 1424–1430, 2014.

[5] N. M. Clark, A. J. Podolski, E. D. Brooks et al., "Prevalence of polycystic ovary syndrome phenotypes using updated criteria for polycystic ovarian morphology: an assessment of over 100 consecutive women self-reporting features of polycystic ovary syndrome," *Reproductive Sciences*, vol. 21, no. 8, pp. 1034–1043, 2014.

[6] Y. Shi, H. Zhao, Y. Cao et al., "Genome-wide association study identifies eight new risk loci for polycystic ovary syndrome," *Nature Genetics*, vol. 44, no. 9, pp. 1020–1025, 2012.

[7] R. González-Fernández, Ó. Peña, J. Hernández, P. Martín-Vasallo, A. Palumbo, and J. Ávila, "Patients with endometriosis and patients with poor ovarian reserve have abnormal follicle-stimulating hormone receptor signaling pathways," *Fertility and Sterility*, vol. 95, no. 7, pp. 2373–2378, 2011.

[8] B.-H. Gu, J.-M. I. Park, and K.-H. Baek, "Genetic variations of follicle stimulating hormone receptor are associated with polycystic ovary syndrome," *International Journal of Molecular Medicine*, vol. 26, no. 1, pp. 107–112, 2010.

[9] L. Sun, Y. Peng, A. C. Sharrow et al., "FSH Directly Regulates Bone Mass," *Cell*, vol. 125, no. 2, pp. 247–260, 2006.

[10] B. O. Yildiz, S. Bolour, K. Woods, A. Moore, and R. Azziz, "Visually scoring hirsutism," *Human Reproduction Update*, vol. 16, no. 1, pp. 51–64, 2009.

[11] G. Plewig, "Classification of acne vulgaris," in *Acne: Morphogenesis and Treatment*, Springer, Berlin, Germany, 1975.

[12] M. Simoni, J. Gromoll, and E. Nieschlag, "The follicle-stimulating hormone receptor: biochemistry, molecular biology, physiology, and pathophysiology," *Endocrine Reviews*, vol. 18, no. 6, pp. 739–773, 1997.

[13] S. Catteau-Jonard, S. P. Jamin, A. Leclerc, J. Gonzalès, D. Dewailly, and N. Di Clemente, "Anti-Mullerian hormone, its receptor, FSH receptor, and androgen receptor genes are overexpressed by granulosa cells from stimulated follicles in women with polycystic ovary syndrome," *Journal of Clinical Endocrinology and Metabolism*, vol. 93, no. 11, pp. 4456–4461, 2008.

[14] M. Durlej, K. Knapczyk-Stwora, M. Duda, J. Galas, and M. Slomczynska, "The expression of FSH receptor (FSHR) in the neonatal porcine ovary and its regulation by flutamide," *Reproduction in Domestic Animals*, vol. 46, no. 3, pp. 377–384, 2011.

[15] A. P. N. Themmen and I. T. Huhtaniemi, "Mutations of gonadotropins and gonadotropin receptors: elucidating the physiology and pathophysiology of pituitary-gonadal function," *Endocrine Reviews*, vol. 21, no. 5, pp. 551–583, 2000.

[16] T. Minegish, K. Nakamura, Y. Takakura, Y. Ibuki, and M. Igarashi, "Cloning and sequencing of human FSH receptor cDNA," *Biochemical and Biophysical Research Communications*, vol. 175, no. 3, pp. 1125–1130, 1991.

[17] C. K. Bose, "Follicle stimulating hormone receptor in ovarian surface epithelium and epithelial ovarian cancer," *Oncology Research*, vol. 17, no. 5, pp. 231–238, 2008.

[18] J. H. Yun, B.-H. Gu, Y.-B. Kang, B.-C. Choi, S. Song, and K.-H. Baek, "Association between INS-VNTR polymorphism and polycystic ovary syndrome in a Korean population," *Gynecological Endocrinology*, vol. 28, no. 7, pp. 525–528, 2012.

[19] J. Yang, H. Gong, W. Liu, and T. Tao, "The association of Pro12ala polymorphism in the peroxisome proliferator-activated receptor-gamma2 gene with the metabolic characteristics in chinese women with polycystic ovary syndrome," *International Journal of Clinical and Experimental Pathology*, vol. 6, no. 9, pp. 1894–1902, 2013.

[20] A. Ben-Salem, M. Ajina, M. Suissi, H. S. Daher, W. Y. Almawi, and T. Mahjoub, "Polymorphisms of transcription factor-7-like 2 (TCF7L2) gene in Tunisian women with polycystic ovary syndrome (PCOS)," *Gene*, vol. 533, no. 2, pp. 554–557, 2014.

[21] D. J. Chen, R. Ding, J. Y. Cao, J. X. Zhai, J. X. Zhang, and D. Q. Ye, "Two follicle-stimulating hormone receptor polymorphisms and polycystic ovary syndrome risk: a meta-analysis," *European Journal of Obstetrics & Gynecology and Reproductive Biology*, vol. 182, pp. 27–32, 2014.

[22] L. Li, X. Chen, Z. He, X. Zhao, L. Huang, and D. Yang, "Clinical and metabolic features of polycystic ovary syndrome among Chinese adolescents," *Journal of Pediatric and Adolescent Gynecology*, vol. 25, no. 6, pp. 390–395, 2012.

[23] S. Jonard and D. Dewailly, "The follicular excess in polycystic ovaries, due to intra-ovarian hyperandrogenism, may be the main culprit for the follicular arrest," *Human Reproduction Update*, vol. 10, no. 2, pp. 107–117, 2004.

[24] K. A. Vendola, J. Zhou, O. O. Adesanya, S. J. Weil, and C. A. Bondy, "Androgens stimulate early stages of follicular growth in the primate ovary," *Journal of Clinical Investigation*, vol. 101, no. 12, pp. 2622–2629, 1998.

[25] N. Xita and A. Tsatsoulis, "Review: fetal programming of polycystic ovary syndrome by androgen excess: Evidence from experimental, clinical, and genetic association studies," *Journal of Clinical Endocrinology and Metabolism*, vol. 91, no. 5, pp. 1660–1666, 2006.

[26] S. Catteau-Jonard and D. Dewailly, "Pathophysiology of polycystic ovary syndrome: the role of hyperandrogenism," *Frontiers of Hormone Research*, vol. 40, pp. 22–27, 2013.

[27] B. C. J. M. Fauser, T. D. Pache, S. W. J. Lamberts, W. C. J. Hop, F. H. de Jong, and K. D. Dahl, "Serum bioactive and immunoreactive luteinizing hormone and follicle-stimulating hormone levels in women with cycle abnormalities, with or without polycystic ovarian disease," *Journal of Clinical Endocrinology and Metabolism*, vol. 73, no. 4, pp. 811–817, 1991.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 901303, 11 pages
http://dx.doi.org/10.1155/2015/901303

Research Article

Deciphering the Correlation between Breast Tumor Samples and Cell Lines by Integrating Copy Number Changes and Gene Expression Profiles

Yi Sun and Qi Liu

Department of Central Laboratory, Shanghai Tenth People's Hospital, School of Life Sciences and Technology, Tongji University, Shanghai 200092, China

Correspondence should be addressed to Qi Liu; qiliu@tongji.edu.cn

Received 28 November 2014; Revised 16 January 2015; Accepted 26 January 2015

Academic Editor: Xiao Chang

Breast cancer is one of the most common cancers with high incident rate and high mortality rate worldwide. Although different breast cancer cell lines were widely used in laboratory investigations, accumulated evidences have indicated that genomic differences exist between cancer cell lines and tissue samples in the past decades. The abundant molecular profiles of cancer cell lines and tumor samples deposited in the Cancer Cell Line Encyclopedia and The Cancer Genome Atlas now allow a systematical comparison of the breast cancer cell lines with breast tumors. We depicted the genomic characteristics of breast primary tumors based on the copy number variation and gene expression profiles and the breast cancer cell lines were compared to different subgroups of breast tumors. We identified that some of the breast cancer cell lines show high correlation with the tumor group that agrees with previous knowledge, while a big part of them do not, including the most used MCF7, MDA-MB-231, and T-47D. We presented a computational framework to identify cell lines that mostly resemble a certain tumor group for the breast tumor study. Our investigation presents a useful guide to bridge the gap between cell lines and tumors and helps to select the most suitable cell line models for personalized cancer studies.

1. Introduction

Breast cancer is one of the most frequently diagnosed life-threatening cancers in women with about 235,000 new cases expected in the United States in 2014. Breast cancer is a complex and heterogeneous disease such that they may have different prognoses. It responds to therapy differently despite similarities in histological types, grade, and stage. In the laboratory, the breast cancer is often modelled using established breast cancer cell lines due to their ease of being acquired and used [1].

However, accumulated evidences have pointed out the genomic differences between cancer cell lines and tissue samples in the past decades [2–4]. In the review of Holliday and Speirs [1], they demonstrated that cell lines are prone to genotypic and phenotypic drift during their continual culture. This is particularly common in the more frequently used cell lines, especially those that have been deposited in cell banks for many years [5]. Subpopulations may arise and cause phenotypic changes over time by the selection of specific, more rapidly growing clones within a population. Considering these findings, it is essential for researchers to choose the decent cell lines models when designing experiments and interpreting results, especially if such cell lines are regarded as valid models in evaluating the pathobiology of breast cancer and/or the likely response to novel drug therapies [1].

With the quick development of the whole genome sequencing and other "-omics" techniques, now it becomes possible to systematically explore the relationship between tumor tissues and cancer cell lines and identify the cell lines that most closely resemble particular tumor subtypes. In The Cancer Genome Atlas (TCGA), the genome and expression profiles of at least 500 tissue samples per tumor type are being comprehensively characterized [6]. The Broad-Novartis Cancer Cell Line Encyclopedia (CCLE) contains a compilation of gene expression, chromosomal copy number, and massively parallel sequencing data from 947 human cancer cell lines that are used as models for various tumor

types [7]. These huge data accumulated regarding tumor samples and cell lines have provided a great potential to mine their associations and characterize the cancer mechanisms.

Traditionally, breast cancer was diagnosed into luminal A, luminal B, HER2+/ER−, basal-like, and normal-like subtypes based on gene expression profiling or immunohistochemical (IHC) characteristic [6]. However, classification criteria defined by using only this information may be not sufficient and likely overly general. In this study, we focus on the primary tumors of breast and try to depict the genomic characteristics of these tumors based on their gene expression profiles. Besides, previous studies have suggested that DNA copy number variations (CNVs) are important influential factors for altered gene expression levels in cancer [8–10]. In a lung cancer study, approximately 78% genes showed a positive correlation between CNV and gene expression level [11]. Considering the potential key constitution of CNVs associated with the gene expression variations in breast tumors, copy number profiles were also incorporated in this study.

Using the genomic information, the relationship between these primary breast tumors and the breast cancer cell lines was explored. Furthermore, as intrinsic differences exist among the breast tumor, we also attempt to figure out the correlation between the cell lines and different breast tumor groups and design an efficient computational framework which helps to select the most suitable cell line models for a specified tumor type.

2. Materials and Methods

2.1. Data Collection and Tumor Sample Classification. In our study, we only reserved breast tumor samples or cancer cell lines with both genome-wide DNA copy number information and mRNA expression profiles available. As a curation result, 543 primary breast tumor samples (including 52 normal samples) profiled by TCGA [6] and 59 breast cancer cell lines from the CCLE [7] were obtained.

Generally, breast cancer may be categorized into luminal A, luminal B, HER2+/ER−, basal-like, and normal-like subtypes based on gene expression profiling or immunohistochemical (IHC) characteristics [12, 13]. However, large-scale genomics projects have revealed heterogeneities exist within the same class of breast cancer patients defined by the classic grouping [6]. Here, in order to make a relatively consistent molecular background for the tumor samples in the same group, we subdivided the 491 breast tumors into 8 groups according to the presence or absence of expression of the estrogen receptor (ER), the human epidermal growth factor receptor 2 (ERBB2/HER2), and progesterone receptor (PR) in combination, and there are ER group (ER+, PR−, and HER−; $n = 46$), PR group (ER−, PR+, and HER−; $n = 5$), HER group (ER−, PR−, and HER+; $n = 19$), ERPR group (ER+, PR+, and HER−; $n = 282$), ERHER group (ER+, PR−, and HER+; $n = 14$), PRHER group (ER−, PR+, and HER+; $n = 1$), TP group (ER+, PR+, and HER+; $n = 38$), and TN group (ER+, PR+, and HER+; $n = 86$). The PRHER group was removed in the further study as there was only one sample in the group. The expression pattern of the three marker genes in all tumor samples was shown in supplementary Figure 1 in the Supplementary Material available online at http://dx.doi.org/10.1155/2015/901303.

2.2. Copy Number Data Analysis. Level 3 copy number data was downloaded for breast tumor samples from TCGA (platform: Affymetrix SNP6) [6]. As the CNV sizes are quite different across the tumor samples, the CNV profiles were further broken into gene basis. To enable the gene based analysis, the Bioconductor package CNTools was used to map the segmented copy number data of TCGA samples to genes [14], and each gene corresponds to only one CNV segment. The mean copy number profile of each group of the TCGA samples was obtained by calculating the mean signal of each gene across all tumor samples in this group. Copy number data (gene level) for cancer cell lines was obtained from CCLE (platform: Affymetrix SNP6) [7]. As reported by TCGA and CCLE, the significant focal copy number alterations in individual tumor samples/cancer cell lines were identified from segmented data using GISTIC [15].

Four classes of abnormal segments were considered based on their estimated copy number [16]:

(1) single copy deletion (copy number < 1.5),

(2) double copy deletion (copy number < 0.5),

(3) gain of copy number (copy number > 2.5),

(4) amplification (copy number > 3.5).

2.3. Gene Expression Data Analysis. We used data from the Agilent G4502A_07 platform for TCGA, with measurements of 17,814 genes. Differentially expressed genes were selected based on the fold change of gene expression between each groups of tumor samples and the control (normal group) under the cutoff of $|\log 2\text{foldchange}| > 1$ [17]. The overexpression/underexpression frequency was calculated for each gene in each tumor group. For example, gene A was overexpressed in ER group as compared to the normal group, and then the proportion of tumor samples in ER group with expression value of gene A higher than the mean expression value of gene A in normal group was defined as the overexpression frequency of gene A in ER group.

CCLE expression data was obtained using Affymetrix U133 Plus 2.0 Arrays, with measurements of 18,926 genes. Differentially expressed genes were selected based on the fold change of gene expression between each cell line and the average of expression value of all the cell lines [17].

For the comparison between gene expression data from TCGA and CCLE, robust z-scores (median-centered expression values divided by the median absolute deviation) were derived separately for the two data sets from CCLE and TCGA, and only common genes were remained.

2.4. Gene Set Functional Enrichment Analysis. Gene set enrichment analyses were performed for the functional annotation of the differential expressed genes. Functional Annotation Tools in DAVID Bioinformatics Resources [18] were used to carry out these analyses. Those gene ontology biological process terms with P value less than 0.05 and genes more than two were considered as significant enriched functions for further analysis.

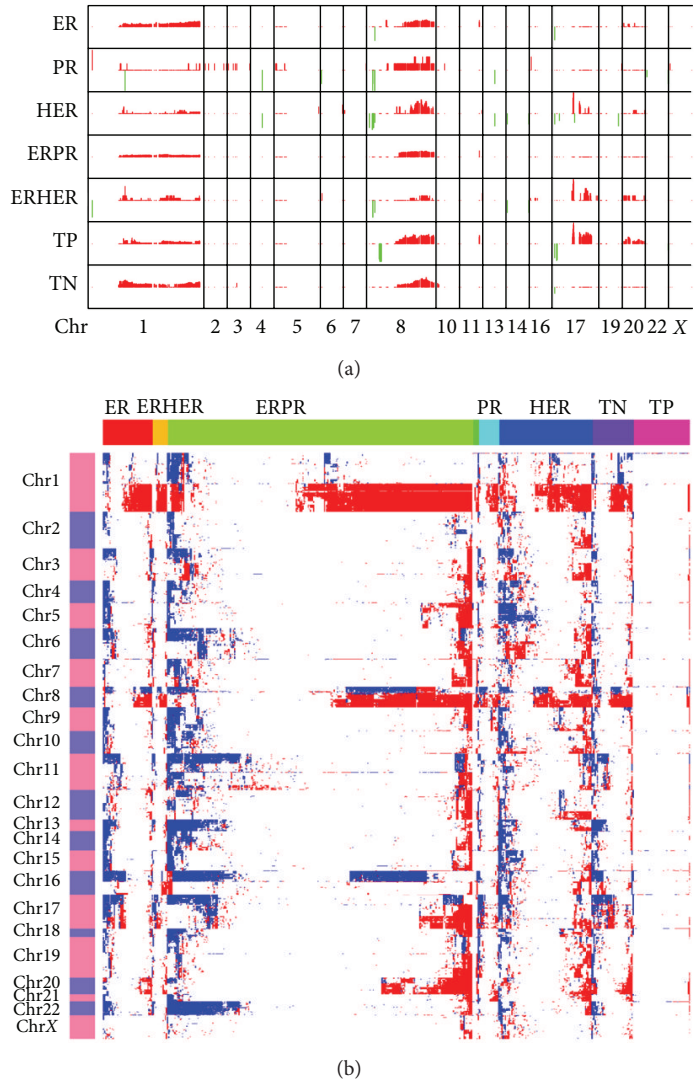


Figure 1: (a) DNA copy number change profiles in each group of breast tumor samples. The CNVs frequency of the whole genome was calculated, the gains of copy number were marked in red, and the losses were marked in green. The y-axis in each subgraph represents the frequency of the copy number gain/loss of the corresponding gene. (b) Clustering of the CNV data. The CNVs on each chromosome in each sample group were clustered separately. The gains of copy number were marked in red and the losses were marked in blue.

2.5. The Construction of "Pathway of the 384 Genes in Breast Tumors". First, pathways closely related to breast cancer were collected via NCI website (http://www.cancer.gov/) and literature review, and they are Estrogen Signaling pathway, ERBB pathway, PI3K/Akt/mTOR Signaling pathway, p53 Signaling pathway, Ras Signaling pathway [19], Notch Signaling pathway [20], Wnt Signaling pathway [21], and NFkB pathway [22]. These pathways were retrieved from KEGG pathway database [23] and compiled into a big pathway via the overlapping elements.

2.6. Rank Aggregation. Two ranking lists derived from copy number profiles and gene expression profiles were fused into one ranking list using R package RankAggreg [24]. Cross Entropy Monte Carlo (CE) algorithm together with Spearman distance was used to perform the rank fusion. The maximum number of iterations was set as 1000.

3. Result

3.1. Genomic Characteristics of Breast Tumor

3.1.1. Copy Number Variations in Breast Tumors. The TCGA and other groups have made great effort to explore the genomic landscape of breast cancer [6, 25]. After classifying the tumor samples from TCGA, we found that, as compared to the normal samples, the tumor samples in other groups show similar copy number variation (CNV) pattern (supplementary Figure 2). Then, we obtained 2,426 genes with CNVs for all groups (supplementary Table 1). It is noteworthy that, for all the groups, the majority of the genes are undergoing frequent copy number gain (Figure 1). Chromosomes 1, 8, 17, and 20 contained most of the genes with CNVs. According to previous studies [6, 26], many genes on chromosomes 8 and 17 show copy number gain, such as MYC on chromosome 8q24, and HER2 as well TOP2A on chromosome 17q21.1.

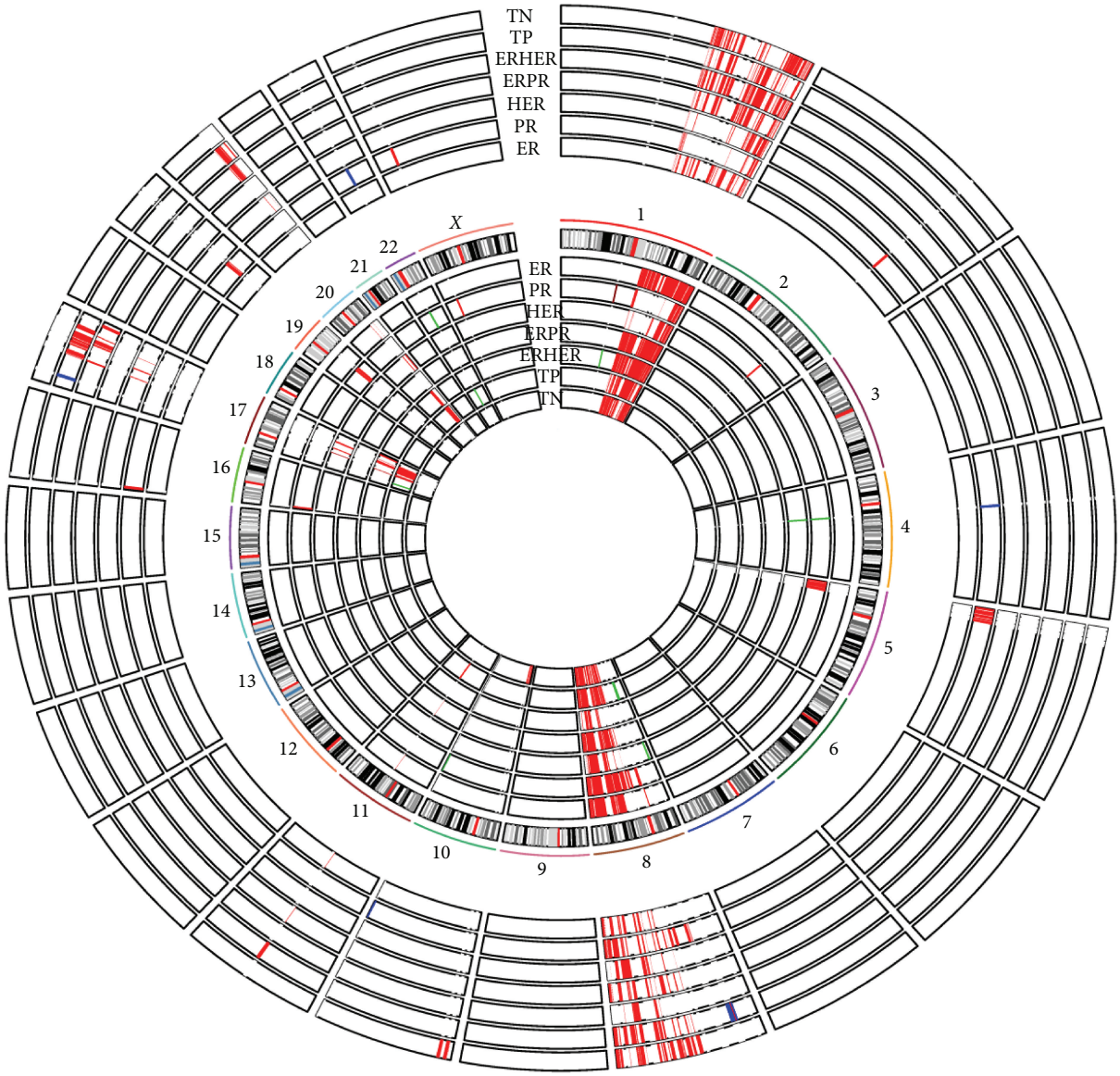


Figure 2: Copy number variation value and expression value of the 384 genes. The 7 circles inside represent the copy number variation of the 384 genes in the 7 tumor groups. The gains of copy number were marked in red and the losses were marked in green. The genes were arranged in chromosomal order (chr1 → chr *X*). The circular rings denote different tumor groups (from outside to inside: ER, PR, HER, ERPR, ERHER, TP, and TN). The 7 circles outside represent the expression value of the 384 genes in the 7 tumor groups. The overexpressed genes were in red and the underexpressed ones were in blue. The genes were arranged in chromosomal order (chr1 → chr *X*). The circular rings denote different tumor groups (from outside to inside: ER, PR, HER, ERPR, ERHER, TP, and TN).

3.1.2. Differentially Expressed Genes in Breast Tumors. Totally, there were 4,843 differentially expressed genes (DEGs) for all groups of tumor samples from TCGA (supplementary Table 2). 399 of the DEGs were overexpressed in all the tumor groups, while 588 of them were underexpressed in all groups (supplementary Figures 3 and 4). There were only 5 overexpressed genes and 14 underexpressed genes unique for ER group, while there were 254 overexpressed genes and 219 underexpressed genes unique for TN group (supplementary Figures 3 and 4). Then, the overexpression/underexpression frequency was calculated for each of the 4,843 genes in each group. Notably, 413 of the genes differ greatly in these tumor sample groups (the deviation between the highest and the lowest frequency of the gene across the groups is bigger than 1), and they were significantly enriched in regulation of hormone levels and cell adhesion.

3.1.3. Genes with Correlations between Copy Number and Expression. We found that totally 384 individual genes show copy number change associated with the alteration in their expression for all tumor sample groups (Figure 2 and supplementary Table 3). The majority of these genes were distributed in chromosomes 1, 8, and 17, which is not surprising, as most of the genes with CNVs were concentrated in these

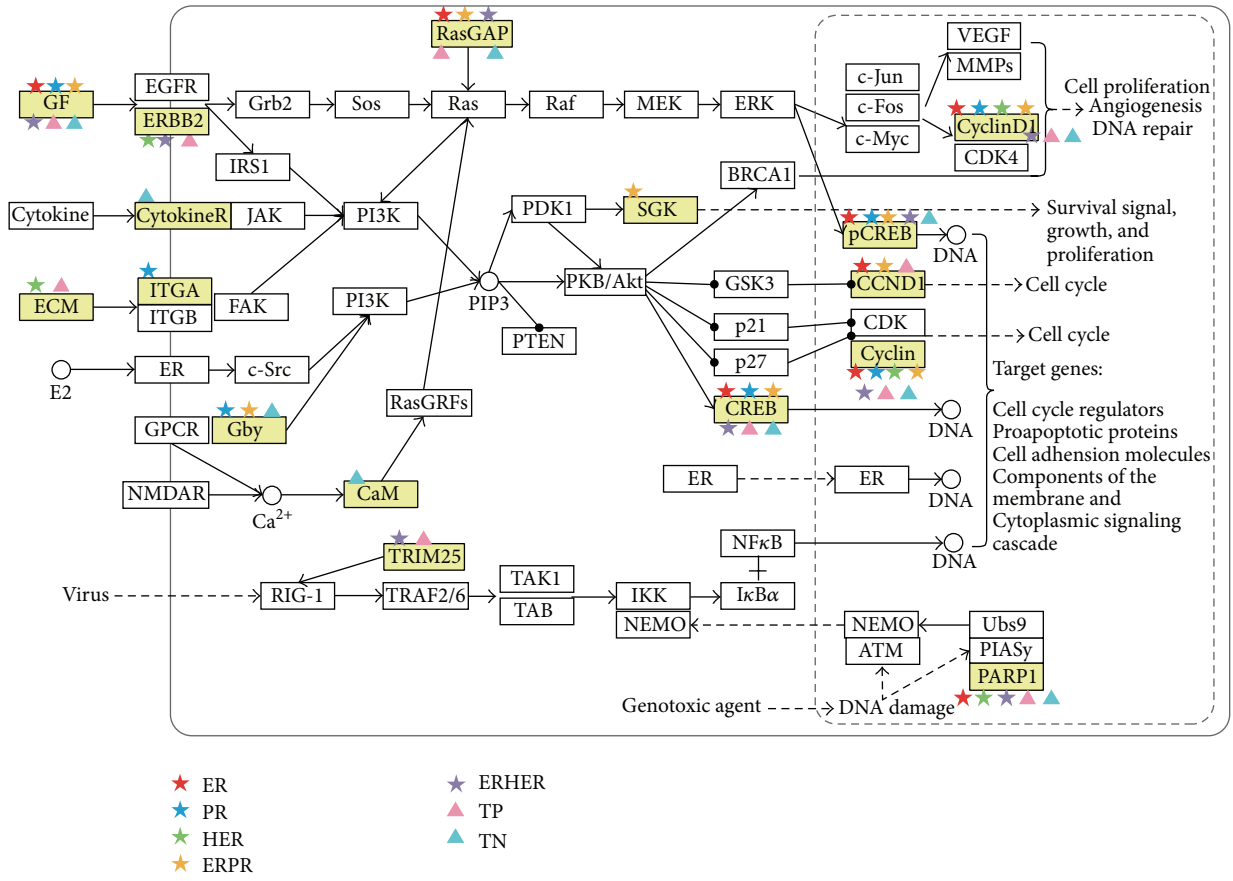


Figure 3: Pathway of the 384 genes in breast tumors. The yellow boxes are the genes that showed copy number change associated with alteration in their expression for all tumor sample groups. The five-pointed stars or triangles with different colors denote the genes in different breast tumor groups.

chromosomes. The genes with high copy number change also show high gene expression change, such as ERBB2, PSMD3, and TCAP. Altogether these genes are significantly enriched in biological processes related to cell cycle. Amplified (and overexpressed) genes are prime therapeutic targets. For example, the use of the drug trastuzumab against ERBB2 has been shown to improve breast cancer survival rates alone or in combination with other treatments [27–29]. The amplified genes with overexpression in each tumor sample groups might be the potential therapeutic targets for the specific tumor type, such as CCND1, CCNE2 for the ER group and E2F5, EIF2C2 for the PR group. 23 of these genes are distributed in the pathways which are closely related to breast cancer: ERBB pathway, PI3K/Akt Signaling pathway, NFkB pathway, and so forth (Figure 3) whereas whether these genes are druggable needs further exploration.

3.2. Correlation between Breast Cancer Cell Lines and Tumor Samples

3.2.1. Characteristics of the Breast Cancer Cell Lines. Among the 59 breast cancer cell lines in CCLE dataset, MCF7, MDA-MB-231, and T-47D are the three most used cell line models for breast cancer account for 82% of current PubMed citations out of the 59 analyzed cell lines (Figure 4). The presence or absence of expression of ER, HER2, and PR in these cell lines was shown in Figure 4, and accordingly, the cell lines were clustered into three parts. These cell lines were also classified into 7 groups as for the breast tumors. The cell lines within the same group show quite different copy number pattern (supplementary Figure 5). The number of overexpressed/underexpressed genes and the count of genes with copy number changes in each cell line were also shown in Figure 4. In general, most of the cell lines have more genes with CNVs rather than DEGs, while CAL51, HS343T, HS606T, HS281T, HMEL, HS274T, HS739T, and HS742T have more DEGs rather than genes with CNVs.

3.2.2. Comparing the Breast Cancer Cell Lines to the Tumor Groups. For each cell line, the copy number profiles were compared with the mean copy number profile of each tumor sample group by calculating Spearman correlation coefficients using the 2,426 genes with CNVs profiles in the tumor samples (Figure 5). In this way, we obtained the correlation between each of the cell lines and the different tumor groups. 12 cell lines (e.g., BT20, BT474, EFM19, etc.) show high correlation with their preclassification indicated by

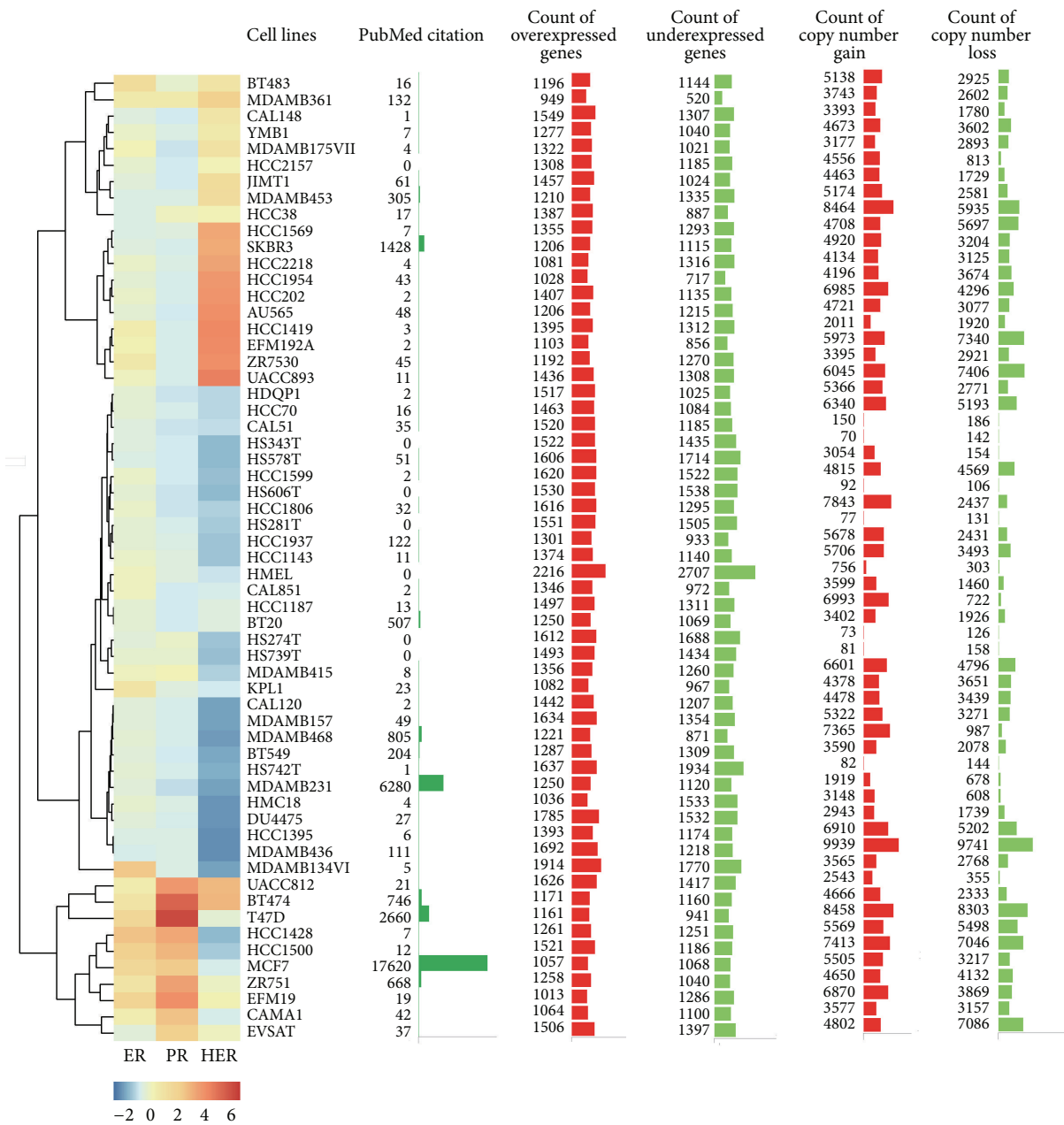


FIGURE 4: General information of the 59 breast cancer cell lines. The fold change values of ER, HER2, and PR in these cell lines were summarized in a heat map, with blue indicating low fold change value and orange indicating high fold change value.

the presence or absence of expression of ER, HER2, and PR in the cell line. We surprisingly found that the most cited three cell lines MCF7, MDA-MB-231, and T-47D do not show high correlation with the preclassified tumor group. Additionally, some cell lines (HS343T, HS606T, HS739T, and HS742T) show low correlation to any one of the tumor groups. This is probably due to the fact that these established breast cancer cell lines are not derived from the primary breast tumors but from tumor metastases. This means that these cell lines are derived from more aggressive metastatic tumors, rather than the primary lesion [1]. Besides, cell lines are purer than tumor samples, which tend to be contaminated with stromal cells [4]. In addition, cell lines are prone to genotypic and phenotypic drift during their continual culture, especially those that have been deposited in cell banks for many years [30].

Then, the gene expression profiles of each cancer cell line were compared with the mean gene expression profile of each tumor sample group by calculating Spearman correlation coefficients using the 4,843 genes differentially expressed in the tumor samples (Figure 6). In this way, we obtained the correlation between each of the cell lines and the different tumor groups. As a whole, the correlation between cell lines and tumor sample groups using gene expression profiles is in accordance with that revealed using CNVs data.

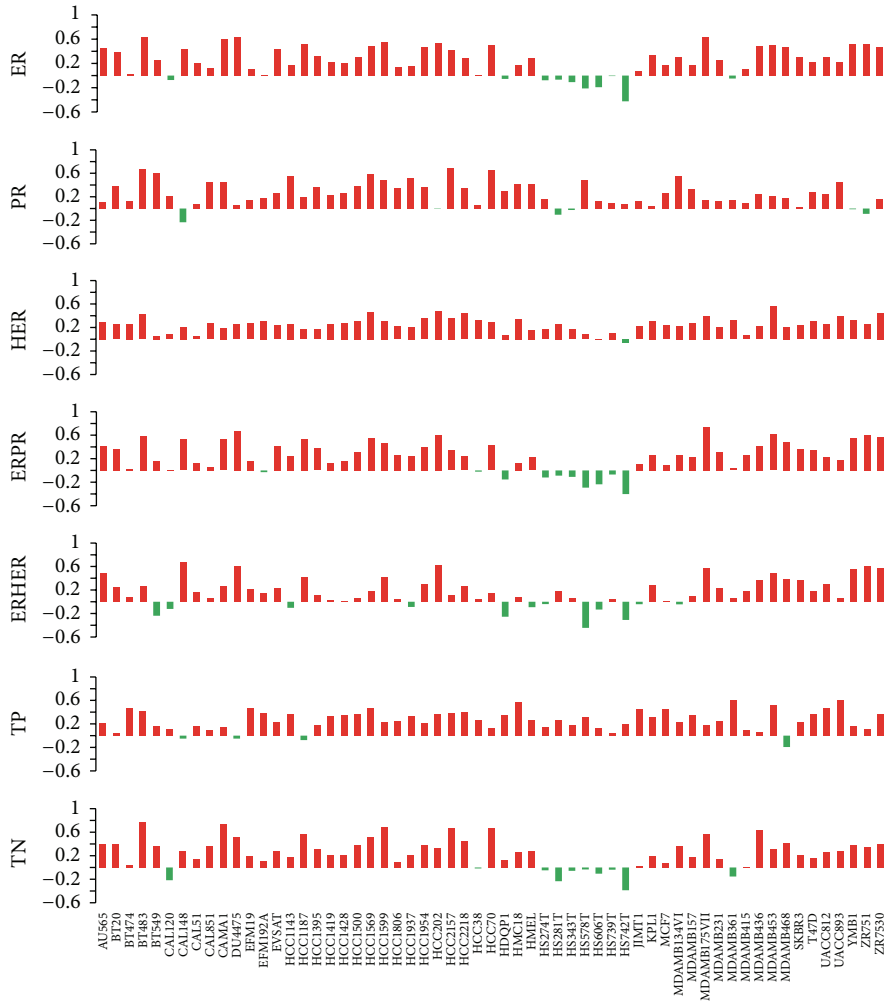


Figure 5: Correlation between the 59 breast cancer cell lines and the tumor sample groups using copy number data.

The difference is that the correlation values are lower than those calculated using CNVs information but with higher concordance (29 cell lines) with the classification indicated by the presence or absence of expression of ER, HER2, and PR in the cell line. HS343T, HS606T, HS739T, and HS742T also show low correlation to any one of the tumor groups.

Additionally, we examined the overlap ratio of genes that showed copy number change associated with alteration in their expression between each breast cancer cell line and each tumor sample group (Figure 7). This ratio could also indicate the correlation between cancer cell lines and different tumor samples, as it shows high consistency with that only by copy number profiles or gene expression profiles.

3.3. Ranking of the Breast Cancer Cell Lines as Candidate Models for Certain Tumor Group Study. Breast cancer is a complex disease that manifests as a result of coordinated alterations on genomic, epigenomic, and proteomic levels. Therefore, it is important to take into account the multiple datasets together to optimize strength of biological information across multiple assays relevant to breast cancer. With the accumulated copy number profiles and gene expression profiles for different cancer cell lines and tumor samples, we could evaluate whether a certain breast cancer cell line is a good model for a specific tumor group by integration of these two aspects of information. We designed a ranking aggregation model of the cell lines according to their correlation with each tumor group based on the integration of copy number profiles and gene expression profiles. First, the breast cancer cell lines were ranked in descending order of their similarity with each tumor group using copy number profiles and gene expression profiles, respectively. Then, for each tumor group, the two derived ranking lists of the breast cancer cell lines were fused into one ranking list using R package RankAggreg [24]. In this way, the good cell line models for each tumor groups were picked out from the 59 breast cancer cell lines.

For each tumor group, the cell lines ranked in the top resemble the tumor group best and might be the best cell line models for laboratory studies. Suggested by the final ranking list, the best three cell line models for each breast tumor group were listed as follows: CAMA1, BT483, and HCC202 (ER group); HCC70, HCC1143, and HCC1937 (PR group); MDAMB453, HCC2218, and UACC893 (HER group); MDAMB453, CAL148, and ZR751 (ERHER group); HCC202,

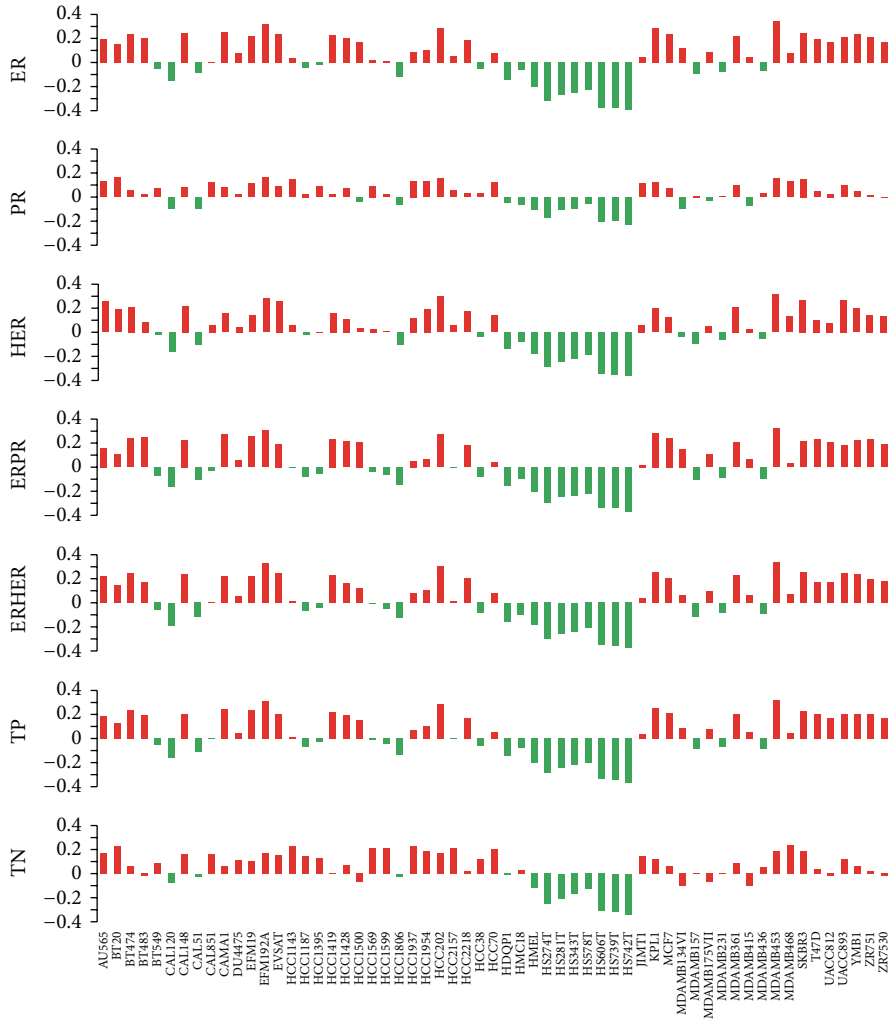


Figure 6: Correlation between the 59 breast cancer cell lines and the tumor sample groups using gene expression profiles.

BT483, and ZR751 (ERPR group); MDAMB453, MDAMB361, and UACC893 (TP group); HCC1599, HCC70, and HCC1569 (TN group). There results may provide useful clues for future personalized breast cancer study.

3.4. Comparing All the Cancer Cell Lines with Breast Tumor Samples. Similarly, we also evaluated the correlation between all the cancer cell lines in CCLE and breast tumor sample groups, using the copy number information and gene expression profiles (supplementary Tables 4 and 5). From the perspective of either copy number or gene expression profiles, respectively, some breast cancer cell lines were ranked with high correlation with any of the breast tumor groups while, interestingly, we also identified that some lung cancer cell lines and ovary cancer cell lines also present high correlation with at least one of the breast tumor groups.

4. Discussion

4.1. Breast Tumor Sample Groups Differ Greatly in the Regulation of Hormone Levels and Cell Adhesion. 413 of the DEGs differ greatly in the frequency of overexpression/underexpression across the different tumor sample groups. After conducting the gene set functional enrichment analysis, we found these genes were significantly enriched in biological processes including the regulation of hormone levels and cell adhesion. The enrichment in the regulation of hormone levels is expected. As cell adhesion is related to cancer metastasis, we checked the literatures and found that different breast cancer subtypes show disparity in metastasizing to different sites [31, 32]. However, the classification of breast cancer into subtypes does not typically inform about metastatic behavior. These genes (COL9A1, ITGB8, ITGB6, TTYH1, RET, etc.) enriched in the cell adhesion may serve as important indicators of different types of breast cancer. Due to limited information in this field, the roles of these genes in manipulating the tendency of breast cancer metastasis to different sites need to be further studied.

4.2. Correlation between Breast Cancer Cell Lines and the Tumor Samples. The integrative genomic study of copy number profiles and gene expression profiles collected on the same set of cancer cell lines and patient samples could serve as an efficient way to depict the characteristics of different tumor types and cancer cell lines. Besides, the integrated

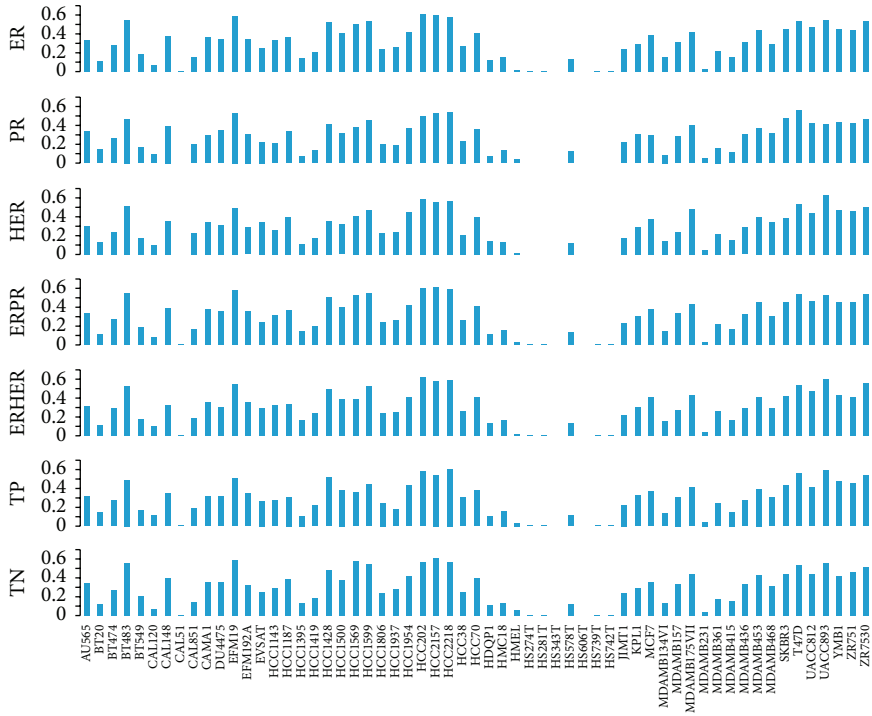


FIGURE 7: The overlap ratio of genes shows copy number change associated with alteration in their expression between each breast cancer cell lines and the tumor sample groups.

investigation of the two perspectives also provides a guide to reveal the relationship between breast cancer cell lines and the tumor samples, as well as selecting the suitable cell lines for the corresponding breast tumor group. In general, the correlation between the cancer cell lines and the tumor sample groups indicated by the two aspects was consistent with each other. The association between copy number variations and gene expression has been investigated by several research groups [33, 34]. As DNA copy number variations (CNVs) are important influential factors for altered gene expression levels in cancer, the observed high consistency was expected.

Some of the cancer cell lines have high correlation with the preclassified tumor group based on the presence or absence of expression of ER, HER2, and PR in the cell line, while a big part of them does not show this tendency, including the most used MCF7, MDA-MB-231, and T-47D. According to ATCC (http://www.atcc.org/) which is one of the largest biosources in the world and offers investigators a complex array of human, animal, insect, fish, and stem cell lines, these three cell lines are not from primary breast cancer but are metastatic breast cancer cell lines derived from pleural effusion. Some of the cell lines (HS343T, HS606T, HS739T, and HS742T) have low correlation to any one of the primary tumor groups either calculated using copy number profiles or gene expression profiles. The low correlation probably lies in that they are not originated from primary tumors, or maybe these cell lines were contaminated during their continual culture.

Indicated by the fused rank based on the similarity of copy number profiles and gene expression profiles, the most resemble breast cancer cell lines were picked out as the good models for different tumor groups. Further evidences might be identified by investigating mutation profiles, proteomics data, and so forth.

4.3. Lung Cancer Cell Lines and Ovary Cancer Cell Lines Show High Correlation with the Breast Tumor Samples. By evaluating the correlation between other cancer cell lines in CCLE with the breast tumor groups, we found some of the lung cancer cell lines and the ovary cancer cell lines show high relevance with the breast tumors. In the systematic analysis of the genomic characteristics of breast tumors, the similarity between ovary tumors and lung tumors was observed [6]. The high correlation between some of the ovary/lung cancer cell lines and the breast tumors was understandable. In addition to the similar CNV profile (e.g., common gains in chromosomes 1, 8, 17, and 20) and gene expression profile (e.g., overexpression of AKT3, MYC) between the breast tumor samples and the ovary/lung cancer cell lines, there are some other commonalities between them. For example, breast tumors and ovary tumors have common risk factors including hormone therapy, obesity, and inherited genetic risk such as BRCA1 and BRCA2 [35, 36]. For breast tumors and lung tumors, they have high frequency of TP53 mutations, EGFR mutation, and so forth [37].

5. Conclusion

In this paper, we investigated the correlation between different groups of primary breast tumors and breast cancer cell

lines using copy number profiles and gene expression profiles. Although the relevance between tumors and cancer cell lines seems not very high, while considering their ease of use, there is no doubt that established cell lines will continue to be used as models for breast cancer. Our study is expected to provide a useful guide for researchers to understand the limitations of the cells and select the suitable cell lines as the tumor model for better investigation of cancer mechanism.

Conflict of Interests

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

Acknowledgments

This work was supported by the National High Technology Research and Development Program ("863" Program) of China (Grant 2012AA020405), the National Natural Science Foundation of China (Grants 31100956 and 61173117), the Fundamental Research Funds for the Central Universities of China (Grant 2000219084), and the key project of Scientific Research Innovation, Shanghai (Grant 20002360059).

References

[1] D. L. Holliday and V. Speirs, "Choosing the right cell line for breast cancer research," *Breast Cancer Research*, vol. 13, no. 4, article 215, 2011.

[2] J.-P. Gillet, A. M. Calcagno, S. Varma et al., "Redefining the relevance of established cancer cell lines to the study of mechanisms of clinical anti-cancer drug resistance," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 46, pp. 18708–18713, 2011.

[3] R. Sandberg and I. Ernberg, "Assessment of tumor characteristic gene expression in cell lines using a tissue similarity index (TSI)," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 6, pp. 2052–2057, 2005.

[4] S. Domcke, R. Sinha, D. A. Levine, C. Sander, and N. Schultz, "Evaluating cell lines as tumour models by comparison of genomic profiles," *Nature Communications*, vol. 4, article 2126, 2013.

[5] S. E. Burdall, A. M. Hanby, M. R. J. Lansdown, and V. Speirs, "Breast cancer cell lines: friend or foe?" *Breast Cancer Research*, vol. 5, no. 2, pp. 89–95, 2003.

[6] The Cancer Genome Atlas Network, "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, no. 7418, pp. 61–70, 2012.

[7] J. Barretina, G. Caponigro, N. Stransky et al., "The cancer cell Line encyclopedia enables predictive modelling of anticancer drug sensitivity," *Nature*, vol. 483, no. 7391, pp. 603–607, 2012.

[8] E. Chaignat, E. A. Yahya-Graison, C. N. Henrichsen et al., "Copy number variation modifies expression time courses," *Genome Research*, vol. 21, no. 1, pp. 106–113, 2011.

[9] L. D. Orozco, S. J. Cokus, A. Ghazalpour et al., "Copy number variation influences gene expression and metabolic traits in mice," *Human Molecular Genetics*, vol. 18, no. 21, pp. 4118–4129, 2009.

[10] B. E. Stranger, M. S. Forrest, M. Dunning et al., "Relative impact of nucleotide and copy number variation on gene phenotypes," *Science*, vol. 315, no. 5813, pp. 848–853, 2007.

[11] T.-P. Lu, L.-C. Lai, M.-H. Tsai et al., "Integrated analyses of copy number variations and gene expression in lung adenocarcinoma," *PLoS ONE*, vol. 6, no. 9, Article ID e24829, 2011.

[12] P. Maisonneuve, D. Disalvatore, N. Rotmensz et al., "Proposed new clinicopathological surrogate definitions of luminal A and luminal B (HER2-negative) intrinsic breast cancer subtypes," *Breast Cancer Research*, vol. 16, no. 3, article R65, 2014.

[13] M. Yanagawa, K. Ikemot, S. Kawauchi et al., "Luminal A and luminal B (HER2 negative) subtypes of breast cancer consist of a mixture of tumors with different genotype," *BMC Research Notes*, vol. 5, article 376, 2012.

[14] R. C. Gentleman, V. J. Carey, D. M. Bates et al., "Bioconductor: open software development for computational biology and bioinformatics," *Genome Biology*, vol. 5, no. 10, article R80, 2004.

[15] C. H. Mermel, S. E. Schumacher, B. Hill, M. L. Meyerson, R. Beroukhim, and G. Getz, "GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers," *Genome Biology*, vol. 12, no. 4, article R41, 2011.

[16] G. Lenz, G. W. Wright, N. C. T. Emre et al., "Molecular subtypes of diffuse large B-cell lymphoma arise by distinct genetic pathways," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 36, pp. 13520–13525, 2008.

[17] R. Agren, S. Bordel, A. Mardinoglu, N. Pornputtapong, I. Nookaew, and J. Nielsen, "Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT," *PLoS Computational Biology*, vol. 8, no. 5, Article ID e1002518, 2012.

[18] W. da Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources," *Nature Protocols*, vol. 4, no. 1, pp. 44–57, 2009.

[19] C. Choi and D. M. Helfman, "The Ras-ERK pathway modulates cytoskeleton organization, cell motility and lung metastasis signature genes in MDA-MB-231 LM2," *Oncogene*, vol. 33, no. 28, pp. 3668–3676, 2013.

[20] H. Al-Hussaini, D. Subramanyam, M. Reedijk, and S. S. Sridhar, "Notch signaling pathway as a therapeutic target in breast cancer," *Molecular Cancer Therapeutics*, vol. 10, no. 1, pp. 9–15, 2011.

[21] R. Lamb, M. P. Ablett, K. Spence, G. Landberg, A. H. Sims, and R. B. Clarke, "Wnt pathway activity in breast cancer sub-types and stem-like cells," *PLoS ONE*, vol. 8, no. 7, Article ID e67811, 2013.

[22] Y. Zhou, S. Eppenberger-Castori, U. Eppenberger, and C. C. Benz, "The NFkappaB pathway and endocrine-resistant breast cancer," *Endocrine-Related Cancer*, vol. 12, supplement 1, pp. S37–S46, 2005.

[23] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, 2000.

[24] V. Pihur and S. Datta, "RankAggreg, an R package for weighted rank aggregation," *BMC Bioinformatics*, vol. 10, article 62, 2009.

[25] S. P. Shah, A. Roth, R. Goya et al., "The clonal and mutational evolution spectrum of primary triple-negative breast cancers," *Nature*, vol. 486, no. 7403, pp. 395–399, 2012.

[26] Y. Kimura, T. Noguchi, K. Kawahara, K. Kashima, T. Daa, and S. Yokoyama, "Genetic alterations in 102 primary gastric cancers by comparative genomic hybridization: gain of 20q and loss of 18q are associated with tumor progression," *Modern Pathology*, vol. 17, no. 11, pp. 1328–1337, 2004.

[27] L. A. Emens and N. E. Davidson, "Trastuzumab in breast cancer," *Oncology*, vol. 18, no. 9, pp. 1117–1128, 2004.

[28] J. Baselga, "Herceptin alone or in combination with chemotherapy in the treatment of HER2-positive metastatic breast cancer: pivotal trials," *Oncology*, vol. 61, supplement 2, pp. 14–21, 2001.

[29] C. L. Vogel, M. A. Cobleigh, D. Tripathy et al., "First-line herceptin monotherapy in metastatic breast cancer," *Oncology*, vol. 61, supplement 2, pp. 37–42, 2001.

[30] D. Ferreira, F. Adega, and R. Chaves, "The importance of cancer cell Lines as in vitro models in cancer methylome analysis and anticancer drugs testing," in *Oncogenomics and Cancer Proteomics—Novel Approaches in Biomarkers Discovery and Therapeutic Targets in Cancer*, InTech, Rijeka, Croatia, 2013.

[31] H. Kennecke, R. Yerushalmi, R. Woods et al., "Metastatic behavior of breast cancer subtypes," *Journal of Clinical Oncology*, vol. 28, no. 20, pp. 3271–3277, 2010.

[32] R. Yerushalmi, S. Tyldesley, H. Kennecke et al., "Tumor markers in metastatic breast cancer subtypes: frequency of elevation and correlation with outcome," *Annals of Oncology*, vol. 23, no. 2, pp. 338–345, 2012.

[33] L. Cheng, P. Wang, S. Yang et al., "Identification of genes with a correlation between copy number and expression in gastric cancer," *BMC Medical Genomics*, vol. 5, article 14, 2012.

[34] R. X. Menezes, M. Boetzer, M. Sieswerda, G.-J. B. van Ommen, and J. M. Boer, "Integrated analysis of DNA copy number and gene expression microarray data using gene sets," *BMC Bioinformatics*, vol. 10, article 203, 2009.

[35] P. Peterlongo, J. Chang-Claude, K. B. Moysich et al., "Candidate genetic modifiers for breast and ovarian cancer risk in BRCA1 and BRCA2 mutation carriers," *Cancer Epidemiology, Biomarkers & Prevention*, vol. 24, no. 1, pp. 308–316, 2014.

[36] I. M. Collins, R. L. Milne, P. C. Weideman et al., "Preventing breast and ovarian cancers in high-risk *BRCA1* and *BRCA2* mutation carriers," *The Medical Journal of Australia*, vol. 199, no. 10, pp. 680–683, 2013.

[37] G. Ciriello, M. L. Miller, B. A. Aksoy, Y. Senbabaoglu, N. Schultz, and C. Sander, "Emerging landscape of oncogenic signatures across human cancers," *Nature Genetics*, vol. 45, no. 10, pp. 1127–1133, 2013.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 205247, 6 pages
http://dx.doi.org/10.1155/2015/205247

Research Article

Network Comparison of Inflammation in Colorectal Cancer and Alzheimer's Disease

Sungjin Park,[1] Seok Jong Yu,[2] Yongseong Cho,[2] Curt Balch,[3] Jinhyuk Lee,[4] Yon Hui Kim,[1] and Seungyoon Nam[1]

[1]*New Experimental Therapeutics Branch, National Cancer Center, Goyang-si, Gyeonggi-do 410-769, Republic of Korea*
[2]*Supercomputing R&D Center, Korea Institute of Science and Technology Information (KISTI), Daejeon 305-806, Republic of Korea*
[3]*Bioscience Advising, Indianapolis, IN 46227, USA*
[4]*Korean Bioinformation Center (KOBIC), Korea Research Institute of Bioscience and Biotechnology, Daejeon 305-806, Republic of Korea*

Correspondence should be addressed to Seungyoon Nam; seungyoon.nam@ncc.re.kr

Received 17 December 2014; Revised 16 February 2015; Accepted 16 February 2015

Academic Editor: Junwen Wang

Recently, a large clinical study revealed an inverse correlation of individual risk of cancer versus Alzheimer's disease (AD). However, no explanation exists for this anticorrelation at the molecular level; however, inflammation is crucial to the pathogenesis of both diseases, necessitating a need to understand differing signaling usage during inflammatory responses distinct to both diseases. Using a subpathway analysis approach, we identified numerous well-known and previously unknown pathways enriched in datasets from both diseases. Here, we present the quantitative importance of the inflammatory response in the two disease pathologies and summarize signal transduction pathways common to both diseases that are affected by inflammation.

1. Introduction

Epidemiological evidence has revealed an inverse incidence between Alzheimer's disease (AD) and cancer that increases exponentially among aged cohorts [1, 2]. However, despite the clear differences in the etiology of the two diseases, including the premature death of neurons in AD and evasion of apoptosis in cancer, it has been suggested that common signaling pathways are involved in the two age-associated diseases [3]. Molecular comparative surveys of the two disease states have led to speculation of roles for p53 and the Wnt signaling pathway in both cancer and AD [4]. However, a global transcriptomic network comparison between the two diseases has yet to be completed [2].

Of interest, immune response is intimately related to both diseases [5–7]. In fact, based on an early colorectal cancer (CRC) transcriptome dataset [8], our previous study [9] found immunosuppression and immune cell infiltration even within normal-appearing cells in CRC patients. Similarly, in the brain, microglia and astrocytes involved in inflammation play a critical role in neurodegeneration [6, 7].

Despite continuous efforts to understand the individual molecular mechanisms of the two diseases, distinction of the global effects of immune response toward specific signal transduction usage in the two diseases has not been established. Here, we systematically inspected the two diseases representing phenotypically opposite cell fates, death and survival, by utilizing functional enrichment analysis and a systems biology approach [9]. This functional enrichment indicated that inflammatory response was significantly involved in both diseases. Subsequently, we found, by the systems biology approach, that various pathways within each disease network were comprised of common inflammation-associated genes.

2. Materials and Methods

2.1. Functional Enrichment Comparison of CRC and AD. Throughout the paper, we compared one colorectal cancer (CRC) dataset (GEO accession GSE4107) [8] with two AD datasets (GEO accessions GSE1297, GSE12685) [10, 11] from

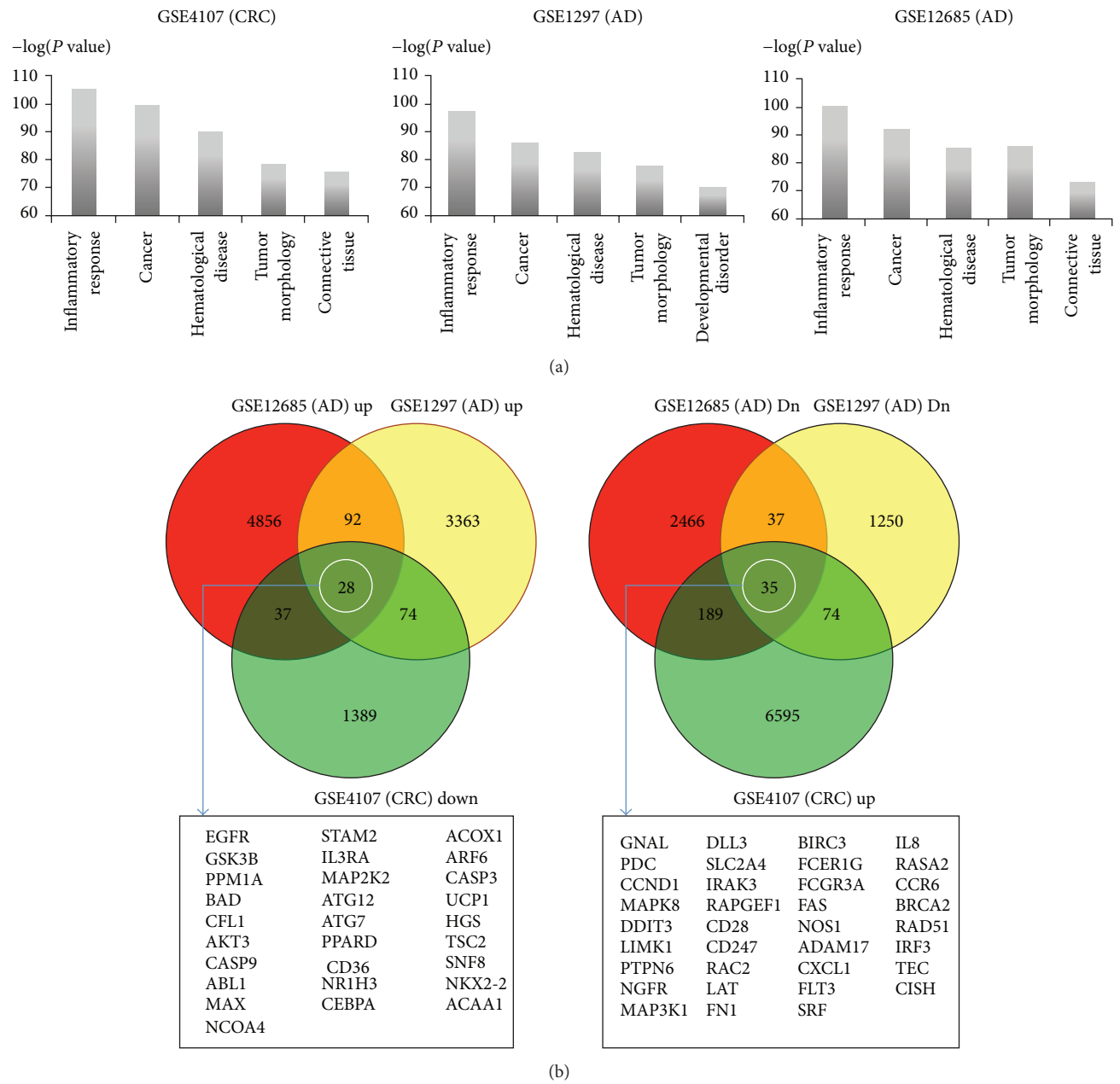


Figure 1: IPA functional enrichment of the CRC and the AD datasets. (a) Top 5 functional categories from "Diseases and Functions" ontology for the datasets are represented. The y-axis represents the minus logarithms of the P values. The higher the value on the y-axis is, the more statistically significant it becomes. The x-axis represents the functional categories. (b) The common genes inversely expressed between the two diseases are indicated by white ovals (see details in Section 2). In the Venn diagrams, "GSE12685 (AD) Dn" is the downregulated gene set in AD patients versus controls. "GSE12685 (AD) Up" is the upregulated gene set in AD patients versus controls. The notation is similar to the GSE1297 (AD) dataset and the GSE4107 (CRC) dataset.

GEO (see details in Supplementary Table S1 in Supplementary Material available online at http://dx.doi.org/10.1155/2015/205247). We used Ingenuity Pathway Analysis (IPA, Qiagen, Valencia, CA, USA) to inspect functionally enriched terms within the IPA "Diseases and Functions" ontology, revealing the top 5 significant terms for the three datasets (Figure 1(a)). For functional enrichment analysis, we uploaded the expression fold-changes of all the genes for the three datasets into IPA: in the CRC dataset, the expression fold-changes of patients versus controls were obtained and in AD, the fold-changes of AD patients versus controls were obtained.

Since cancer and AD are phenotypically opposite (cell survival versus cell death), we obtained oppositely expressed common genes between the two diseases. Based on all the genes' fold-changes from the three datasets, we obtained the common genes as shown in Figure 1(b).

Table 1: Inflammation-associated genes common to both AD and CRC show opposite expression patterns. The 16 oppositely expressed common genes (in Figure 1(b)) between AD and CRC were assigned to inflammation-associated functional terms in IPA.

Functional category	Downregulated in AD and upregulated in CRC	Upregulated in AD and downregulated in CRC
Chemokine	PTPN6$^{+*\#}$, IRAK3$^{+*\#}$, FLT3$^{+*\#}$	BAD$^{+*\#}$, CD36$^{+*\#}$
Inflammation relating to CRC	DDIT3$^{+*\#}$, FAS$^{+*\#}$, IRF3$^{+*\#}$	
Inflammation relating to brain	CCR6$^{+*\#}$, CD28$^{+*\#}$, DDIT3$^{+*\#}$, FAS$^{+*\#}$, FCER1G$^{+*\#}$, NGFR$^{+*\#}$	PPARD$^{+*\#}$
Cytokines relating to cancer	CD28$^{+*\#}$, FN1$^{+*\#}$	ABL1$^{+*\#}$, EGFR$^{+*\#}$
Cytokines relating to brain		CD36$^{+*\#}$

$^{+}$Genes detected in the CRC network from GSE4107 dataset.
*Genes detected in the AD network from GSE1297 dataset.
$^{\#}$Genes detected in the AD network from GSE12685 dataset.

2.2. Network Construction of CRC and AD. For generating networks from the three datasets, we applied our previous subpathway-based systems biology approach [9]. In brief, KEGG pathways were decomposed to all their possible paths (i.e., subpathways). In a given dataset, we applied a statistical test to each subpathway to determine whether the gene expression levels agreed with edge types (e.g., activation, inhibition) of the subpathway. Subsequently, in the dataset, we gathered the statistically significant subpathways (P values <0.05) that comprised the network.

3. Results and Discussion

3.1. Overview. While cancer and AD are two of the most common diseases worldwide (15.6 million versus 7.7 million new cases per year) relating to aging, their phenotypes are opposite: cell death (neurons) in AD versus survival (mostly epithelial cells) in cancer. Also, AD patients are less susceptible to cancer and vice versa [1]. Consequently, we aimed at understanding changes at the molecular level between the two diseases. First, we inspected functional enrichment comparison of a cancer dataset (from our previous study) and the two AD datasets. Second, due to the involvement of inflammation in both pathologies [12, 13], we aimed to identify global network differences between the two diseases to possibly identify differential inflammation environments and differential chemokine/cytokine receptor usages. For this purpose, we selected colorectal cancer (CRC) as the cancer dataset to extend our previous result [9]. We also obtained the two independent AD datasets from GEO (Supplementary Table S1).

3.2. Functional Enrichment Comparison of CRC and AD: Inflammation-Related Genes. We used Ingenuity Pathway Analysis (IPA) to perform functional pathway enrichment of early CRC and AD. IPA reported the top 5 functional categories from its "Diseases and Functions" ontology. In Figure 1(a), inflammatory response-related genes, as well as cancer-associated genes, were significantly enriched in the CRC and the AD datasets.

Figure 1(b) shows common genes that were inversely expressed between the two phenotypically opposite diseases. Considering that the biopsy tissues for the datasets contain immune cells, inflammatory response is reasonable for functional enrichment.

Out of the common genes in Figure 1(b), *ARF6* was upregulated in the AD datasets but downregulated in the CRC dataset. ARF6, a small GTPase [14–16], regulates early endosome internalization of the protease BACE1, Beta-Site APP-Cleaving Enzyme 1. This internalization enables BACE1 to encounter and cleave intracellular amyloid precursor proteins (APPs), leading to amyloidogenic processing for the accumulation Aβ dimers in neurons, a hallmark of AD pathology [17].

CCR6 (in Figure 1(b)) was upregulated in CRC but downregulated in both AD datasets. CCR6 is an important surface marker of immunosuppressive immune cells in the CRC tumor microenvironment [18]. Regulatory T cells (T_{Reg} cells) expressing CCR6 are recruited to a tumor mass by tumor-associated macrophages (TAMs), and tumor development is enhanced by CCR6 binding to its ligand CCL20 (CRC 1.721-fold of overcontrol in the GSE4107 dataset) secreted by tumor cells [18]. This scenario agrees with our previous result, indicating T_{Reg} cell infiltration into normal-appearing mucosa in CRC patients [9]. Considering that T and B cells do not exist in brain, the low expression of the T_{Reg} cell surface markers in AD patients is not surprising.

We further dissected the common genes (28 and 35 genes in white circles in the Venn diagram in Figure 1(b)) in terms of inflammation, considering that inflammatory response was the highest enrichment in all three datasets. For this purpose, we selected several terms involved in inflammation from the IPA "Diseases and Functions" ontology (see the terms and entries in Supplementary Table S2). Out of the genes common to the three datasets, 16 were oppositely expressed between the two diseases in terms of IPA inflammation-related terms (Table 1).

3.3. Network Construction of CRC and AD. Next, we constructed molecular networks for the two diseases. By applying our previous systems biology method to the three disease

Table 2: KEGG pathways associated with the 16 oppositely expressed common genes (in Table 1) in the AD and the CRC networks. From the AD and the CRC networks, pathway information of the 16 genes was obtained. The 16 genes were inversely expressed in the pathways between the AD and the CRC networks.

Gene symbols	Pathways	CRC (GSE4107)	AD (GSE12685)	AD (GSE1297)
PTPN6	hsa04662_B_cell_receptor_signaling_pathway; hsa04630_Jak-STAT_signaling_pathway; hsa05140_Leishmaniasis	Up	Down	Down
IRAK3	hsa04722_Neurotrophin_signaling_pathway			
FLT3	hsa05221_Acute_myeloid_leukemia			
DDIT3	hsa04010_MAPK_signaling_pathway			
FAS	hsa04115_p53_signaling_pathway; hsa04650_Natural_killer_cell_mediated_cytotoxicity			
IRF3	hsa04622_RIG-I-like_receptor_signaling_pathway; hsa04623_Cytosolic_DNA-sensing_pathway			
CCR6	hsa04060_Cytokine-cytokine_receptor_interaction; hsa04062_Chemokine_signaling_pathway			
CD28	hsa04660_T_cell_receptor_signaling_pathway; hsa05416_Viral_myocarditis			
FCER1G	hsa04650_Natural_killer_cell_mediated_cytotoxicity			
NGFR	hsa04722_Neurotrophin_signaling_pathway			
FN1	hsa04512_ECM-receptor_interaction			
BAD	hsa04510_Focal_adhesion; hsa05223_Non-small_cell_lung_cancer; hsa05210_Colorectal_cancer	Down	Up	Up
CD36	hsa03320_PPAR_signaling_pathway; hsa04512_ECM-receptor_interaction			
PPARD	hsa05221_Acute_myeloid_leukemia; hsa04310_Wnt_signaling_pathway			
ABL1	hsa04012_ErbB_signaling_pathway; hsa04722_Neurotrophin_signaling_pathway			
EGFR	hsa05214_Glioma; hsa04012_ErbB_signaling_pathway			

datasets, we obtained CRC and AD pathogenesis networks (Supplementary Figures S1–S3). We summarized the most significant 100 subpathways for each network (Supplementary Tables S3–S5) in order to see the signaling in detail. These subpathways were assigned to various pathways in CRC and AD (Supplementary Tables S3–S5), suggesting that, in addition to inflammatory response inferred by our functional enrichment comparison, those pathways (not assigned to inflammation) remain largely unexplored in CRC or AD. Of interest, we found pathways previously unassociated with the two diseases, including Hedgehog signaling, axon guidance, ECM-receptor interaction, and WNT signaling (Table 3). In CRC, WNT3 facilitates crosstalk between the Hedgehog and Wnt signaling pathways (Table 3). Similarly, ECM-receptor interaction was oppositely regulated between the two diseases.

3.4. Opposite Signaling Pathway Expression between CRC and AD by Inflammation-Related Genes. The AD datasets were prepared from frontal cortex synaptoneurosomes and hippocampi. Both brain regions include neurons, as well as astrocytes and microglia [19, 20]. In our previous analysis [9] of the CRC dataset, immune cells were infiltrating. Considering immune cell involvement in the two diseases and their two opposite phenotypes, different inflammation-related molecule usage in signaling is self-evident.

So, we inspected the 16 genes' (in Table 1) differential usage of the CRC and AD networks (from Supplementary Figures S1–S3). Table 2 indicates that 16 genes were involved in extensive signaling transduction in both the CRC and AD networks, and all were inversely expressed between the two diseases.

Out of the 16 gene products, CD36 (a class B scavenger receptor) was found in microglia and vascular endothelial cells of AD patient brains [21]. Activation of CD36 and PPAR delta (gene symbol: PPARD, upregulated in both AD datasets in Table 2) resulted in FoxO1 activation in a functional study of muscle cells [22]. Considering that microglia are activated by FoxO1 [23], the two genes (*CD36* and *PPARD*) could be involved in inflammation of AD patient brains.

Another intriguing observation was the opposite expression of a cell growth (antiapoptosis) gene, *MAPK*, which was upregulated in CRC and downregulated in AD, while two apoptosis pathways genes, *FAS* (part of the extrinsic apoptosis pathway) and *BAD*, showed the opposite pattern (up in AD and down in CRC) (Table 2). This apoptosis versus cell survival relationship has been previously postulated to explain the inverse risk correlation between malignant and neurodegenerative diseases.

Table 3: Subpathways previously not associated with the two diseases. These subpathways were selected from the most significant 100 subpathways in each network. Subpathway (linear signaling flow) with fold-change (the numeral in parenthesis) of the disease group over the control group is represented in each dataset. The most significant 100 subpathways for each dataset are provided in Supplementary Tables S3–S5. The notation in the flow is "B <- A: A activates B" and "B |- A: A represses B."

KEGG pathway	GSE4107 (CRC) subpathway; *P* value	GSE1297 (AD) subpathway; *P* value	GSE12685 (AD) subpathway; *P* value
Hedgehog signaling (hsa04340)	PTCH1 (1.863) <- GLI2 (2.878) |- CSNK1G1 (0.587); 0.000035 WNT3 (3.147) <- GLI2 (2.878) |- CSNK1G1 (0.587); 0.000223		PTCH2 (0.938) <- GLI3 (0.682) |- GSK3B (1.513); 0.0015
Axon guidance (hsa04360)		PAK3 (0.732) <- RAC1 (0.943) |- PLXNB3 (1.627) <- SEMA4C (1.283); 0.0008	CFL1 (1.157) |- LIMK1 (0.896) <- PAK4 (0.871) <- RAC3 (0.892) <- PLXNA3 (0.954) <- FES (0.841); 0.0011
WNT signaling (hsa04310)	JUN (4.179) <- TCF7L1 (2.735) <- CTNNB1 (2.562) |- GSK3B (0.735) |- DVL3 (1.608) <- FZD10 (6.256) <- WNT3 (3.147) <- PORCN (1.279); 0.000114 JUN (4.179) <- TCF7L1 (2.735) <- CTNNB1 (2.562) |- GSK3B (0.735) |- DVL3 (1.608) <- APC2 (2.201) <- AXIN2 (2.307) <- CSNK1A1 (1.963); 0.00016		
Pathways in cancer (hsa05200)	MMP2 (3.031) <- JUN (4.179) <- MAPK1 (2.425) <- MAP2K1 (1.162) <- ARAF (4.631) <- HRAS (1.027) <- SOS1 (1.624) <- GRB2 (1.613) <- IGF1R (2.299) <- IGF1 (2.529); 0.000022		
ECM-receptor interaction (hsa04512)	SDC2 (3.091) <- TNC (9.557); 0.000026 SDC2 (3.091) <- FN1 (5.594); 0.000125	SDC3 (0.849) <- COL5A2 (0.162); 0.003	SDC1 (0.865) <- COL3A1 (0.865); 0.0017
Neurotrophin signaling (hsa04722)			BAD (1.279) |- AKT2 (0.856) <- PDK1 (0.943) <- PIK3CD (0.576) <- GAB1 (0.997) <- SHC2 (0.844) <- NTRK1 (0.945) <- NTF3 (0.784); 0.0008

4. Conclusions

In general, single gene expression analysis looks into highly differentially expressed genes under a certain cutoff (e.g., *P* value, fold-change). However, in real biological problems, signaling proteins involved in phenotype differences may not show a drastic expression-level change [9, 24]. Also, considering that phenotype change or disease results from dysregulation of complex relationships between biological components [25, 26], a strict cutoff usage in single gene analysis can miss signal flow. For example, some biological entities belonging to the flow would be filtered out under a certain cutoff. Along that line, we applied our previous systems biology method [9] for describing the interdependency underlined in the diseases. In summary, we found that inflammatory response was a very important mechanism in two diseases of opposite phenotypes, that is, cancer (cell survival) and Alzheimer's disease (cell death). The inflammation-related common genes between the diseases regulated opposite gene expression in various cell signaling in the two-disease networks. In other words, the inflammation-related genes in Table 2 utilized different pathways according to the disease states, leading to different signaling transductions. Further investigation of such networks could provide knowledge into the immunological bases for the progression of both of these devastating diseases.

Conflict of Interests

Curt Balch is the Chair of Bioscience Advising, IN, USA. This does not alter the result of the study and adherence to the journal publication policy. All the authors declare no potential competing interests.

Acknowledgments

This work was supported by grants from the National Cancer Center, Republic of Korea (Grant nos. NCC-1210460 and NCC-1510141-1 to Seungyoon Nam), and Korea Institute of Science and Technology Information (KISTI) (Grant no. P-14-SI-IA27 to Seok Jong Yu and Yongseong Cho), Republic of Korea. KISTI kindly provided Seungyoon Nam with a high-performance computing resource.

References

[1] M. Musicco, F. Adorni, S. di Santo et al., "Inverse occurrence of cancer and Alzheimer disease: a population-based incidence study," *Neurology*, vol. 81, no. 4, pp. 322–328, 2013.

[2] K. Ibáñez, C. Boullosa, R. Tabarés-Seisdedos, A. Baudot, and A. Valencia, "Molecular evidence for the inverse comorbidity between central nervous system disorders and cancers detected by transcriptomic meta-analyses," *PLoS Genetics*, vol. 10, no. 2, Article ID e1004173, 2014.

[3] M. J. Devine, H. Plun-Favreau, and N. W. Wood, "Parkinson's disease and cancer: two wars, one front," *Nature Reviews Cancer*, vol. 11, no. 11, pp. 812–823, 2011.

[4] M. I. Behrens, C. Lendon, and C. M. Roe, "A common biological mechanism in cancer and Alzheimer's disease?" *Current Alzheimer Research*, vol. 6, no. 3, pp. 196–204, 2009.

[5] G. P. Dunn, C. M. Koebel, and R. D. Schreiber, "Interferons, immunity and cancer immunoediting," *Nature Reviews Immunology*, vol. 6, no. 11, pp. 836–848, 2006.

[6] N. J. Maragakis and J. D. Rothstein, "Mechanisms of disease: astrocytes in neurodegenerative disease," *Nature Clinical Practice Neurology*, vol. 2, no. 12, pp. 679–689, 2006.

[7] V. H. Perry and C. Holmes, "Microglial priming in neurodegenerative disease," *Nature Reviews Neurology*, vol. 10, no. 4, pp. 217–224, 2014.

[8] Y. Hong, S. H. Kok, W. E. Kong, and Y. C. Peh, "A susceptibility gene set for early onset colorectal cancer that integrates diverse signaling pathways: implication for tumorigenesis," *Clinical Cancer Research*, vol. 13, no. 4, pp. 1107–1114, 2007.

[9] S. Nam and T. Park, "Pathway-based evaluation in early onset colorectal cancer suggests focal adhesion and immunosuppression along with epithelial-mesenchymal transition," *PLoS ONE*, vol. 7, no. 4, Article ID e31685, 2012.

[10] C. Williams, R. M. Shai, Y. Wu et al., "Transcriptome analysis of synaptoneurosomes identifies neuroplasticity genes overexpressed in incipient Alzheimer's disease," *PLoS ONE*, vol. 4, no. 3, Article ID e4936, 2009.

[11] E. M. Blalock, J. W. Geddes, K. C. Chen, N. M. Porter, W. R. Markesbery, and P. W. Landfield, "Incipient Alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 7, pp. 2173–2178, 2004.

[12] T. Wyss-Coray, "Inflammation in Alzheimer disease: driving force, bystander or beneficial response?" *Nature Medicine*, vol. 12, no. 9, pp. 1005–1015, 2006.

[13] A. Mantovani, P. Allavena, A. Sica, and F. Balkwill, "Cancer-related inflammation," *Nature*, vol. 454, no. 7203, pp. 436–444, 2008.

[14] S. Jiang, Y. Li, X. Zhang, G. Bu, H. Xu, and Y.-W. Zhang, "Trafficking regulation of proteins in Alzheimer's disease," *Molecular Neurodegeneration*, vol. 9, no. 1, article 6, 2014.

[15] L. Rajendran and W. Annaert, "Membrane trafficking pathways in Alzheimer's disease," *Traffic*, vol. 13, no. 6, pp. 759–770, 2012.

[16] C. F. Bento, C. Puri, K. Moreau, and D. C. Rubinsztein, "The role of membrane-trafficking small GTPases in the regulation of autophagy," *Journal of Cell Science*, vol. 126, no. 5, pp. 1059–1069, 2013.

[17] R. Sannerud, I. Declerck, A. Peric et al., "ADP ribosylation factor 6 (ARF6) controls amyloid precursor protein (APP) processing by mediating the endosomal sorting of BACE1," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 34, pp. E559–E568, 2011.

[18] J. Liu, N. Zhang, Q. Li et al., "Tumor-associated macrophages recruit CCR6$^+$ regulatory T cells and promote the development of colorectal cancer via enhancing CCL20 production in mice," *PLoS ONE*, vol. 6, no. 4, Article ID e19495, 2011.

[19] D. Most, L. Ferguson, Y. Blednov, R. D. Mayfield, and R. A. Harris, "The synaptoneurosome transcriptome: a model for profiling the emolecular effects of alcohol," *The Pharmacogenomics Journal*, 2014.

[20] J. H. Choi and M.-H. Won, "Microglia in the normally aged hippocampus," *Laboratory Animal Research*, vol. 27, no. 3, pp. 181–187, 2011.

[21] I. S. Coraci, J. Husemann, J. W. Berman et al., "CD36, a class B scavenger receptor, is expressed on microglia in Alzheimer's disease brains and can mediate production of reactive oxygen species in response to β-amyloid fibrils," *The American Journal of Pathology*, vol. 160, no. 1, pp. 101–112, 2002.

[22] Z. Nahlé, M. Hsieh, T. Pietka et al., "CD36-dependent regulation of muscle FoxO1 and PDK4 in the PPAR delta/beta-mediated adaptation to metabolic stress," *The Journal of Biological Chemistry*, vol. 283, no. 21, pp. 14317–14326, 2008.

[23] H. Dong, X. Zhang, X. Dai et al., "Lithium ameliorates lipopolysaccharide-induced microglial activation via inhibition of toll-like receptor 4 expression by activating the PI3K/Akt/FoxO1 pathway," *Journal of Neuroinflammation*, vol. 11, no. 1, article 140, 2014.

[24] E. Kotelnikova, N. Ivanikova, A. Kalinin, A. Yuryev, and N. Daraselia, "Atlas of signaling for interpretation of microarray experiments," *PLoS ONE*, vol. 5, no. 2, Article ID e9256, 2010.

[25] A.-L. Barabási and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," *Nature Reviews Genetics*, vol. 5, no. 2, pp. 101–113, 2004.

[26] S. Nam, H. R. Chang, K.-T. Kim et al., "PATHOME: an algorithm for accurately detecting differentially expressed subpathways," *Oncogene*, vol. 33, pp. 4941–4951, 2014.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 143712, 18 pages
http://dx.doi.org/10.1155/2015/143712

Review Article

The Current and Future Use of Ridge Regression for Prediction in Quantitative Genetics

Ronald de Vlaming[1] and Patrick J. F. Groenen[2]

[1]*Erasmus University Rotterdam Institute for Behavior and Biology, Department of Applied Economics, Erasmus School of Economics, Erasmus University Rotterdam, Postbus 1738, 3000 DR Rotterdam, Netherlands*
[2]*Econometric Institute, Erasmus School of Economics, Erasmus University Rotterdam, Postbus 1738, 3000 DR Rotterdam, Netherlands*

Correspondence should be addressed to Ronald de Vlaming; devlaming@ese.eur.nl

Received 28 November 2014; Accepted 24 December 2014

Academic Editor: Junwen Wang

In recent years, there has been a considerable amount of research on the use of regularization methods for inference and prediction in quantitative genetics. Such research mostly focuses on selection of markers and shrinkage of their effects. In this review paper, the use of *ridge regression* for prediction in quantitative genetics using *single-nucleotide polymorphism* data is discussed. In particular, we consider (i) the theoretical foundations of ridge regression, (ii) its link to commonly used methods in animal breeding, (iii) the computational feasibility, and (iv) the scope for constructing prediction models with nonlinear effects (e.g., *dominance* and *epistasis*). Based on a simulation study we gauge the current and future potential of ridge regression for prediction of human traits using genome-wide SNP data. We conclude that, for outcomes with a relatively simple genetic architecture, given current sample sizes in most cohorts (i.e., $N < 10{,}000$) the predictive accuracy of ridge regression is slightly higher than the classical *genome-wide association study* approach of *repeated simple regression* (i.e., one regression per SNP). However, both capture only a small proportion of the heritability. Nevertheless, we find evidence that for large-scale initiatives, such as biobanks, sample sizes can be achieved where ridge regression compared to the classical approach improves predictive accuracy substantially.

1. Introduction

The advent of large-scale molecular genetic data has paved the way for using these data to help predict, diagnose, and treat complex human diseases [1]. In recent years, the use of such data for the prediction of polygenic diseases and traits has become increasingly popular (e.g., [2–4]). This venue has proved successful even for traits such as educational attainment and cognitive performance [5, 6]. The vast majority of research into the genetic architecture of human traits and diseases is exploratory and considers the effects of at least hundreds of thousands of *single-nucleotide polymorphisms* (SNPs) on the outcome of interest [7].

Predictions based on molecular genetic data are typically constructed as a weighted linear combination of the available SNPs. This yields a so-called *polygenic risk score* [3] (*polygenic score*, *genetic risk score*, and *genome-wide score* [8]). *Multiple regression* (*ordinary least squares*, OLS) is a natural technique for estimating the weights of the predictors (SNPs) in this context but cannot be applied here: in general, the number of samples (N) available is far lower than the number of SNPs (P); typically, $N < 10{,}000$ and $P > 100{,}000$. OLS would yield a perfect in-sample prediction without any predictive value out of sample and would not allow drawing inferences on the weights of the SNPs, as they are nonunique. A commonly accepted solution to this problem is to carry out a *genome-wide association study* (GWAS), where one regresses the outcome of interest on each SNP separately. In this paper, we call this the *repeated simple regression* (RSR) approach.

Polygenic scores are typically constructed as the weighted sum of the SNPs with weights resulting from a GWAS using RSR. We raise four points of critique regarding this method. The first problem with this approach is that, in contrast to multiple regression, there is no search for the best linear combination over all SNPs jointly for predicting the outcome. A second, related, problem is that highly correlated SNPs (i.e.,

SNPs in strong *linkage disequilibrium*) repeatedly contribute very similar information, thereby distorting the risk score. For example, consider a set of ten perfectly correlated SNPs. In the RSR, they receive exactly the same weight. As the polygenic risk score is a weighted linear sum of the SNPs with the weights coming from RSR, these perfectly correlated SNPs contribute a factor ten stronger to the risk score than a single SNP capturing all information from that region does. This factor ten does not depend on the predictive power of the information in that region. A third problem is that the polygenic risk score can theoretically be correlated with *confounding variables* (*confounders*, *control variables*, and *controls*). For instance, SNPs can be correlated with the population structure. Therefore, the polygenic risk—being a linear combination of SNPs—can be correlated with the confounders. Usually, confounders, such as age and gender, are included as regressors in order to control for spurious relations through these covariates. However, we find that often in empirical work researchers do not control properly for the confounders in at least one of the many steps that lead from phenotype and genotype data to evaluation of the out-of-sample predictive accuracy of the polygenic risk score. A fourth problem is that the RSR approach is not able to handle even two-way interactions between the SNPs, as it would lead to a number of weights to be estimated that is quadratic in the number of SNPs, which is clearly computationally infeasible.

In this paper, we review the use of ridge regression (RR) [9] to tackle the four problems discussed above. The purpose of this paper is threefold. First, we discuss how prediction using RR can address the aforementioned four points of critique pertaining to a typical polygenic score, that is, how RR can be used to search for the best linear combination of SNPs jointly, to address the multicollinearity of SNPs [10, 11], and to account for the presence of confounding variables and of nonlinear SNP effects (e.g., [12–17]). Second, we review relevant work on ridge regression both in and outside the field genetics. Third, we assess the merits of prediction using ridge regression in the new domain of biobanks. That is, we predict the expected accuracy of ridge regression in large scale initiatives with over a 100,000 observations.

An important property of RR is that it cannot select a subset of predictors (e.g., SNPs). Other regularization methods related to RR are able to select a subset of predictors from a large set of predictors. Examples of such methods are the *least absolute shrinkage and selection operator* (LASSO), group LASSO [18], adaptive LASSO [19], and the elastic net [20].

In a GWAS, SNP selection is a desirable property when trying to find regions in the DNA that bear a causal influence on the outcome. However, there is mixed evidence for the claim that selection techniques in general improve the overall predictive accuracy of the polygenic score. Some studies suggest that preselection of markers (e.g., SNPs), based on either linkage disequilibrium or (in-sample) univariate association results, is detrimental to predictive accuracy (e.g., [3, 8, 11, 21]). Moreover, there is no conclusive evidence on the relative performance of RR-type methods and LASSO-type methods. For instance, using a simulation study, Ogutu et al. [22] find that LASSO-type methods outperform classic RR, whereas other studies find that RR outperforms LASSO and similar variable selection methods (e.g., [23–25]). A reasonable proposition is that the relative performance of RR and LASSO depends on trait architecture (e.g., [21, 26]). In particular, a low number of causal SNPs favor LASSO-type methods, whereas an intermediate or high number of causal variants favor RR-type methods. Regularization methods performing selection are computationally more involved and less amenable to incorporate nonlinear SNP effects than RR. For the above reasons, as well as our aim to provide a clear overview of RR, we focus in this paper primarily on RR.

The remainder of this paper is organized as follows. In Section 2, we present the theory underlying RR. In Section 3, we show that RR can be perceived as a method between OLS and RSR, leveraging the advantages of these two methods. Subsequently, in Section 4, we discuss the relation between RR and the best linear unbiased prediction used in animal breeding and the relation between RR and LASSO-type methods. In Section 5, we pay special attention to the effect standardization of SNP data has on the implicit assumptions about the genetic architecture of traits. As indicated, the feasibility of RR depends critically on the use of computationally efficient approaches. These will be discussed in Section 6. Related to this, in Section 7, we will discuss methods to tune the penalty parameter of RR. Following that, in Section 8, advanced RR techniques will be discussed, such as modelling nonlinear effects using RR, weighting SNPs differently, and incorporating information from earlier studies.

In order to assess the current and future use of ridge regression for prediction in quantitative genetics, we run a suite of simulations. The design of the simulations and the results are presented in Section 9. Based on these results we will estimate the effect sample size, the number of SNPs, the number of causal SNPs, and trait heritability have on the predictive accuracy of RR and the classical RSR approach. Using these estimates we will extrapolate how RR and RSR are expected to perform relative to each other in large scale studies (e.g., $N \geq 100{,}000$). Finally, in Section 10, we summarize the most important aspects of RR in the context of prediction in quantitative genetics and discuss our expectations for its future uses.

2. Ridge Regression

Using ridge regression (RR) for prediction in quantitative genetics was first proposed by Whittaker et al. [27]. RR can be understood as follows. Like regular *least-squares* methods RR minimizes a loss function that includes the sum of squared regression residuals. However, opposed to least squares, the loss function also includes a term consisting of positive penalty parameter λ times the model complexity, measured by the sum of squared regression weights [9]. This penalty prevents overfitting by shrinking the weights towards zero, ensuring that, even in case of multicollinearity and $P \gg N$, the estimator has a solution. The RR estimator has a simple analytical solution.

More formally, given a set of N individuals, P SNPs, and K confounders, a linear model for quantitative outcome vector

$\mathbf{y}$ ($N \times 1$), with a matrix of SNP data $\mathbf{X}$ ($N \times P$), and a matrix of confounders $\mathbf{Z}$ ($N \times K$) as predictors, is given by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \quad (1)$$

where $\boldsymbol{\beta}$ is the vector of SNP effects, $\boldsymbol{\gamma}$ the vector of effects of the confounders, and $\boldsymbol{\varepsilon}$ the phenotype noise.

In this particular case, we consider a large set of SNPs and a small set of potential confounders. Since one of our aims is to prevent any spurious relations via the confounders, we use a loss function that does not apply shrinkage to these. Therefore, the RR estimator minimizes

$$\mathscr{L}_{\text{RR}}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\gamma})^{\top} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\gamma}) + \lambda\boldsymbol{\beta}^{\top}\boldsymbol{\beta}. \quad (2)$$

Under this loss function, the RR estimator of $\boldsymbol{\beta}$ is given by

$$\widehat{\boldsymbol{\beta}}_{\text{RR}} = (\mathbf{X}^{\top}\mathbf{M}_{\mathbf{Z}}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\top}\mathbf{M}_{\mathbf{Z}}\mathbf{y}, \quad (3)$$

where $\mathbf{M}_{\mathbf{Z}} = \mathbf{I} - \mathbf{Z}(\mathbf{Z}^{\top}\mathbf{Z})^{-1}\mathbf{Z}^{\top}$ is the projection matrix, removing the effects of the confounding variables. The larger the λ is, the more the shrinkage is applied. When $\lambda = 0$, RR corresponds to OLS. The OLS estimator only exists if $\text{rank}(\mathbf{X}^{\top}\mathbf{M}_{\mathbf{Z}}\mathbf{X}) = P$, meaning that there is no perfect collinearity amongst the SNPs and that $P \leq N$. However, in a GWAS, almost invariably $P \gg N$. Therefore, OLS cannot be applied in this context. However, the RR estimator has a solution for any $\lambda > 0$, even if $P \gg N$.

Heteroskedastic ridge regression (HRR) is a generalization of RR, where each SNP p receives a different amount of shrinkage, $\lambda_p \geq 0$. The loss function of HRR is given by

$$\mathscr{L}_{\text{HRR}}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\gamma})^{\top} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\gamma}) + \lambda\boldsymbol{\beta}^{\top}\boldsymbol{\Lambda}\boldsymbol{\beta}, \quad (4)$$

where $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_P)$. The corresponding estimator is given by

$$\widehat{\boldsymbol{\beta}}_{\text{HRR}} = (\mathbf{X}^{\top}\mathbf{M}_{\mathbf{Z}}\mathbf{X} + \lambda\boldsymbol{\Lambda})^{-1}\mathbf{X}^{\top}\mathbf{M}_{\mathbf{Z}}\mathbf{y}. \quad (5)$$

The $P \times P$ matrix $\mathbf{X}^{\top}\mathbf{M}_{\mathbf{Z}}\mathbf{X}$ in (3) and (5) can be regarded as a map of the estimated correlation (linkage disequilibrium) between markers. OLS takes this linkage disequilibrium fully into account at the expense of overfitting the data, whereas RSR completely ignores it. For this reason, when constructing a polygenic score, RSR is often used in combination with a heuristic procedure, known as linkage disequilibrium pruning, which selects SNPs that are not too strongly correlated. As is shown in the next section, RR leverages the two extremes of OLS and RSR. Therefore, opposed to RSR, RR does not require the *a priori* selection of SNPs; RR is able to handle linkage disequilibrium between markers [10, 11].

RR is expected to perform particularly well under a scenario where a substantial proportion of the SNPs is expected to contribute to the phenotype and where each contribution is small.

3. The Limiting Cases of Ridge Regression

Varying the penalty weight, λ, allows specifying special cases of RR. Prediction by RR can be perceived as a method that lies between prediction based on OLS estimates considering all SNPs jointly and OLS estimates considering each SNP separately. By definition of RR [9], for sufficiently low shrinkage, the RR estimates converge to the multiple regression estimates [10], provided these are unique. For sufficiently high shrinkage a RR prediction score is equivalent to an RSR prediction score, in terms of the proportion of variance accounted for by the respective scores. For ease of notation, we assume in this section that there are no confounders $\mathbf{Z}$.

To establish the aforementioned relations, two conditions are needed. First, the measure of predictive accuracy is independent of scale. That is, given an out-of-sample quantitative outcome vector ($\mathbf{y}_2$) and its prediction ($\widehat{\mathbf{y}}_2$), the accuracy measure should be such that for any coefficient $b > 0$ the accuracy of prediction $\widehat{\mathbf{y}}_2$ is identical to that of prediction $\widehat{\mathbf{y}}_2^{*} = b\widehat{\mathbf{y}}_2$. An example of such a measure is the R^2 of an outcome and its prediction. The second condition is that SNP data are standardized, such that each SNPs p has mean zero ($\mathbf{x}_p^{\top}\boldsymbol{\iota} = 0$, where $\boldsymbol{\iota}^{\top} = (1, \ldots, 1)$) and equal standard deviation ($\mathbf{x}_p^{\top}\mathbf{x}_p = c$, where c is a scalar).

Consider the prediction of $\mathbf{y}_2$ based on $N_2 \times P$ out-of-sample genotype matrix $\mathbf{X}_2$, using in-sample RR estimates $\widehat{\boldsymbol{\beta}}_{\text{RR}}$. This prediction is given by $\widehat{\mathbf{y}}_2 = \mathbf{X}_2\widehat{\boldsymbol{\beta}}_{\text{RR}}$. Based on the first condition, we can multiply the prediction $\widehat{\mathbf{y}}_2$ by $b = (1 + \lambda)$. This is equivalent to inflating the RR estimates by $(1 + \lambda)$ instead of inflating the predictions. Thus, we can take $\widehat{\boldsymbol{\beta}}_{\text{RR}}^{*} = (1 + \lambda)\widehat{\boldsymbol{\beta}}_{\text{RR}}$. This yields

$$\widehat{\boldsymbol{\beta}}_{\text{RR}}^{*} = (\alpha\mathbf{I} + (1 - \alpha)\mathbf{X}^{\top}\mathbf{X})^{-1}\mathbf{X}^{\top}\mathbf{y}, \quad (6)$$

where $\alpha = (1 + \lambda)^{-1}\lambda \in (0, 1)$. The OLS estimator considering all SNPs jointly is given by

$$\widehat{\boldsymbol{\beta}}_{\text{OLS}} = (\mathbf{X}^{\top}\mathbf{X})^{-1}\mathbf{X}^{\top}\mathbf{y}. \quad (7)$$

Thus, it follows that when α goes to zero (i.e., λ goes to zero), the RR estimator goes to the OLS estimator. Moreover, as α goes to one (i.e., λ becomes sufficiently large), the inflated RR estimator goes to $\mathbf{X}^{\top}\mathbf{y}$.

Using the condition of having standardized SNPs, we can rewrite the RSR for SNP p as $\widehat{\beta}_p = \mathbf{x}_p^{\top}\mathbf{y}$, where $\mathbf{x}_p$ is the standardized genotype vector of SNP p. This expression can be vectorized over all SNPs as $\widehat{\boldsymbol{\beta}}_{\text{RSR}} = \mathbf{X}^{\top}\mathbf{y}$. From this, it follows that the inflated RR estimates approach the RSR estimates as λ becomes sufficiently large.

4. Related Methods

Prediction using RR is related to the predictions that arise under a widely used simple mixed linear model, commonly referred to as the *animal model*. In such a model, expected genetic relatedness is mapped to phenotypic relatedness. Usually pedigree information is used to infer genetic relatedness. However, with the advent of genome-wide molecular data, mixed models that use SNPs to estimate genetic relatedness have been proposed (e.g., see Yang et al. [28]). In most mixed models using SNPs, the prior assumption is that SNP effects

are normally distributed with mean zero and variance $\sigma^2_{\boldsymbol{\beta}}$, and the error terms in the phenotype are also normally distributed with variance $\sigma^2_{\boldsymbol{\varepsilon}}$.

To understand the relation between RR and mixed models, consider the following mixed linear model

$$\begin{aligned}\mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon},\\ \boldsymbol{\beta} &\sim \mathcal{N}\left(\mathbf{0}, \sigma^2_{\boldsymbol{\beta}}\mathbf{I}_P\right),\\ \boldsymbol{\varepsilon} &\sim \mathcal{N}\left(\mathbf{0}, \sigma^2_{\boldsymbol{\varepsilon}}\mathbf{I}_N\right),\end{aligned} \tag{8}$$

where $\sigma^2_{\boldsymbol{\beta}}$ is the SNP effect variance and $\sigma^2_{\boldsymbol{\varepsilon}}$ the noise variance. In this model the effects of the confounders, $\mathbf{Z}$, are assumed to be fixed. For the remainder of this section we ignore the confounders for ease of notation. The parameters $\sigma^2_{\boldsymbol{\varepsilon}}$ and $\sigma^2_{\boldsymbol{\beta}}$ can be estimated using, for instance, *maximum likelihood*, *restricted maximum likelihood* [29], or *expectation maximization* [30]. Alternatively, these parameters can be fixed by using prior information from other data sets; see, for instance, Hofheinz et al. [31].

Consider conditional expectations $\mathbb{E}[\boldsymbol{\beta} \mid \mathbf{y}]$ and $\mathbb{E}[\mathbf{y}_2 \mid \mathbf{y}]$. In a mixed linear model such expectations are known as the best linear unbiased prediction (BLUP) [32–36]. BLUP was first proposed by Henderson [32] in order to obtain estimates of the so-called *breeding values*, that is, the part of the phenotype that can be attributed to genetic variation.

Provided that the RR penalty $\lambda = \sigma^2_{\boldsymbol{\varepsilon}}/\sigma^2_{\boldsymbol{\beta}}$, the BLUP of SNP effects [28, 37, 38] is equivalent to the RR estimator. Under that same condition, the BLUP of the SNP-based breeding values is equivalent to RR prediction. Such *genomic estimated breeding values* [38] contain the part of the phenotype that can be attributed to the genetic variation in the genotyped markers.

To understand this equivalence, first we rewrite the RR estimator in (3). By applying the *Sherman-Morrison-Woodbury formula* [39, 40] to the $P \times P$ inverse of $\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}$, we obtain

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{\text{RR}} &= \frac{1}{\lambda}\left[\mathbf{I}_P - \mathbf{X}^\top\left(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_N\right)^{-1}\mathbf{X}\right]\mathbf{X}^\top\mathbf{y}\\ &= \frac{1}{\lambda}\mathbf{X}^\top\left[\mathbf{I}_N - \left(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_N\right)^{-1}\mathbf{X}\mathbf{X}^\top\right]\mathbf{y}\\ &= \frac{1}{\lambda}\mathbf{X}^\top\left(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_N\right)^{-1}\left[\left(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_N\right) - \mathbf{X}\mathbf{X}^\top\right]\mathbf{y}\\ &= \mathbf{X}^\top\left(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I}_N\right)^{-1}\mathbf{y}.\end{aligned} \tag{9}$$

Second, by rewriting (8) in terms of the joint distribution of $\mathbf{y}$ and $\boldsymbol{\beta}$:

$$\begin{pmatrix}\mathbf{y}\\ \boldsymbol{\beta}\end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\sigma^2_{\boldsymbol{\beta}}\mathbf{X}\mathbf{X}^\top + \sigma^2_{\boldsymbol{\varepsilon}}\mathbf{I}_N & \sigma^2_{\boldsymbol{\beta}}\mathbf{X}\\ \sigma^2_{\boldsymbol{\beta}}\mathbf{X}^\top & \sigma^2_{\boldsymbol{\beta}}\mathbf{I}_P\end{bmatrix}\right), \tag{10}$$

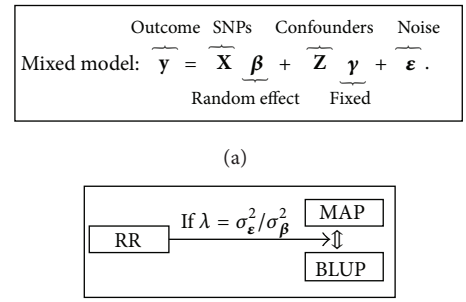


Figure 1: Diagram (b) showing the relation between estimation of SNP effects using *ridge regression* (RR), the *best linear unbiased prediction* (BLUP), and *maximum a posteriori* (MAP) estimation, under the specified *mixed linear model* (a).

the BLUP of $\boldsymbol{\beta}$ is given by the expectation of $\boldsymbol{\beta}$ conditional on $\mathbf{y}$ [17]. This yields

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{\text{BLUP}} &= \sigma^2_{\boldsymbol{\beta}}\mathbf{X}^\top\left(\sigma^2_{\boldsymbol{\beta}}\mathbf{X}\mathbf{X}^\top + \sigma^2_{\boldsymbol{\varepsilon}}\mathbf{I}_N\right)^{-1}\mathbf{y}\\ &= \mathbf{X}^\top\left(\mathbf{X}\mathbf{X}^\top + \frac{\sigma^2_{\boldsymbol{\varepsilon}}}{\sigma^2_{\boldsymbol{\beta}}}\mathbf{I}_N\right)^{-1}\mathbf{y}.\end{aligned} \tag{11}$$

Clearly, when $\lambda = \sigma^2_{\boldsymbol{\varepsilon}}/\sigma^2_{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\beta}}_{\text{RR}} = \widehat{\boldsymbol{\beta}}_{\text{BLUP}}$.

In addition, from a Bayesian perspective the posterior mode of the distribution of SNP effects (i.e., the mode of the distribution conditional on a training set) can also be used as point estimator. Estimation using the posterior mode is known as maximum a posteriori (MAP) estimation. However, due to the normality of $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$ the mode coincides with the conditional expectation $\mathbb{E}[\boldsymbol{\beta} \mid \mathbf{y}]$. Therefore, MAP estimation of $\boldsymbol{\beta}$ in (8) is equivalent to BLUP.

Consequently, there exists a λ such that the RR estimator of SNP effects is equivalent to its BLUP [16, 41] and by extension to the MAP estimator. The diagram in Figure 1 summarizes the relations between RR, BLUP, and MAP.

4.1. SNP Selection Using LASSO-Type Methods. An important feature that RR lacks is the selection of SNPs. LASSO-type methods, such as the LASSO, group LASSO, adaptive LASSO, and the elastic net, are able to select SNPs. The key to achieving SNP selection is to include an L_1 penalty, that is, adding a penalty consisting of a penalty parameter, λ, times $\|\boldsymbol{\beta}\|_1 = |\beta_1| + \cdots + |\beta_P|$. The loss function of the LASSO is given by

$$\mathscr{L}_{\text{LASSO}}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\gamma}\right)^\top\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\gamma}\right) + \lambda\left\|\boldsymbol{\beta}\right\|_1. \tag{12}$$

This function is highly similar to the RR loss function in (2). The most important property of the LASSO is that it performs variable selection; that is, for a sufficiently large λ many of the SNP coefficients β_p will be zero. The higher the λ is, the fewer the nonzero SNP effects are obtained by the LASSO. Moreover, this method also shrinks the nonzero coefficients, that is, the estimated effects of the selected SNPs.

The loss function of the elastic net [20] is obtained by taking a convex combination of $\boldsymbol{\beta}^\top\boldsymbol{\beta}$ and $\|\boldsymbol{\beta}\|_1$ as penalty; that is,

$$\mathscr{L}_{\text{net}}(\boldsymbol{\beta},\boldsymbol{\gamma}) = (\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\boldsymbol{\gamma})^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\boldsymbol{\gamma}) + \lambda\left(\alpha\boldsymbol{\beta}^\top\boldsymbol{\beta}+(1-\alpha)\|\boldsymbol{\beta}\|_1\right), \tag{13}$$

with $\lambda \geq 0$ and $\alpha \in [0,1]$. The elastic-net method preserves SNP selection, while allowing more than N of P SNPs to be selected. Taking a convex combination of the two norms hardly increases the computational costs of solving this problem, when compared to solving the LASSO problem [20]. Typically, the LASSO solution is obtained by means of the least-angle regression algorithm [42]. This algorithm entails an iterative procedure, where at most one SNP can enter the model at a time. Therefore, LASSO-type methods are computationally far more involved than RR-type methods.

Finally, the group LASSO [18] splits the P predictors in G mutually disjoint groups, with p_g predictors in group g, and associated effects $\boldsymbol{\beta}_g$, for groups $g = 1,\ldots,G$. The group LASSO minimizes

$$\mathscr{L}_{\text{group}}(\boldsymbol{\beta},\boldsymbol{\gamma}) = (\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\boldsymbol{\gamma})^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\boldsymbol{\gamma}) + \lambda\sum_{g=1}^{G}\sqrt{\boldsymbol{\beta}_g^\top\boldsymbol{\beta}_g}. \tag{14}$$

Each group can be chosen, for instance, to represent a single gene in terms of its SNPs. The group LASSO induces sparsity at the group level (e.g., a gene is either included as a whole or wholly excluded), whereas within a group the individual regressors receive an L_2 penalty. To the best of our knowledge, Sabourin et al. [43] provide the first, and so far only, application of a (modified) group LASSO using SNP data to construct polygenic scores. In this study, each SNP is considered as a group, with two effects: an additive and a dominance effect. In a simulation with mild to strong dominance, this method improves accuracy, compared to an RSR-type approach [43]. For a detailed comparison of LASSO-type methods and RR, we refer to Hastie et al. [44].

5. The Implications of Standardizing SNPs

In the preceding sections, we have only considered SNP standardization as a tool to show that RR can be perceived as a method between the classical GWAS approach and the OLS approach considering all SNPs jointly. However, SNP standardization is often used in the mixed linear model in (8).

The reason for this is that standardization has a profound effect on the implicit assumptions about the effect sizes of SNPs. We show in this section that the standardization we use is equivalent to HRR applied to raw genetic data, where SNPs measuring rare variants receive less shrinkage than SNPs measuring common variants.

More specifically, let $\mathbf{G}$ (resp., $\mathbf{G}_2$) denote raw SNP data in sample (out of sample) that has already been mean-centered but not yet standardized to have the same variance. The standardized data $\mathbf{X}$ in Section 3 can now be obtained by postmultiplying $\mathbf{G}$ by a diagonal matrix $\mathbf{D}$. That is, $\mathbf{X} = \mathbf{GD}$, where

$$\mathbf{D} = \operatorname{diag}\left(\left\{\sqrt{\frac{N-1}{\mathbf{x}_p^\top\mathbf{x}_p}}\right\}_{p=1,\ldots,P}\right). \tag{15}$$

Under the reasonable assumption that only SNPs are considered for which in-sample variation occurs, this matrix $\mathbf{D}$ is invertible.

By applying this transformation in both the training and test set, RR prediction based on standardized data is given by

$$\begin{aligned}\hat{\mathbf{y}}_2 &= \mathbf{X}_2\left(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I}\right)^{-1}\mathbf{X}^\top\mathbf{y}\\ &= \mathbf{G}_2\mathbf{D}\left(\mathbf{D}\mathbf{G}^\top\mathbf{G}\mathbf{D}+\lambda\mathbf{I}\right)^{-1}\mathbf{D}\mathbf{G}^\top\mathbf{y}\\ &= \mathbf{G}_2\left(\mathbf{G}^\top\mathbf{G}+\lambda\mathbf{D}^{-2}\right)^{-1}\mathbf{G}^\top\mathbf{y}.\end{aligned} \tag{16}$$

This shows that RR applied to standardized SNP data is equivalent to HRR, with $\boldsymbol{\Lambda} = \mathbf{D}^{-2}$, applied to raw genotype data. Here, the SNP-specific shrinkage depends on the amount of SNP variation. This type of shrinkage implicitly assumes that the standardized SNPs have homoskedastic effects, whereas the underlying raw genotypes (i.e., the count data) have effects of which the variance decreases with minor allele frequency. That is, rare alleles are assumed to have larger effects on average than common variants. For a qualitative treatment of the relation between allele frequency and expected effect sizes, see, for instance, Manolio et al. [45].

To be more precise, this type of shrinkage corresponds to the implicit assumption that the variance of the effect of raw SNP p, denoted by $\sigma^2_{\beta_p}$, with allele frequency f_p, is proportional to $(2f_p(1-f_p))^{-1}$. This assumption implies that when f_p is close to one or zero, the variance of the effect size is expected to be large, whereas for f_p close to 50% the variance of the effect size attains its minimum.

Naturally, raw SNP effect variances responding differently to allele frequency can be conceived. As indicated by Manolio et al. [45] such relations depend on the effect of the trait under consideration, on the fitness of the individual. Therefore, a natural extension would be to consider HRR with $\boldsymbol{\Lambda} = \mathbf{D}^\alpha$. Here $\alpha = 0$ corresponds to a trait for which allele frequency is independent of effect size and $\alpha = -2$ corresponds to the relation described before. Moreover, $-2 < \alpha < 0$ describes a trait for which there is a slight relation between allele frequency and effect size. It is interesting to note that $\alpha > 0$ corresponds to a trait where diversity is an asset, that is, a trait in which variants causing phenotypic divergence between individuals tend to become common. Finally, $\alpha < -2$ would correspond to a trait for which there has been strong selection pressure causing convergence; only very rare variants are expected to have a large effect. Thus, in future work α can be considered as an additional hyperparameter which might boost predictive accuracy and of which the estimate would reveal something about the selection pressure regarding the trait under consideration. The same type of transformation has been proposed by Speed et al. [46] for improving estimation of SNP-based heritability in a mixed linear model.

6. Computational Costs

The main hurdle in computing RR predictions is estimating the P parameters, when $P \gg N$. In particular, a naive approach requires solving a system with P unknowns. However, RR can be implemented in a computationally efficient way. When $P > N$, using dimensionality reduction techniques the complexity of RR can be reduced from $\mathcal{O}(P^3)$ to $\mathcal{O}(PN^2)$ in case one is interested in the estimated effects [47].

Moreover, if the focus lies solely on obtaining predictions, a nonparametric representation of RR reveals the fact that a dual formulation exists, which can be perceived as solving a linear model with N unknowns [48]. Solving such a system has a complexity slightly less than $\mathcal{O}(N^3)$. Building on this computationally efficient approach, RR can also efficiently control for confounders, both in sample and out of sample.

Finally, when considering a wide array of values of λ, RR can be reformulated to generate predictions for all values of λ jointly by exploiting the properties of the eigendecomposition of an $N \times N$ matrix, thereby yielding a complexity of $\mathcal{O}(N^3)$.

To understand these reductions in computational costs, consider the RR estimator in (9), used to show equivalence of RR and the BLUP. Premultiplying this expression by $\mathbf{X}_2$, the out-of-sample prediction is given by

$$\widehat{\mathbf{y}}_2 = \mathbf{X}_2\mathbf{X}^\top \left(\mathbf{X}\mathbf{X}^\top + \lambda \mathbf{I}_N\right)^{-1} \mathbf{y}. \tag{17}$$

As discussed, accounting for confounding variables is important. Let $\mathbf{Z}$ be the in-sample $N \times K$ matrix of confounders and $\mathbf{Z}_2$ the out-of-sample $N_2 \times K$ matrix of confounders. By replacing $\mathbf{X}$ by $\mathbf{X}^* = \mathbf{M}_\mathbf{Z}\mathbf{X}$ and $\mathbf{X}_2$ by $\mathbf{X}_2^* = \mathbf{M}_{\mathbf{Z}_2}\mathbf{X}_2$, where $\mathbf{M}_\mathbf{C}$ is the projection matrix removing the effects of $\mathbf{C}$, we find that

$$\widehat{\mathbf{y}}_2 = \mathbf{A}_{21}^* \left(\mathbf{A}^* + \lambda_{\text{GRM}} \mathbf{I}_N\right)^{-1} \mathbf{y}, \tag{18}$$

where $\mathbf{A}^* = \mathbf{M}_\mathbf{Z}\mathbf{A}\mathbf{M}_\mathbf{Z}$ and $\mathbf{A}_{21}^* = \mathbf{M}_{\mathbf{Z}_2}\mathbf{A}_{21}\mathbf{M}_\mathbf{Z}$, $\mathbf{A} = P^{-1}\mathbf{X}\mathbf{X}^\top$ and $\mathbf{A}_{21} = P^{-1}\mathbf{X}_2\mathbf{X}^\top$, and $\lambda_{\text{GRM}} = P^{-1}\lambda$. Therefore, one can correct for covariates by simply pre- and postmultiplying $N_{(2)} \times N$ matrices, by appropriate projection matrices.

Matrices $\mathbf{A}$ and $\mathbf{A}_{21}$ both have the interpretation of a SNP-based *genetic relationship matrix* (GRM) [28], measuring the genetic similarity of individuals in the space of additive SNP effects.

Given the eigendecomposition

$$\mathbf{A}^* = \mathbf{Q}\,\text{diag}\left(\{\theta_i\}_{i=1,\ldots,N}\right)\mathbf{Q}^\top, \tag{19}$$

RR prediction can be written as

$$\widehat{\mathbf{y}}_2 = \mathbf{A}_{21}^*\mathbf{Q}\,\text{diag}\left(\left\{\frac{1}{\theta_i + \lambda_{\text{GRM}}}\right\}_{i=1,\ldots,N}\right)\mathbf{Q}^\top\mathbf{y}. \tag{20}$$

If $P \gg N$, this approach is far more efficient than the naive approach to RR prediction. GRMs can be computed efficiently in packages such as `PLINK` 1.9 [49] and `GCTA` [28]. The most involved step in the prediction procedure is finding the eigendecomposition of $\mathbf{A}^*$.

7. Tuning and Interpreting λ

So far, it was assumed that the penalty strength parameter λ is given. However, in most applications of RR the optimal λ is not known in advance. Here, we discuss three ways for choosing λ.

The dominant approach in the machine learning literature for tuning λ is by maximizing out-of-sample predictive accuracy of RR using *cross-validation* (CV). In CV one considers a fine grid $\mathscr{L}$ of potential values of λ. The data are randomly split in a (small) test set (e.g., 10% of the sample) and CV set (90%). To the CV set one applies K-fold CV (e.g., $K = 10$), meaning that one splits the CV sample randomly in K blocks of (approximately) equal size. In each fold $K - 1$ blocks are considered as CV training set and the remaining block as CV test set. Using RR for all values of $\lambda \in \mathscr{L}$, predictions in the CV test set are generated. Each block is the CV test set precisely once. After the K-folds, the predictive accuracy over all CV test sets is evaluated for all $\lambda \in \mathscr{L}$. Now, $\widehat{\lambda}$ is set to maximize the cross-validation accuracy. Finally, using $\widehat{\lambda}$ the predictive accuracy in the final test is considered, using the full CV set as training data. For a more detailed treatment of CV, see, for instance, Hastie et al. [44].

Nested cross-validation (NCV) is a natural extension of CV, where the sample is randomly split in S "super"-blocks of approximately equal size (e.g., $S = 10$) and where there are S "super"-folds. In each superfold, one block is considered as final test set and $S - 1$ other blocks as CV set. To this CV set and test set one applies regular K-fold CV. Each superblock is used as final test set precisely once.

Classical CV is used to fit the model and to assess its predictive accuracy; one can judge the merits of a set of values of the hyperparameter by means of the CV procedure and apply the optimal value to a new part of the sample which has not yet been considered. Using NCV one can test whether the hyperparameter and accuracy that result from classical CV are robust; NCV can show the amount of variation in either of these over the "super"-folds.

CV requires a computationally efficient strategy since a different set of RR predictions will result for each different value of λ. However, a large set of different values of λ can be evaluated in one step at nearly the same costs of evaluating a single value of λ. This approach avoids computing a full RR solution for each λ separately. To see this, the formulation of RR prediction in (20) is highly relevant. In this equation, the eigendecomposition of $\mathbf{A}^*$ is independent of λ. Thus, predictions for each $\lambda \in \{\lambda_1, \ldots, \lambda_L\}$ can be obtained by the following equation:

$$\widehat{\mathbf{Y}}_2 = \mathbf{A}_{2,1}^*\mathbf{Q} \cdot \left[\begin{pmatrix} (\theta_1 + \lambda_1)^{-1} & \cdots & (\theta_1 + \lambda_L)^{-1} \\ \vdots & \ddots & \vdots \\ (\theta_N + \lambda_1)^{-1} & \cdots & (\theta_N + \lambda_L)^{-1} \end{pmatrix} \circ \left((\mathbf{Q}^\top\mathbf{y})\,\boldsymbol{\iota}^\top\right)\right], \tag{21}$$

where $\boldsymbol{\iota}^\top = (1, \ldots, 1)$ and "$\circ$" denotes the element-wise (Hadamard) product. A `MATLAB` implementation of this approach

```
(1) % rdgpred  Efficient prediction using ridge regression.
(2) %  rdgpred(Y,A,A21,L) returns ridge regression predictions in test set.
(3) %  Vector Y contains outcome training set, matrix A similarity measures
(4) %  in training set (e.g., A=XX' for N-by-P input matrix X), matrix A21
(5) %  similarity individuals in test set (rows) and training set (columns),
(6) %  and vector L the penalty values to consider.
(7) %
(8) %  rdgpred(Y,A,A21,L,Z,Z2) first corrects A and A21 for confounders.
(9) %  Matrix Z contains confounders in traing set and Z2 those in test set.
(10) %
(11) %   Author    : R de Vlaming and PJF Groenen
(12) %  Institute: Erasmus School of Economics    Date: November 25, 2014
(13) function Y2 = rdgpred(Y,A,A21,L,Z,Z2)
(14)
(15) P = numel(L);      % find size of set of penalties
(16) N = numel(Y);      % find size of training set
(17) N2 = size(A21,1); % find size of test set
(18)
(19) if nargin > 5      % correct similarities if confounders present
(20)   M = eye(N) - Z*inv(Z'*Z)*Z';       % anti projection matrix of Z
(21)   M2 = eye(N2) - Z2*inv(Z2'*Z2)*Z2'; % anti projection matrix of Z2
(22)   A = M*A*M;                          % adjust similarties A
(23)   A21 = M2*A21*M;                     % adjust similarties A21
(24) end
(25)
(26) [Q,D] = eig(A); D = diag(D); % obtain eigenvecs Q and eigenvalues D of A
(27)
(28) % for each eigenvalue E (rows) and lambda S (cols) find 1/(E+S)
(29) D = 1./(repmat(D,1,P) + repmat(L(:)',N,1));
(30)
(31) % predict for observations in test set (rows) for each lambda (cols)
(32) QTY = repmat((Y'*Q)',1,P);
(33) Y2 = A21*(Q*(D.*QTY));
(34) Y2 = real(Y2); % remove imaginary part due to numerical imprecession
(35)
(36) end
```

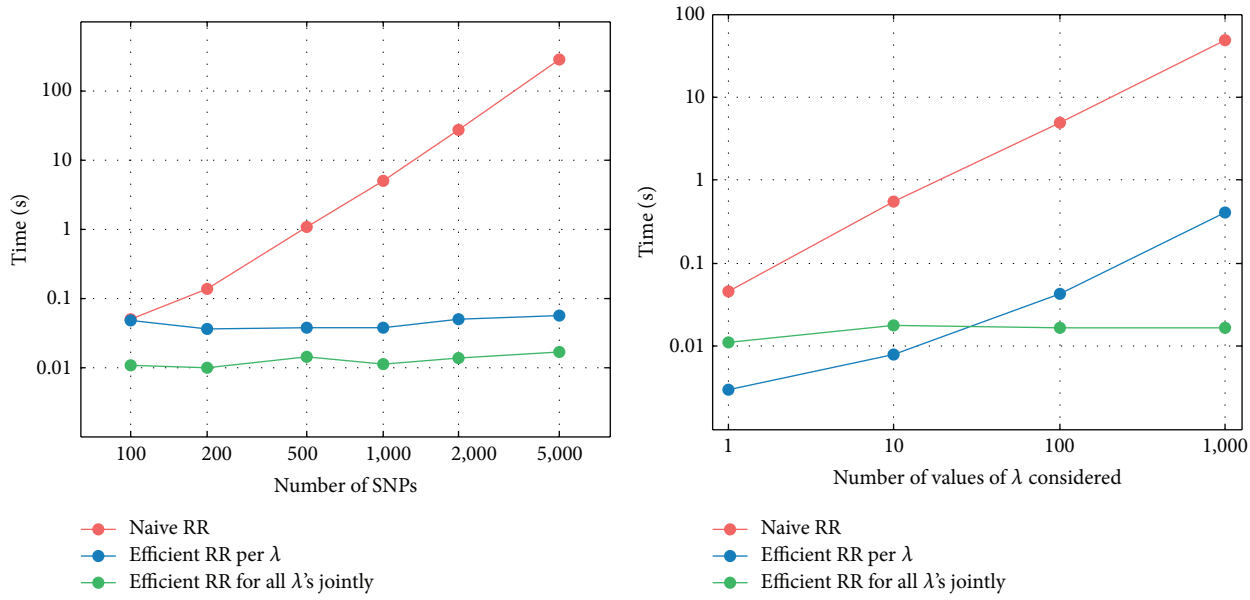
Algorithm 1: MATLAB code for efficient ridge regression prediction: rdgpred.m.

to RR prediction is provided in Algorithm 1. The computation of the eigendecomposition of $\mathbf{A}^*$ has a computational complexity of $\mathcal{O}(N^3)$. Given this decomposition, the prediction consists of $(N_2 + 3)NL + (L + 1)N^2$ simple operations such as multiplication and addition of scalars.

To illustrate the differences in the respective approaches to RR, Figure 2 shows the CPU time for (i) the naive approach in (3) which involves solving P unknowns, (ii) the dual formulation in (18) which requires solving L systems with N unknowns each, and (iii) the dual formulation in (21) solving for all values of λ jointly. These results are obtained by applying the approaches to simulated data, with baseline settings $N = 100$, $N_2 = 10$, $P = 1000$, and $L = 100$, and by varying the levels of the factors N and L, one factor at a time. In order to ensure no approach has an advantage in terms of preprocessing of the data (e.g., constructing $P^{-1}\mathbf{X}\mathbf{X}^\top$ and its eigendecomposition) all reported CPU times include these preprocessing steps.

In Figure 2(a), we see that as the number of SNPs P increases the time required by the naive approach keeps growing at a fixed rate, whereas the time required by the dual approaches remains unchanged. Moreover, the approach considering all values of λ jointly outperforms the dual approach solving L separate systems. When sample size N is relatively large compared to P the dual formulations lose their advantage compared to the naive approach. This is not surprising: when $N > P$ the dual formulation requires solving more unknowns than the naive approach. Concordantly, when faced with data in which $N \leq P$ one can apply the dual approach, and when $N > P$ one can use the classical approach to RR. Figure 2(b) shows that for a very small set of λ's the dual formulation solving L systems with N unknowns is faster than the formulation solving for all values of λ jointly. However, the CPU time required by the former approach increases continuously with L, whereas the CPU time of the method considering all λ's jointly hardly changes. When $L \geq 10$, the latter method attains a better CPU time than the former method does.

The second method for setting λ is based on the mixed model in (8). In this model, the optimal hyperparameter is



(a) Varying number of SNPs P, $N = 100$, $N_2 = 10$, and $L = 100$

(b) Varying number of λ's, $N = 100$, $N_2 = 10$, and $P = 1{,}000$

FIGURE 2: CPU time in seconds of prediction using naive RR (red), efficient RR for each λ separately (blue), and efficient RR considering all λ's jointly (green).

a function of σ^2_ε and σ^2_β. Therefore, one can estimate the mixed linear model using methods such as (restricted) maximum likelihood [28, 29] and take $\lambda = \sigma^2_\varepsilon / \sigma^2_\beta$.

Finally, one can use an existing heritability estimate of the trait under consideration. Given the following definition of SNP-based heritability,

$$h^2_{\text{SNP}} = \frac{P\sigma^2_\beta}{P\sigma^2_\beta + \sigma^2_\varepsilon}, \tag{22}$$

provided the SNP data are standardized as Z-scores, it is shown by Hofheinz et al. [31] that the RR shrinkage parameter λ can be written as a function of the SNP-based heritability. Specifically, simple algebra shows that, under the above definition of SNP-based heritability,

$$\lambda = P\left(\frac{1}{h^2_{\text{SNP}}} - 1\right). \tag{23}$$

This implies that heritability estimates can be used to set λ [31]. When using a GRM ($P^{-1}\mathbf{X}\mathbf{X}^\top$) to carry out RR prediction, the corresponding shrinkage parameter $\lambda_{\text{GRM}} = P^{-1}\lambda$. This implies the relation between λ_{GRM} and h^2_{SNP} is given by $\lambda_{\text{GRM}} = (h^2_{\text{SNP}})^{-1} - 1$.

8. Advanced Ridge Regression Methods

8.1. Heteroskedastic Ridge Regression. A point of critique regarding the use of RR is the lack of SNP selection. However, for highly polygenic traits, given current sample sizes, there is evidence that SNP selection is sometimes detrimental to predictive accuracy (e.g., [3, 8, 21]). Nevertheless, since RR can be used for inference just as well as RSR, the approach of selecting SNPs that attain a p value below some threshold τ in the GWAS can also be extended to RR.

In a spirit similar to that of SNP selection, one can argue in favor of a heteroskedastic ridge regression (HRR), where each SNP receives a different amount of shrinkage [50, 51]. As with homoskedastic shrinkage, this SNP-specific shrinkage might be based on either results from the training set or prior information from different data sets. Depending on the size of SNP-specific shrinkage, this method can leverage between SNP selection and full inclusion. Based on prior evidence or in-sample evidence the weight assigned to a SNP can be made arbitrarily small or arbitrarily large given the amount of evidence for association with the outcome. SNP-specific shrinkage opens up the door for a whole array of HRR methods (e.g., [50, 51]).

The HRR estimator in (5) and resulting predictions can be rewritten as

$$\widehat{\boldsymbol{\beta}}_{\text{HRR}} = \boldsymbol{\Lambda}^{-1}\mathbf{X}^\top\left(\mathbf{X}\boldsymbol{\Lambda}^{-1}\mathbf{X}^\top + \lambda\mathbf{I}\right)^{-1}\mathbf{y}, \tag{24}$$

$$\widehat{\mathbf{y}}_2 = \mathbf{X}_2\boldsymbol{\Lambda}^{-1}\mathbf{X}^\top\left(\mathbf{X}\boldsymbol{\Lambda}^{-1}\mathbf{X}^\top + \lambda\mathbf{I}\right)^{-1}\mathbf{y}, \tag{25}$$

where $\boldsymbol{\Lambda} = \text{diag}(\{\lambda_p\}_{p=1,\ldots,P})$ is a diagonal matrix with SNP-specific shrinkage effects.

It is implied by (24) and (25) that HRR can be carried out using the same machinery as homoskedastic RR, by first weighting the SNPs appropriately. More specifically, take $\mathbf{X}^* = \mathbf{X}\boldsymbol{\Lambda}^{1/2}$ and $\mathbf{X}^*_2 = \mathbf{X}_2\boldsymbol{\Lambda}^{1/2}$ and construct corresponding weighted GRMs by taking

$$\begin{aligned}\mathbf{A}^* &= \mathbf{M}_{\mathbf{Z}}\left(\frac{1}{P}\mathbf{X}^*\mathbf{X}^{*\top}\right)\mathbf{M}_{\mathbf{Z}},\\ \mathbf{A}^*_{21} &= \mathbf{M}_{\mathbf{Z}_2}\left(\frac{1}{P}\mathbf{X}^*_2\mathbf{X}^{*\top}\right)\mathbf{M}_{\mathbf{Z}}.\end{aligned} \tag{26}$$

Now, using the eigendecomposition defined in (19) of the weighted GRM defined in (26) and by subsequently applying (21) to resulting eigenvectors in $\mathbf{Q}$ and eigenvalues, $\{\theta_i\}_{i=1,\ldots,N}$, we obtain efficient out-of-sample HRR predictions.

8.2. Incorporating Information from Earlier Studies. Using HRR prediction it is possible to include results from a GWAS in other samples as prior information. Consider SNP-specific shrinkage, given by $\lambda_p = \sigma^2_{\boldsymbol{\varepsilon}}/\sigma^2_{\beta_p}$, and a set of GWAS t-test statistics from another study without the presence of confounding variables. Given that $\widehat{\sigma}_{\boldsymbol{\varepsilon}}$ is approximately constant over the SNPs in the GWAS, the t-test statistic of SNP p can be written as

$$t_p \approx \frac{1}{\widehat{\sigma}_{\boldsymbol{\varepsilon}}}\left(\frac{\mathbf{x}_p^\top}{\sqrt{\mathbf{x}_p^\top\mathbf{x}_p}}\right)\mathbf{y} = \frac{1}{\widehat{\sigma}_{\boldsymbol{\varepsilon}}}\mathbf{x}_p^{*\top}\mathbf{y} = \frac{1}{\widehat{\sigma}_{\boldsymbol{\varepsilon}}}\widehat{\beta}_p, \tag{27}$$

where $\mathbf{x}_p^*$ denotes SNP p standardized to unit length and $\widehat{\beta}_p$ the estimated effect of the standardized SNP. It follows from this equation that these statistics are proportional to the estimated effects of standardized SNPs. Therefore, the square t-test statistics are approximately proportional to the square standardized GWAS estimates. Now, under the prior probability distribution that $\beta_p \sim \mathcal{N}(0, \sigma^2_{\beta_p})$ we have that $\widehat{\beta}^2_p$ is a consistent estimator of $\sigma^2_{\beta_p}$. Correspondingly, the square t-test statistics are proportional to this estimator of the SNP-specific effect variance. Therefore, for a suitable choice of λ a consistent estimator of λ_p is given by $\lambda t_p^{-2} = \sigma^2_{\boldsymbol{\varepsilon}}/\widehat{\beta}^2_p$. In the framework of HRR, this entails setting $\widehat{\boldsymbol{\Lambda}} = \mathrm{diag}(\{t_p^{-2}\}_{p=1,\ldots,P})$. This definition of $\widehat{\boldsymbol{\Lambda}}$ implies that SNPs are weighted according to t_p. From these weighted SNPs we can construct the weighted GRM and apply (25) to obtain out-of-sample HRR predictions which incorporate information from a GWAS in another dataset.

8.3. Nonlinear Prediction Methods. An important question in genetics is how nonlinear effects (e.g., *dominance* and *epistasis*) contribute to the variation of complex traits. RR can efficiently implement such nonlinear SNP effects using the *kernel trick* from machine learning. Resulting *kernel ridge regression* (KRR) extends the nonparametric approach to RR, where genetic "similarities" in the space of additive effects are replaced by genetic "similarities" in a larger (potentially infinite) feature space, for instance, including two- or three-way interactions.

The efficient RR predictions in (17) are in essence a weighted average of the observed phenotypes in the training set. Weights are based on the genetic similarity of individuals in the test set and the training set. The more genetically similar two individuals are in the test and training set, the more weight will be given to the phenotype of the similar individual in the training set.

Classical RR measures genetic similarity of individuals in the space of additive effects and assigns weights accordingly. KRR, however, can measure genetic similarity in the space of more than just additive effects. This extended space can include, for instance, d-way interactions between SNPs. Now, a GWAS estimating all potential d-way interactions between SNPs is not feasible. However, with KRR, rather than having to estimate all coefficients of all nonlinear combinations of regressors, one can instead obtain the measure of genetic similarity in this higher-dimensional space by applying a simple kernel function $k(\mathbf{x}_i, \mathbf{x}_j)$ to any two genotype vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ corresponding to individuals i and j.

In this context, classical RR corresponds to $k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^\top\mathbf{x}_j$. Similarly, a function measuring similarity in the space consisting only of two-way linear interactions is given by

$$k\left(\mathbf{x}_i, \mathbf{x}_j\right) = \left(\mathbf{x}_i^\top\mathbf{x}_j\right)^2. \tag{28}$$

To see why this is so, consider expanding (28). We then have

$$\begin{aligned} k\left(\mathbf{x}_i, \mathbf{x}_j\right) &= \left(\sum_{p=1}^{P} x_{ip}x_{jp}\right)^2 = \sum_{p=1}^{P}\sum_{q=1}^{P} x_{ip}x_{jp}x_{iq}x_{jq} \\ &= \sum_{p=1}^{P}\sum_{q=1}^{P}\left(x_{ip}x_{iq}\right)\left(x_{jp}x_{jq}\right) = \boldsymbol{\phi}(\mathbf{x}_i)^\top\boldsymbol{\phi}\left(\mathbf{x}_j\right), \end{aligned} \tag{29}$$

where $\boldsymbol{\phi}(\mathbf{x}_i)^\top = (\{\{x_{ip}x_{iq}\}_{q=1,\ldots,P}\}_{p=1,\ldots,P})$. Thus, $\boldsymbol{\phi}(\mathbf{x}_i)$ is a vector that contains all possible two-way interactions between the P markers. Kernel function $k(\mathbf{x}_i, \mathbf{x}_j)$ represents the genetic similarity of individuals i and j in this space of all two-way interactions between SNPs.

The essence of KRR is the so-called kernel trick that allows one to efficiently compute the higher-dimensional similarity measure by applying a simple kernel function $k(\mathbf{x}_i, \mathbf{x}_j)$ to any two input vectors for individuals i and j [52]. Provided the kernel is positive definite it constitutes the reproducing kernel of a unique *reproducing kernel Hilbert space* (RKHS) [53]. KRR then is equivalent to a so-called RKHS regression.

In the case of d-way interactions the associated kernel function $k(\mathbf{x}_i, \mathbf{x}_j)$ can be evaluated for all pairs of individuals by raising each element of the GRM, $P^{-1}\mathbf{X}\mathbf{X}^\top$, to the power d. An alternative is the nonhomogeneous polynomial kernel of degree d, given by $k(\mathbf{x}_i, \mathbf{x}_j) = (c + \mathbf{x}_i^\top\mathbf{x}_j)^d$. This kernel, similar to the regular polynomial kernel of degree d, includes d-way interactions but also lower-order interaction terms including the "one-way interactions," that is, simple additive linear effects.

The preceding example of the polynomial kernel of degree two shows how KRR can include dominance and epistasis in the prediction model. For frequently used kernels, such as the Gaussian (radial basis function) kernel, there exists a representation in which classical RR is applied to a model with infinitely many predictors, nevertheless yielding finite predictions. Obtaining the weights for infinitely many predictors is not possible. Hence, rather than aiming to obtain point estimates of $\boldsymbol{\beta}$, KRR only aims to obtain predictions.

BLUP and, by extension, RR are special cases of prediction using KRR (e.g., [54, 55]). There has been a substantial amount of work in plant and animal breeding, aiming to

improve predictive accuracy using KRR (e.g., [12, 15, 16]). A generally used kernel is the aforementioned Gaussian kernel, defined as

$$k\left(\mathbf{x}_i, \mathbf{x}_j\right) = \exp\left[-\frac{d^2\left(\mathbf{x}_i, \mathbf{x}_j\right)}{\eta}\right], \tag{30}$$

where $d^2(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^\top(\mathbf{x}_i - \mathbf{x}_j)$ and hyperparameter $\eta > 0$. This type of kernel includes all conceivable linear interactions between the P SNPs and with themselves. Endelman [16] finds that the Gaussian kernel outperforms accuracy of RR and a Bayesian approach to LASSO, used to predict wheat and maize traits in samples, typically with about 300 observations and 3000 SNPs. Similarly, using a Bayesian approach, Crossa et al. [15] find in samples of about 250 observations, with 1100 SNPs, that both the Gaussian kernel and the LASSO outperform predictive accuracy of RR for grain yield and maize flowering traits. However, when comparing the LASSO with the Gaussian KRR, which of two the methods is better, depends on the trait. An efficient implementation of KRR based on maximum likelihood, using the Gaussian kernel, is available in the R package `rrBLUP` [16].

Morota and Gianola [17] compare a wide range of kernels, such as the exponential [12, 16, 56], Matérn, diffusion (e.g., [57]), and t kernel [58], for the purpose of obtaining genomic estimated breeding values [38]. Though it is argued that selecting a suitable kernel is the most precarious step (e.g., [14]), current evidence suggests that most considered kernels attain a predictive accuracy similar to that of the Gaussian kernels [17]. Thus, it appears that the Gaussian KRR is a robust prediction method for quantitative traits, able to handle nonlinear genetic architectures. Moreover, Endelman [16] finds little evidence supporting the hypothesis that a Gaussian kernel is likely to overfit the data [56].

Given the current evidence, KRR using an appropriate kernel (e.g., the Gaussian kernel) is a promising prediction technique, especially for traits where epistatic effects and dominance are expected to contribute to trait variation. De los Campos et al. [14] suggest an interesting venue for further research on the use of KRR for prediction in quantitative genetics, by combining multiple kernels in a single model, each kernel representing a single variance component (e.g., additive, dominance, or epistasis). For a more detailed treatment of KRR and its uses in quantitative genetics, see Morota and Gianola [17].

Regarding the computation of predictions using KRR, let $\mathbf{K}$ denote the matrix of similarities in the higher-dimensional feature space in the training set, such that an element of this matrix k_{ij} is given by $k(\mathbf{x}_i, \mathbf{x}_j)$ and let $\mathbf{K}_{21}$ be defined similarly for individuals in the test set versus individuals in the training set. Now, KRR prediction without confounders is given by $\widehat{\mathbf{y}}_2 = \mathbf{K}_{21}(\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y}$ and with confounders by

$$\widehat{\mathbf{y}}_2 = \mathbf{M}_{\mathbf{Z}_2}\mathbf{K}_{21}\mathbf{M}_{\mathbf{Z}}\left(\mathbf{M}_{\mathbf{Z}}\mathbf{K}\mathbf{M}_{\mathbf{Z}} + \lambda\mathbf{I}\right)^{-1}\mathbf{y}, \tag{31}$$

where, as before, $\mathbf{M}_{\mathbf{C}}$ is the projection matrix removing the effects of $\mathbf{C}$.

In the case of the nonhomogeneous polynomial kernel of degree d, given the GRMs, $P^{-1}\mathbf{X}\mathbf{X}^\top$ and $P^{-1}\mathbf{X}_2\mathbf{X}^\top$, the matrices $\mathbf{K}$ and $\mathbf{K}_{21}$ can be obtained efficiently by adding a constant c to each element of the GRMs and by raising each resulting element to the power d. When $c > 0$ and $d \in \{1, 2, \ldots\}$ are not fixed, these are additional hyperparameters which can be tuned via (N)CV.

9. Simulation Study

An important question is under what circumstances can we expect RR to yield more accurate predictions than RSR? The answer to this question can help us assess the merits of RR in quantitative genetics. As discussed, prediction using RR is intimately related to the BLUP of the phenotype under a mixed linear model in which SNP effects are assumed to be all drawn from a normal distribution. This corresponds to idea of each SNP making a tiny contribution to phenotype. Therefore, it is reasonable to assume that RR will perform well when the SNP effects are as such. However, given that not all SNPs in existence are causally affecting the outcome, an open question is how does RR perform when only a subset of SNPs affects the outcome?

Moreover, an important factor influencing predictive accuracy of a classical polygenic score is the training sample size. Therefore, RR is likely also to be very sensitive to the sample size. Finally, the more heritable a trait is, the easier it should be to detect the effects of SNPs. Thus, an additional question is how do RR and RSR perform under different levels of heritability?

In short, we want to know the relative predictive accuracy of RR and RSR (i) for a wide range of trait architectures and (ii) under particular combinations of sample size and the number of genotyped SNPs. To answer this question we run a suite of simulations. In these analyses, we vary sample size of the training set (N), the number of genotyped SNPs (P), the fraction of SNPs exerting a causal influence (f_C), and the SNP-based heritability (h^2_{SNP}).

Table 1 shows the levels we consider for these factors. In addition, a range of values for λ on the interval $[10^{-6}; 10^{9}]$ is considered. Each unique combination of levels of these factors constitutes a scenario. The total number of scenarios is $S = 7 \times 12 \times 37 \times 20 = 62{,}160$. We consider $R = 21$ runs, yielding $S \times R = 1{,}305{,}360$ combinations of levels and runs.

For a combination of sample size, the number of SNPs, trait heritability, and a fraction of causal SNPs chosen from the levels shown in Table 1, let C be the corresponding number of causal SNPs. Now, the data generating process for this combination of levels is given by

$$\begin{aligned}
y_i &= \sum_{p=1}^{C} x_{ip}\beta_p + \varepsilon_i, \quad \text{for } i = 1, \ldots, N_{\text{total}},\\
x_{ip} &= \frac{g_{ip} - 2f_p}{\sqrt{2f_p\left(1 - f_p\right)}}, \quad \text{for } i = 1, \ldots, N_{\text{total}},\ p = 1, \ldots, P,\\
g_{ip} &\sim \text{Binom}\left(2, f_p\right), \quad \text{for } i = 1, \ldots, N_{\text{total}},\ p = 1, \ldots, P,\\
f_p &\sim \mathcal{U}\left(0.05, 0.95\right), \quad \text{for } p = 1, \ldots, P,\\
\beta_p &\sim \mathcal{N}\left(0, \sigma^2_{\boldsymbol{\beta}}\right), \quad \text{for } p = 1, \ldots, P,\\
\varepsilon_i &\sim \mathcal{N}\left(0, \sigma^2_{\boldsymbol{\varepsilon}}\right), \quad \text{for } i = 1, \ldots, N_{\text{total}},
\end{aligned} \tag{32}$$

Table 1: Factors and levels of the simulation study. h^2_{SNP}: heritability simulated phenotype, N: sample size training set, P: number of SNPs, and f_C: fraction of SNPs causal.

Factors	Levels	Number of levels
N	$\{200; 500; 1,000; \ldots; 10,000; 20,000\}$	7
P	$\{100; 200; 500; \ldots; 100,000; 200,000; 500,000\}$	12
f_C (%)	$\{0.1; \ldots, 100\}$ (Linear increases on logarithmic scale)	37
h^2_{SNP} (%)	$\{5; 10; 15; \ldots; 100\}$	20

where $\text{Binom}(a, b)$ denotes the binomial distribution with a draws each with probability of success b and $\mathcal{U}(a, b)$ denotes the uniform distribution on the interval (a, b). This data generating process corresponds to a quantitative trait which is normally distributed and has only additive genetic variation to which common variants contribute (i.e., minor allele frequency above 5%).

The total number of observations N_{total} includes the individuals in the test set. The size of the test set is 10% of the size of the training set, hence, yielding $N_{\text{total}} = \lfloor 1.1N \rfloor$. Here, $\lfloor x \rfloor$ denotes the nearest smaller integer.

In order not to be dependent on a single generated dataset, the entire simulation consists of $R = 21$ independent runs (replications). In each run we simulate only one set of genotype data for $N_{\max} = 22{,}000$ individuals and $P_{\max} = 500{,}000$ SNPs. Given any combination of N and P listed in Table 1 we can take an appropriate submatrix of the genotype matrix. To this submatrix we apply a set of P weights of which some are zero, such that we attain the desired fraction of SNPs being causal. Moreover, by scaling these weights and the noise vector $\boldsymbol{\varepsilon}$ appropriately we can attain any specified heritability. The result is a four-dimensional phenotype array with individuals along the first dimension and the factors P, f_C, and h^2 along the remaining dimensions.

When computing the out-of-sample predictions based on RR and RSR, the available genotype matrix only depends on N and P, not on h^2 or on f_C. Therefore, given N and P, when $N \leq P$ the eigendecomposition of the $N \times N$ GRM, $P^{-1}\mathbf{X}\mathbf{X}^\top$, can be reused for all combinations of h^2 and f_C. Moreover, the approach has already been amended to reuse the eigendecomposition for different values of λ. Similarly, when $N > P$ the eigendecomposition of $P \times P$ matrix $P^{-1}\mathbf{X}^\top\mathbf{X}$ can be reused. Since there only are 7 unique levels of N and 12 unique levels of P, RR prediction (i) for the 62,160 scenarios per replication reduces to computing $7 \times 12 = 84$ eigendecompositions and (ii) for each scenario to carrying out the matrix multiplications seen in (21).

In a typical run it takes 4.5 hours to predict using RR on a machine with 16 cores at 2.60 GHz per core with 64 GB RAM. The RSR predictions are generated alongside at virtually no costs in terms of CPU and memory. The computing time includes computation of the GRM, $P^{-1}\mathbf{X}\mathbf{X}^\top$, when $N \leq P$ and $P^{-1}\mathbf{X}^\top\mathbf{X}$ when $N > P$. Given N and P, failure to exploit (i) the constancy of the GRM and of $P^{-1}\mathbf{X}^\top\mathbf{X}$ over the 20×37 different combinations of h^2 and f_C and (ii) the properties of the eigendecomposition which enable the joint evaluation of the 151 values of λ we consider dramatically increases the CPU time of RR. In fact, we infer that the less efficient approach yields a CPU time that is at most a factor $20 \times 37 \times 151 = 111{,}740$ larger than the 4.5 hours we attain (i.e., about 57 years per run). Even worse, when the naive RR approach is applied and also when $P \gg N$, RR predictions cannot be obtained for datasets with more than 50,000 SNPs on the machine we use. Thus, using the efficient approach based on the GRM when $N \leq P$ and based on $P^{-1}\mathbf{X}^\top\mathbf{X}$ when $N > P$, combined with the smart use of eigendecompositions and constancy of GRMs over different combinations of f_C and h^2 we are able to reduce CPU times from several decades to several hours.

In each run, for each combination of levels we compute the R^2 of the RSR prediction with the outcome and the R^2 of the RR prediction with the outcome. R^2 is measured by the squared sample correlation coefficient between the polygenic score and the outcome in the test set. Our aim is to assess predictive accuracy of RSR and see whether it differs significantly from zero for a wide range of configurations. Moreover, we want to test whether RR provides a significant improvement compared to RSR. Therefore, the performance of RR is measured relative to RSR. That is, we take the log-ratio of the two, given by $\log(R^2_{\text{RR}}/R^2_{\text{RSR}})$. This measure is continuously distributed over $(-\infty, +\infty)$.

We measure the absolute performance of RSR by the logit transformation of $R^2_{\text{RSR}}/h^2_{\text{SNP}}$; that is,

$$\text{logit}\left(\frac{R^2_{\text{RSR}}}{h^2_{\text{SN}}}\right) = \log\left(\frac{R^2_{\text{RSR}}/h^2_{\text{SNP}}}{1 - R^2_{\text{RSR}}/h^2_{\text{SNP}}}\right). \tag{33}$$

This measure is also distributed over $(-\infty, +\infty)$. The reason for dividing R^2_{RSR} by h^2_{SNP} is that we want to know what part of the genetic variation the polygenic score captures. If h^2_{SNP} is low, for instance, 5%, we consider a polygenic score that attains an R^2 of 4% to be more impressive than a risk score that explains 10% of the variation in a highly heritable trait (e.g., $h^2_{\text{SNP}} = 50\%$). Note that we exclude observations with $R^2_{\text{RSR}} > h^2_{\text{SNP}}$ as these are uninformative outliers; a polygenic score that "explains" more genetic variation than there actually is is simply wrong.

Regarding the RR penalty, let $R^2_{\text{RR}}(\lambda, r)$ denote the accuracy of RR in run r, given penalty λ, conditional on some N, P, f_C, and h^2. Now, let

$$R^2_{\text{RR,med}}(\lambda) = \text{median}\left(\left\{R^2_{\text{RR}}(\lambda, r)\right\}_{r=1,\ldots,R}\right) \tag{34}$$

denote the median of the RR performance over the runs for a specific value of λ. Now, for this combination of N, P, f_C, and h^2 we take

$$\widehat{\lambda} = \underset{\lambda \in \{\lambda_1, \ldots, \lambda_L\}}{\arg\max}\, R^2_{\text{RR,med}}(\lambda). \tag{35}$$

Thus, for a given combination of levels of factors λ is tuned by setting it such that it maximizes the median R^2 of RR over the runs for the given combination of levels. Based on this

Table 2: Summary statistics of $\log(R^2_{RR}/R^2_{RSR})$ for the full set of observed log-ratios and for the subset excluding log-ratios outside $(-1, +1)$, and R^2_{RSR} for the full set and for the subset excluding observations for which $R^2_{RSR} \geq h^2_{SNP}$. Results stem from all combinations of the levels of the factors in the simulation design, with 21 replications per combination.

Outcome	Restriction	Count	(% total)	Mean	Var.	Min	Max
$\log(R^2_{RR}/R^2_{RSR})$	None	1,305,360	(100.0%)	0.065	0.403	−22.2	20.7
$\log(R^2_{RR}/R^2_{RSR})$	$\in (-1, +1)$	1,254,168	(96.1%)	0.060	0.041	−1.00	1.00
R^2_{RSR}	None	1,305,360	(100.0%)	0.177	0.058	0.000	0.997
R^2_{RSR}	$<h^2_{SNP}$	1,239,721	(95.0%)	0.160	0.051	0.000	0.997

procedure, the optimal R^2 of RR in run r is given $R^2_{RR}(\widehat{\lambda}, r)$. This yields a single measure of accuracy of RR per replication and per combination of levels. This procedure results in a value of λ that performs well in 21 independent samples. Hence, it is similar to a value that would result from CV; there is little scope for overfitting. Moreover, since the median is less sensitive to outliers than, for instance, the mean, we make our measure more robust by taking the median over the runs. The reason that we choose for this approach instead of CV is to reduce the computational complexity of the simulation procedure at the expense of having a slightly less elegant approach.

9.1. Simulation Results. Table 2 shows the summary statistics of the measure $\log(R^2_{RR}/R^2_{RSR})$ and of R^2_{RSR}. As can be seen, overall the combinations of levels and runs RR seems to outperform RSR on average by about 6%. However, there is much variation in the log-ratio. The lowest log-ratio is −22.2 and the highest is +20.7. Since this ratio is on a log scale this implies a tremendous difference in R^2. The reason for this is that when either the nominator or the denominator of R^2_{RR}/R^2_{RSR} gets close to zero, the log-ratio can attain a large value (in absolute terms). For this reason we excluded log-ratios outside the interval $(-1, +1)$. This leads to a drop in the variance from about 0.4 to 0.04, only at the expense of losing 3.9% of the observed combinations of levels and runs. Moreover, the mean log-ratio hardly changes by removing the outliers. This reduction in variance allows us to display the results in a more insightful manner and ensures further inferences on the relation between our factors (e.g., sample size) and predictive accuracy are not influenced by aberrant observations. For R^2_{RSR} we see that the average R^2 of about 17% is significantly greater than zero.

Figure 3 shows the histogram of $\log(R^2_{RR}/R^2_{RSR})$ over the combinations of runs and levels inside the range $(-1, +1)$. This histogram confirms that there are long and thin tails. Most mass centers around zero. However, the empirical distribution is slightly skewed to the right, giving rise to the positive average log-ratio. The figure shows that RR often performs better than RSR. Given the fact that RR lies between RSR and OLS, this is not surprising. Using the penalty parameter λ, RR tries to find the optimum between these two extremes. Figure 4 shows the histogram of $\text{logit}(R^2_{RR}/h^2_{SNP})$ excluding observations for which $R^2_{RR} > h^2_{SNP}$. The observations are smoothly distributed. A value of zero corresponds to an R^2 equal to half the heritability. Thus, in a substantial

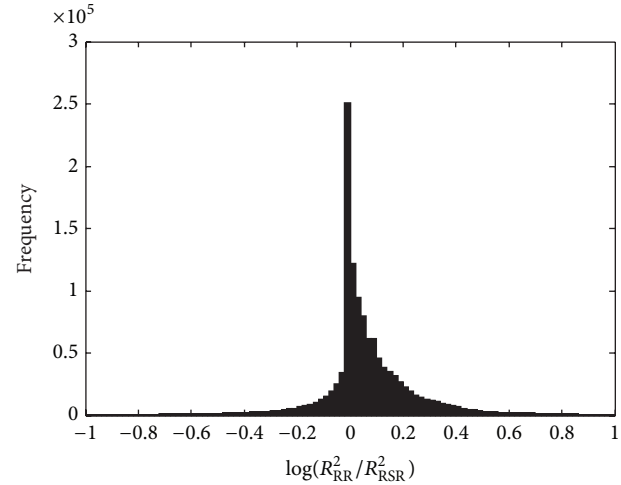


Figure 3: Histogram of $\log(R^2_{RR}/R^2_{RSR})$ for 21 runs of simulated data, for different values of N, P, f_C, and h^2. Ridge parameter λ is chosen to maximize median R^2_{RR}. Values outside $(-1, +1)$ are excluded.

proportion of the cases RSR captures more than half of the genetic variation.

Figure 5 shows the log-ratio of the median R^2 of ridge regression and of RSR, with values outside the interval $(-1, +1)$ truncated to corresponding extremes of this interval. This truncation is necessary in order for the figure not to be dominated by the outliers. For $N \ll P$ (see the lower right block in Figure 5), the performance of RR and RSR is volatile. Sometimes, RR strongly outperforms RSR and sometimes it is the other way round. However, on average RR seems to outperform RSR. As N approaches P (see the lower left and upper right blocks in Figure 5) RR starts to outperform RSR. There are large regions, where the log of the gain in accuracy is consistently between zero and a half. This corresponds to a relative increase between zero and 65%. For example, for $N = P = 20{,}000$, $h^2_{SNP} = 50\%$, and 200 causal SNPs RSR attains a median R^2 of 17% and RR 20%, constituting a relative increase of 16%. This gain in accuracy peaks when $N \approx P$.

When $N \gg P$ (see the upper left block in Figure 5), the gain in accuracy drops to zero. However, it is unlikely that this pattern, where the gain of RR dies out as N keeps increasing, replicates empirically. The reason for this is that the pattern is probably an artefact of the design of the simulation; all SNPs are simulated independent of each other. Even though

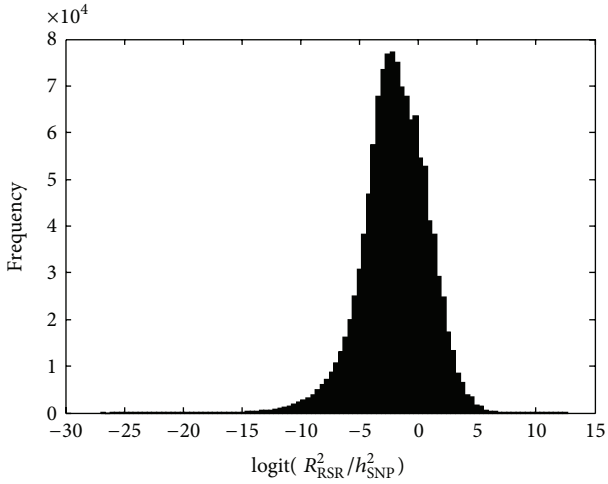


Figure 4: Histogram of logit(R^2_{RSR}/h^2_{SNP}) for 21 runs of simulated data, for different values of N, P, f_C, and h^2, excluding observations for which $R^2_{RSR} \geq h^2$.

Table 3: The observed median of the R^2 of RSR and of RR relative to RSR, over 21 simulations, for different sample sizes (N), number of SNPs (P) of which 1% causal, and heritabilities (h^2_{SNP}). Ridge parameter λ is chosen to maximize the median R^2_{RR}.

N	P	h^2_{SNP}	Median R^2_{RSR}	Median R^2_{RR}/ median R^2_{RSR}
5 k	500 k	0.50	0.003	1.078
10 k	500 k	0.50	0.005	1.029
20 k	500 k	0.50	0.009	1.038
10 k	100 k	0.50	0.027	1.079
10 k	200 k	0.50	0.011	1.000
10 k	500 k	0.50	0.005	1.029
10 k	500 k	0.25	0.001	1.000
10 k	500 k	0.50	0.005	1.029
10 k	500 k	0.75	0.011	1.011

empirical correlations between SNPs can arise in the simulations, asymptotically there is none. Thus, for sufficiently large N (compared to P) the standardized simulated SNP data are such that $\mathbf{X}^\top\mathbf{X}$ approaches the identity matrix and RR becomes equivalent to RSR (see Section 3). Therefore, the accuracy of RR and RSR does not differ for such extremely large values of N. How the performance differs in these large samples when there is linkage disequilibrium in the data remains to be seen.

Table 3 shows the median of the R^2 of RSR and that of RR relative to RSR for combinations of sample size and the number of genotyped SNPs that are typically seen in a GWAS (e.g., N = 10,000, P = 500,000). We see that for these data dimensions a trait with a heritability of 50% has a classical polygenic score which on average only explains 0.5% of the total phenotypic variation. Moreover, RR yields a relative increase of just 2.9%. This increase gives an absolute R^2 of 0.51% for RR. This observation clearly illustrates that the so-called missing heritability [45] is hard to find, even under a very simple data generating process, that is, a process for which we are sure that both RSR and RR should asymptotically capture all genetic variation.

Table 4: Regressors used to explain the predictive accuracy of RR and RSR.

Regressor	Captures
Intercept	Level
$\log(N)$	Effect sample size
$\log(P)$	Effect number of SNPs
$\log(C)$	Effect of number of causal SNPs (C)
$\log(f_C)$	Effect of fraction of SNPs causal
$\log(h^2)$	Effect of heritability

9.2. Modelling the Simulation Results. To understand the relation between the various factors in the simulation study and the gain in predictive accuracy by RR we fit a linear model to the logarithm of the ratio R^2_{RR}/R^2_{RSR} for all replications and for all considered levels of factors, such as sample size. Moreover, in order to obtain the R^2 of RSR as a benchmark we also fit a linear model, the logit transformation of R^2_{RSR}.

The results in the previous section indicate that the relation between sample size N and the performance is nonlinear. The relation seems to exhibit an inverted U-shape. For this purpose, we include $\log(N)$ and its square as regressors. Moreover, the location of the peak depends on the number of SNPs, implying that the parameters of regressors related to sample size depend on P. Consequently, interactions between P and N are added to the model. By symmetry of Figure 5, similar arguments hold for the performance as function of P. Based on this argument we consider up to three-way interactions between the regressors.

In addition, we see in many subplots of Figure 5 that the gain in predictive accuracy differs systematically between low, intermediate, and high heritabilities. Therefore, heritability is included as regressor. Finally, although the effect of the fraction of causal SNPs is hard to judge from Figure 5, we include this factor as regressor as well.

Both outcomes are modelled as a linear function of the aforementioned basic regressors. These regressors are reported in Table 4. We consider models ranging from merely an intercept up to all 3-way interactions between the explanatory variables. We choose the model that minimizes the Bayes information criterion (BIC) [59].

Table 5 reports the BIC values of the respective models. On the basis of these values we find that a model including all three-way interactions is most appropriate, both in case of the log-ratio and in case of the logit of the performance of RSR relative to the heritability. The model for the gain in accuracy of RR relative to RSR can explain approximately 12% of the variation in this measure on the basis of sample size and the other regressors. The model for the accuracy of RSR can explain about 61%.

A likely reason for the fact that we can explain far more variation in the R^2 of RSR than in the gain of RR relative to RSR is the following. In case both the R^2 of RR and RSR are to

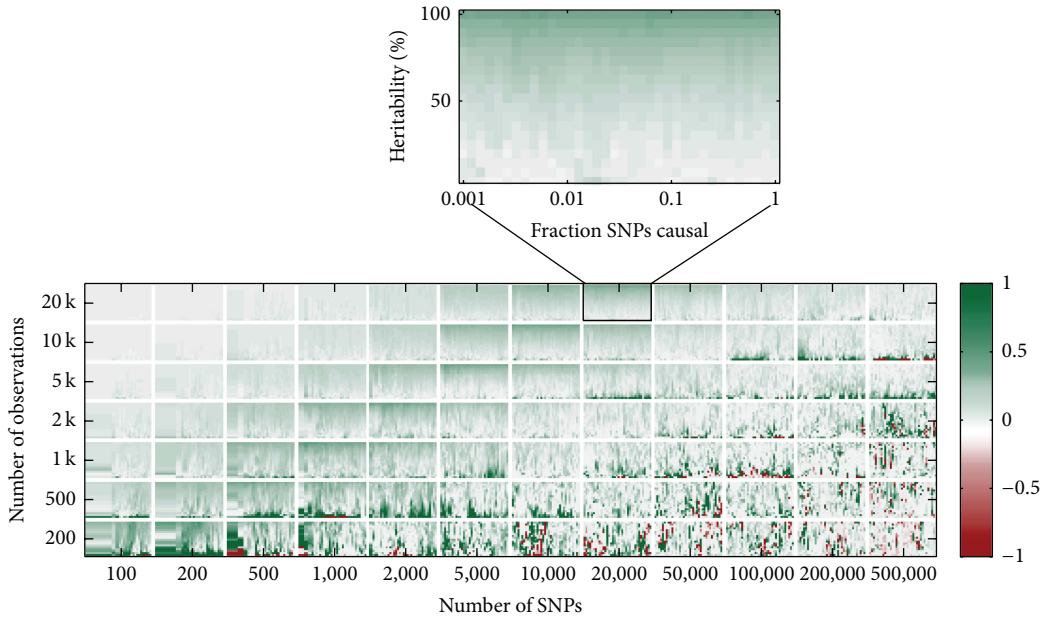


Figure 5: The logarithm of the ratio of the median of the R^2 over 21 simulations, attained by ridge regression and the classical GWAS approach of RSR, for different combinations of training sample size (y-axis), number of SNPs (x-axis), trait heritability (y-axis in the sub-plots), and fraction of SNPs with a causal effect (x-axis in the subplots). Ridge parameter λ is chosen to maximize the median R^2_{RR}. Values truncated to lie between minus one (red) and plus one (green).

Table 5: Bayes information criterion (BIC) and the proportion of explained variance (R^2_{model}) of the model for the gain in predictive accuracy of RR relative to RSR and the model for the performance of RSR, over different combinations of levels of the factors sample size, number of SNPs, fraction of causal SNPs, and heritability. Lowest BIC printed bold.

Outcome	Regressors (number of regressors)	Number of observations	R^2_{model}	BIC
$\log(R^2_{RR}/R^2_{RSR})$	Intercept (1)	1,254,168	0.0%	$-3.998 \cdot 10^6$
$\log(R^2_{RR}/R^2_{RSR})$	Regressors in Table 4 (5)	1,254,168	4.7%	$-4.058 \cdot 10^6$
$\log(R^2_{RR}/R^2_{RSR})$	& 2-way interactions (15)	1,254,168	8.4%	$-4.107 \cdot 10^6$
$\log(R^2_{RR}/R^2_{RSR})$	& 3-way interactions (35)	1,254,168	12.4%	$\mathbf{-4.163 \cdot 10^6}$
$\text{logit}(R^2_{RSR}/h^2_{SNP})$	Intercept (1)	1,239,721	0.0%	$2.542 \cdot 10^6$
$\text{logit}(R^2_{RSR}/h^2_{SNP})$	Regressors in Table 4 (5)	1,239,721	48.6%	$1.717 \cdot 10^6$
$\text{logit}(R^2_{RSR}/h^2_{SNP})$	& 2-way interactions (15)	1,239,721	56.3%	$1.515 \cdot 10^6$
$\text{logit}(R^2_{RSR}/h^2_{SNP})$	& 3-way interactions (35)	1,239,721	60.9%	$1.379 \cdot 10^6$

a large extent influenced by our factors in a similar way, taking the log-ratio basically eliminates these common effects. What then remains is a measure over which the factors have less predictive power than over the absolute R^2 measure.

Using the parameters estimates of the models we predict the log-ratio of R^2_{RR} and R^2_{RSR} as well as R^2_{RSR} for sample sizes between 100,000 and 500,000 individuals and the number of SNPs between 100,000 and 500,000. For heritability and the fraction of causal SNPs we use the ranges considered in the initial simulations. The resulting predictions of the gain in accuracy are displayed in the heatmap in Figure 6.

In addition, point estimates of R^2_{RR}/R^2_{RSR} and R^2_{RSR} are reported together with confidence intervals in Table 6. There are three groups of predictions. In the first group $P = 500{,}000$, $h^2 = 50\%$, and N varies from 100,000 to 500,000. In the second group $N = 500{,}000$ and P varies from 100,000 to 500,000. In the last group $P = N = 500{,}000$ and h^2 ranges from 25 to 75%.

Results from Figure 6 and Table 6 indicate that in most cases RR is expected to yield a relative increase in R^2 between 10% and 20% for sample sizes ranging between 100,000 and 500,000 individuals. All increases in accuracy are greater than zero at a 5% significance level. Moreover, RSR attains values of R^2 ranging between 15% and 75%. As an example, in case of 200,000 individuals and 500,000 SNPs, for a trait with $h^2_{SNP} = 50\%$ the R^2 of RSR is expected to be 33.7% and the R^2 of RR 37.3%.

Regarding these findings, combining the R^2 attained by RSR with the relative increase by RR yields expected values of the R^2 of RR which in some cases surpass h^2. In practice this cannot be true. In case a trait has an h^2 of 50% it is not possible to consistently predict more than 50% of the phenotypic

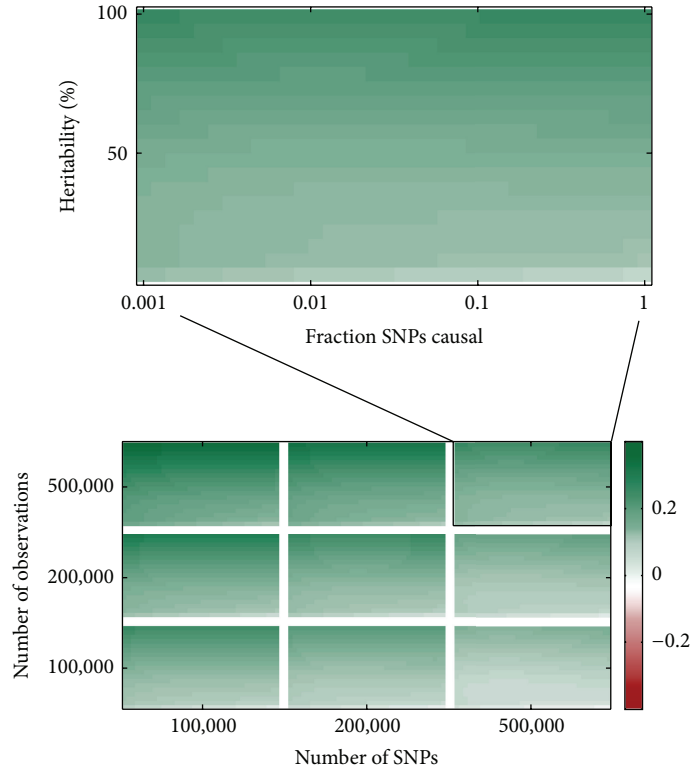


Figure 6: The predicted logarithm of the ratio of the R^2 attained by ridge regression and the classical GWAS approach of RSR, based on a model fitted to this measure in 21 runs of simulations. Predictions are shown for different combinations of training sample size (y-axis), number of SNPs (x-axis), trait heritability (y-axis in the subplots), and fraction of SNPs with a causal effect (x-axis in the subplots). Predicted values are not truncated. Deep green represents the highest prediction.

Table 6: Predictions of the predictive accuracy of RSR and of the gain in predictive accuracy of RR compared to RSR in large scale samples (e.g., $N \geq 100{,}000$) based on a linear model with sample size, number of SNPs, fraction of causal SNPs, and heritability as predictors. 95% confidence intervals (CI) are reported in parentheses, with the middle value indicating the point estimate. 1% of the SNPs are assumed to be causal.

N	P	h^2_{SNP}	95% CI R^2_{RSR}	95% CI $R^2_{\text{RR}}/R^2_{\text{RSR}}$
100 k	500 k	0.50	(0.139; 0.146; 0.153)	(1.062; 1.070; 1.079)
200 k	500 k	0.50	(0.324; 0.337; 0.349)	(1.094; 1.107; 1.121)
500 k	500 k	0.50	(0.473; 0.478; 0.482)	(1.142; 1.167; 1.193)
500 k	100 k	0.50	(0.486; 0.488; 0.490)	(1.218; 1.244; 1.270)
500 k	200 k	0.50	(0.482; 0.485; 0.488)	(1.193; 1.218; 1.244)
500 k	500 k	0.50	(0.473; 0.478; 0.482)	(1.142; 1.167; 1.193)
500 k	500 k	0.25	(0.205; 0.212; 0.218)	(1.110; 1.135; 1.160)
500 k	500 k	0.50	(0.473; 0.478; 0.482)	(1.142; 1.167; 1.193)
500 k	500 k	0.75	(0.733; 0.736; 0.739)	(1.191; 1.218; 1.245)

variation on the basis of SNP data. This seems to indicate that our estimates are somewhat optimistic. Nevertheless, for the ranges in which we actually simulated data (i.e., $N \leq 20{,}000$ and $P \leq 500{,}000$) RSR is able to attain a substantial R^2 when $N \approx P$ and RR is able to considerably increase the R^2. For instance, at $h^2 = 50\%$ and $N = P = 20{,}000$, with 200 causal SNPs the median R^2 of RSR is 17%, and the median R^2 of RR is 20%. This constitutes a relative increase in R^2 of about 16%. As shown in Figure 5, this pattern seems to persist while $N \approx P$. Hence, at the very least, the expectation that RR improves the R^2 of RSR considerably for large samples (e.g., $N \approx P \approx 500{,}000$) is not unreasonable.

10. Conclusions and Discussion

Ridge regression is a flexible technique that can be used to estimate the association between a set of P SNPs and an outcome observed for N individuals, even when $P \gg N$. When the ridge penalty is equal to the ratio of the noise variance and the variance of random SNP effects in a mixed linear

model, prediction using the weights from ridge regression is equivalent to the best linear unbiased prediction used in animal breeding, agricultural science, and more recently also human genetics.

Ridge regression can be perceived as a method that partially accounts for linkage disequilibrium between markers. For a sufficiently low penalty the method fully accounts for linkage disequilibrium and is therefore equivalent to the OLS estimator of the multiple regression problem using all SNPs jointly. On the other hand, for a sufficiently high penalty, in terms of predictions ridge regression ignores linkage disequilibrium and is therefore equivalent to the approach of a simple regression per SNP, which is common in a GWAS.

Using standard results from, for instance, machine learning and animal breeding, prediction using ridge regression can be shown to constitute solving an equation with N unknown weights and applying these weights to a measure of relatedness of individuals out of sample and in sample. Formulating ridge regression this way makes it a computationally efficient technique, even for a large number of SNPs.

As with multiple regression and GWAS predictions, ridge regression can account for the presence of confounding variables, such as age, gender, and population structure. Moreover, such corrections can again be implemented at low computational costs.

When the shrinkage parameter is unknown ridge prediction can be formulated such that predictions for different values of this parameter can be generated in a single step, requiring the eigendecomposition of an $N \times N$ matrix only once. This expression allows the researcher to efficiently carry out procedures, such as cross-validation, to tune this parameter.

Finally, ridge regression prediction is amenable to a wide array of advanced techniques. First, using the kernel trick from machine learning, nonlinear effects such as dominance and epistasis can easily be incorporated in the prediction model. Moreover, in a Bayesian spirit, results from earlier studies can be used to give a prior weight to SNPs in the ridge regression prediction. Similarly, when prior information is not available, in-sample information can be used to discount SNPs differently, yielding a heteroskedastic ridge regression prediction.

Empirical findings so far seem to suggest that for current sample sizes the performance of plain vanilla ridge regression is very similar to that of the repeated simple regression approach used in a GWAS. This raises two questions. First, how do more advanced ridge regression approaches perform? Second, how will the plain version of ridge regression perform in upcoming large scale initiatives, such as biobanks?

Using a suite of simulations we consider the second question. We confirm the finding that for most current studies, with sample sizes usually below 10,000 individuals and more than 500,000 SNPs, ridge regression hardly outperforms the classical GWAS approach. For a sample of 10,000 observations, with 500,000 SNPs of which 5,000 causal, for a trait with a heritability of 50%, the median R^2 in 21 independently simulated datasets is 0.5% for repeated simple regression and 0.51% for ridge regression. This resonates with the finding that the main determinant of predictive accuracy of the polygenic score is the sample size of the training set (e.g., [60, 61]). As long as $N \ll P$, there seems to be little advantage of advanced approaches, such as ridge regression, over the classical GWAS approach [61].

However, by analyzing the difference in accuracy of the classical approach and ridge regression for different values of N, P, trait heritability, and the fraction of causal variants, we are able to extrapolate the performance of ridge regression for large scale initiatives. For a sample size of 200,000 individuals and 500,000 SNPs, we find that in a trait with 50% heritability and with 5,000 causal variants the polygenic score of a GWAS is expected to explain 34% of the phenotypic variation, whereas ridge regression is expected to capture about 37%. Thus, in this scenario ridge regression is expected to capture about 75% of the genetic variation, whereas the classical approach captures 67%.

However, these predictions are rather coarse. They depend highly on the model being fitted (e.g., by including interactions between the number of individuals, SNPs, and heritability). This observation comes as no surprise; we extrapolate quite a bit outside the interior of the levels of the factors that were considered in the simulations (e.g., $N \leq 20{,}000$). However, one thing that remains unchanged even under different specifications of the models that try to explain the accuracy of respective methods is that ridge regression outperforms the repeated simple regression approach in all large scale samples considered.

A final note is concerned with the independence of the loci. In the present simulations at most 500,000 truly independent markers were used. As a result, all carry their own idiosyncratic bit of information about the genetic relationship of individuals in the data. As is shown, however, by Yang et al. [62], in real data with linkage disequilibrium taking a random subset of 60% or more of the SNPs from a grand set of 295 k SNPs yields heritability estimates of human height highly similar to estimates based on the full set; apparently adding more markers hardly changes the genetic relatedness estimates.

The findings of Yang et al. [62] illustrate that there might be a limited number of SNPs that can make a meaningful contribution to the SNP-based measure of genetic relationship. After this "effective number of SNPs" [63], new SNPs are primarily repeating the story that has been told by previous SNPs already. Therefore, even with many millions of SNPs (e.g., in imputed data), the resulting genetic relatedness estimates are highly similar to those obtained from a considerably smaller set of SNPs. Consequently, if this "effective number of SNPs" exists this implies that for large scale initiatives the performance of ridge regression relative to repeated simple regression might be similar to what we have observed in our simulations when $N \approx P$, even when in fact P is far greater still than N. Such a proposition would need to be tested either in empirical work or by means of simulations using actual genotype data in which linkage disequilibrium is present.

The use of GWAS data for the prediction of complex traits based on sample sizes far below 100,000 individuals yields genetic risk scores with little predictive accuracy, regardless of whether one applies the classical GWAS approach or ridge regression. However, as sample sizes approach the "effective

number of SNPs" we expect the polygenic risk score based on repeated simple regression to be able to explain a substantial proportion of the normal genetic variation. Moreover, under this scenario prediction using ridge regression is likely to outperform the classical GWAS predictions significantly. Bearing in mind that ridge regression is amenable to include nonadditive genetic variance in the prediction model it is therefore not unlikely that ridge regression will make an even more substantial contribution to the accuracy of polygenic scores in traits where epistasis and dominance are expected to play an important role.

Conflict of Interests

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

Acknowledgments

The authors thank SURFsara (http://www.surfsara.nl/) for the support in using the Lisa Compute Cluster. In addition, they thank Professor Dr. P. D. Koellinger, Professor Dr. A. R. Thurik, Professor Dr. A. Hofman, Dr. C. A. Rietveld, A. Okbay, and G. J. J. van den Burg for their support and for sharing their insights and expertise. The authors also thank an anonymous referee whose comments have helped to improve this paper.

References

[1] P. D. P. Pharoah, A. Antoniou, M. Bobrow, R. L. Zimmern, D. F. Easton, and B. A. J. Ponder, "Polygenic susceptibility to breast cancer and implications for prevention," *Nature Genetics*, vol. 31, no. 1, pp. 33–36, 2002.

[2] J. B. Meigs, P. Shrader, L. M. Sullivan et al., "Genotype score in addition to common risk factors for prediction of type 2 diabetes," *The New England Journal of Medicine*, vol. 359, no. 21, pp. 2208–2219, 2008.

[3] S. M. Purcell, N. R. Wray, J. L. Stone et al., "Common polygenic variation contributes to risk of schizophrenia and bipolar disorder," *Nature*, vol. 460, no. 7256, pp. 748–752, 2009.

[4] J. W. Smoller, K. Kendler, N. Craddock et al., "Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis," *The Lancet*, vol. 381, no. 9875, pp. 1371–1379, 2013.

[5] C. A. Rietveld, S. E. Medland, J. Derringer et al., "GWAS of 126,559 individuals identifies genetic variants associated with educational attainment," *Science*, vol. 340, no. 6139, pp. 1467–1471, 2013.

[6] C. A. Rietveld, T. Esko, G. Davies et al., "Common genetic variants associated with cognitive performance identified using the proxy-phenotype method," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 111, no. 38, pp. 13790–13794, 2014.

[7] S. M. Purcell, B. Neale, K. Todd-Brown et al., "PLINK: a tool set for whole-genome association and population-based linkage analyses," *The American Journal of Human Genetics*, vol. 81, no. 3, pp. 559–575, 2007.

[8] D. M. Evans, P. M. Visscher, and N. R. Wray, "Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk," *Human Molecular Genetics*, vol. 18, no. 18, pp. 3525–3531, 2009.

[9] A. E. Hoerl and R. W. Kennard, "Ridge regression: biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[10] N. Malo, O. Libiger, and N. J. Schork, "Accommodating linkage disequilibrium in genetic-association analyses via ridge regression," *The American Journal of Human Genetics*, vol. 82, no. 2, pp. 375–385, 2008.

[11] G. Abraham, A. Kowalczyk, J. Zobel, and M. Inouye, "Performance and robustness of penalized and unpenalized methods for genetic prediction of complex human disease," *Genetic Epidemiology*, vol. 37, no. 2, pp. 184–195, 2013.

[12] O. González-Recio, D. Gianola, N. Long, K. A. Weigel, G. J. M. Rosa, and S. Avendaño, "Nonparametric methods for incorporating genomic information into genetic evaluations: an application to mortality in broilers," *Genetics*, vol. 178, no. 4, pp. 2305–2313, 2008.

[13] D. Gianola and J. B. C. H. M. van Kaam, "Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits," *Genetics*, vol. 178, no. 4, pp. 2289–2303, 2008.

[14] G. de los Campos, D. Gianola, and G. J. M. Rosa, "Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation," *Journal of Animal Science*, vol. 87, no. 6, pp. 1883–1887, 2009.

[15] J. Crossa, G. de los Campos, P. Pérez et al., "Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers," *Genetics*, vol. 186, no. 2, pp. 713–724, 2010.

[16] J. B. Endelman, "Ridge regression and other kernels for genomic selection with R package rrBLUP," *The Plant Genome Journal*, vol. 4, no. 3, pp. 250–255, 2011.

[17] G. Morota and D. Gianola, "Kernel-based whole-genome prediction of complex traits: a review," *Frontiers in Genetics*, vol. 5, article 363, 2014.

[18] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, vol. 68, no. 1, pp. 49–67, 2006.

[19] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1418–1429, 2006.

[20] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, vol. 67, no. 2, pp. 301–320, 2005.

[21] A. Benner, M. Zucknick, T. Hielscher, C. Ittrich, and U. Mansmann, "High-dimensional Cox models: the choice of penalty as part of the model building process," *Biometrical Journal*, vol. 52, no. 1, pp. 50–69, 2010.

[22] J. O. Ogutu, T. Schulz-Streeck, and H. P. Piepho, "Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions," *BMC Proceedings*, vol. 6, supplement 2, article S10, 2012.

[23] I. E. Frank and J. H. Friedman, "A statistical view of some chemometrics regression tools," *Technometrics*, vol. 35, no. 2, pp. 109–135, 1993.

[24] H. M. Bøvelstad, S. Nygård, H. L. Størvold et al., "Predicting survival from microarray data—a comparative study," *Bioinformatics*, vol. 23, no. 16, pp. 2080–2087, 2007.

[25] W. N. van Wieringen, D. Kun, R. Hampel, and A.-L. Boulesteix, "Survival prediction using gene expression data: a review and

comparison," *Computational Statistics & Data Analysis*, vol. 53, no. 5, pp. 1590–1603, 2009.

[26] M. G. Usai, M. E. Goddard, and B. J. Hayes, "LASSO with cross-validation for genomic selection," *Genetics Research*, vol. 91, no. 6, pp. 427–436, 2009.

[27] J. C. Whittaker, R. Thompson, and M. C. Denham, "Marker-assisted selection using ridge regression," *Genetical Research*, vol. 75, no. 2, pp. 249–252, 2000.

[28] J. Yang, S. H. Lee, M. E. Goddard, and P. M. Visscher, "GCTA: a tool for genome-wide complex trait analysis," *The American Journal of Human Genetics*, vol. 88, no. 1, pp. 76–82, 2011.

[29] H. D. Patterson and R. Thompson, "Recovery of inter-block information when block sizes are unequal," *Biometrika*, vol. 58, no. 3, pp. 545–554, 1971.

[30] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society Series B: Methodological*, vol. 39, no. 1, pp. 1–38, 1977.

[31] N. Hofheinz, D. Borchardt, K. Weissleder, and M. Frisch, "Genome-based prediction of test cross performance in two subsequent breeding cycles," *Theoretical and Applied Genetics*, vol. 125, no. 8, pp. 1639–1645, 2012.

[32] C. R. Henderson, "Estimation of genetic parameters," in *Biometrics*, vol. 6, pp. 186–187, International Biometric Society, Washington, DC, USA, 1950.

[33] C. R. Henderson, "Estimation of variance and covariance components," *Biometrics*, vol. 9, no. 2, pp. 226–252, 1953.

[34] C. R. Henderson, "Selection index and expected genetic advance," *Statistical Genetics and Plant Breeding*, vol. 982, pp. 141–163, 1963.

[35] C. R. Henderson, "Best linear unbiased estimation and prediction under a selection model," *Biometrics*, vol. 31, no. 2, pp. 423–447, 1975.

[36] C. R. Henderson, "Best linear unbiased prediction of nonadditive genetic merits," *Journal of Animal Science*, vol. 60, no. 1, pp. 111–117, 1985.

[37] T. H. E. Meuwissen, B. J. Hayes, and M. E. Goddard, "Prediction of total genetic value using genome-wide dense marker maps," *Genetics*, vol. 157, no. 4, pp. 1819–1829, 2001.

[38] L. R. Schaeffer, "Strategy for applying genome-wide selection in dairy cattle," *Journal of Animal Breeding and Genetics*, vol. 123, no. 4, pp. 218–223, 2006.

[39] J. Sherman and W. J. Morrison, "Adjustment of an inverse matrix corresponding to a change in one element of a given matrix," *The Annals of Mathematical Statistics*, vol. 21, pp. 124–127, 1950.

[40] M. Woodbury, "Inverting modified matrices," Memorandum Report 42, Statistical Research Group, Princeton University, 1950.

[41] S. R. Searle, G. Casella, and C. E. McCulloch, *Variance Components*, John Wiley & Sons, Hoboken, NJ, USA, 2006.

[42] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *The Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[43] J. Sabourin, A. B. Nobel, and W. Valdar, "Fine-mapping additive and dominant SNP effects using group-LASSO and fractional resample model averaging," *Genetic Epidemiology*, vol. 39, no. 2, pp. 77–88, 2015.

[44] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, vol. 2, Springer, New York, NY, USA, 2009.

[45] T. A. Manolio, F. S. Collins, N. J. Cox et al., "Finding the missing heritability of complex diseases," *Nature*, vol. 461, no. 7265, pp. 747–753, 2009.

[46] D. Speed, G. Hemani, M. R. Johnson, and D. J. Balding, "Improved heritability estimation from genome-wide SNPs," *The American Journal of Human Genetics*, vol. 91, no. 6, pp. 1011–1021, 2012.

[47] T. Hastie and R. Tibshirani, "Efficient quadratic regularization for expression arrays," *Biostatistics*, vol. 5, no. 3, pp. 329–340, 2004.

[48] G. S. Kimeldorf and G. Wahba, "A correspondence between Bayesian estimation on stochastic processes and smoothing by splines," *The Annals of Mathematical Statistics*, vol. 41, pp. 495–502, 1970.

[49] C. C. Chang, C. C. Chow, L. C. A. M. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee, "Second-generation PLINK: rising to the challenge of larger and richer datasets," *Gigascience*, vol. 4, no. 7, 2015.

[50] X. Shen, M. Alam, F. Fikse, and L. Rönnegård, "A novel generalized ridge regression method for quantitative genetics," *Genetics*, vol. 193, no. 4, pp. 1255–1268, 2013.

[51] N. Hofheinz and M. Frisch, "Heteroscedastic ridge regression approaches for genome-wide prediction with a focus on computational efficiency and accurate effect estimation," *G3: Genes, Genomes, Genetics*, vol. 4, no. 3, pp. 539–546, 2014.

[52] A. Aizerman, E. M. Braverman, and L. I. Rozoner, "Theoretical foundations of the potential function method in pattern recognition learning," *Automationand Remote Control*, vol. 25, pp. 821–837, 1964.

[53] N. Aronszajn, "Theory of reproducing kernels," *Transactions of the American Mathematical Society*, vol. 68, pp. 337–404, 1950.

[54] D. A. Harville, "Discussion on a section on interpolation and estimation," in *Statistics: An Appraisal*, pp. 281–286, 1983.

[55] T. Speed, "[That BLUP is a good thing: the estimation of random effects]: comment," *Statistical Science*, vol. 6, no. 1, pp. 42–44, 1991.

[56] H. P. Piepho, "Ridge regression and extensions for genomewide selection in maize," *Crop Science*, vol. 49, no. 4, pp. 1165–1176, 2009.

[57] G. Morota, M. Koyama, G. J. M. Rosa, K. A. Weigel, and D. Gianola, "Predicting complex traits using a diffusion kernel on genetic markers with an application to dairy cattle and wheat data," *Genetics Selection Evolution*, vol. 45, no. 1, article 17, 2013.

[58] L. Tusell, P. Pérez-Rodríguez, S. Forni, and D. Gianola, "Model averaging for genome-enabled prediction with reproducing kernel Hilbert spaces: a case study with pig litter size and wheat yield," *Journal of Animal Breeding and Genetics*, vol. 131, no. 2, pp. 105–115, 2014.

[59] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[60] F. Dudbridge, "Power and predictive accuracy of polygenic risk scores," *PLoS Genetics*, vol. 9, no. 3, Article ID e1003348, 2013.

[61] H. Warren, J.-P. Casas, A. Hingorani, F. Dudbridge, and J. Whittaker, "Genetic prediction of quantitative lipid traits: comparing shrinkage models to gene scores," *Genetic Epidemiology*, vol. 38, no. 1, pp. 72–83, 2014.

[62] J. Yang, B. Benyamin, B. P. McEvoy et al., "Common SNPs explain a large proportion of the heritability for human height," *Nature Genetics*, vol. 42, no. 7, pp. 565–569, 2010.

[63] F. Dudbridge and A. Gusnanto, "Estimation of significance thresholds for genomewide association scans," *Genetic Epidemiology*, vol. 32, no. 3, pp. 227–234, 2008.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 319797, 11 pages
http://dx.doi.org/10.1155/2015/319797

Research Article

FARMS: A New Algorithm for Variable Selection

Susana Perez-Alvarez,[1] Guadalupe Gómez,[2] and Christian Brander[1,3,4]

[1] *AIDS Research Institute IrsiCaixa-HIVACAT, Hospital Universitari Germans Trias i Pujol, Universitat Autònoma de Barcelona, 08916 Badalona, Spain*
[2] *Universitat Politècnica de Catalunya, 08034 Barcelona, Spain*
[3] *Institució Catalana de Recerca Avançada (ICREA), 08010 Barcelona, Spain*
[4] *University of Vic and Central Catalonia (UVIC-UCC), 08500 Vic, Spain*

Correspondence should be addressed to Susana Perez-Alvarez; susana.perez@biokit.com

Received 16 January 2015; Revised 13 March 2015; Accepted 13 March 2015

Academic Editor: Junwen Wang

Large datasets including an extensive number of covariates are generated these days in many different situations, for instance, in detailed genetic studies of outbreed human populations or in complex analyses of immune responses to different infections. Aiming at informing clinical interventions or vaccine design, methods for variable selection identifying those variables with the optimal prediction performance for a specific outcome are crucial. However, testing for all potential subsets of variables is not feasible and alternatives to existing methods are needed. Here, we describe a new method to handle such complex datasets, referred to as FARMS, that combines forward and all subsets regression for model selection. We apply FARMS to a host genetic and immunological dataset of over 800 individuals from Lima (Peru) and Durban (South Africa) who were HIV infected and tested for antiviral immune responses. This dataset includes more than 500 explanatory variables: around 400 variables with information on HIV immune reactivity and around 100 individual genetic characteristics. We have implemented FARMS in *R* statistical language and we showed that FARMS is fast and outcompetes other comparable commonly used approaches, thus providing a new tool for the thorough analysis of complex datasets without the need for massive computational infrastructure.

1. Introduction

An important goal of biosciences research is the identification of traits, markers, or features associated with an outcome of interest. Such analyses may include complex datasets that can include thousands of host and pathogen genetic markers as well as clinical and experimental parameters that need to be probed for possible associations. One such challenging situation is the infection with Human Immunodeficiency Virus (HIV). In this case, the design of an effective HIV vaccine depends on a comprehensive delineation of immune correlates of controlled infection that factors in host genetic determinants, clinical parameters, and viral replication data. Regression models, which interrogate possible mathematical relation(s) between the outcome of interest and a subset of relevant predictive characteristics, provide a convenient framework for this identification [1, 2]. In this paper, we have combined biostatistics and bioinformatics knowledge to present a new regression method, FARMS, and demonstrate its utility in an exemplary study where the relationship between immune responses to HIV and relative disease control is being explored.

A crucial and difficult step when building a regression model is the selection of variables to be included in the final model from a set of plausible and meaningful variables [3]. The selection problem has been widely discussed: Hocking noted the importance and interest of selecting the best subset of variables and concluded that there is no unique answer to this problem [4]. Indeed, variable selection can be seen as a special case of the model selection problem where each

model under consideration could correspond to a distinct subset of explanatory variables [5–7]. There is a rich literature discussing the appropriateness of stepwise methods, computational efficiency aspects, or False Discovery Rate selection control [8–13].

When trying to define the best subset of explanatory variables, one would be tempted to perform an all subsets regression analysis in order to identify the best predictive subset given a certain optimization criterion. However, if the number n of variables in the original dataset is exceedingly large, an exhaustive test of all possible variable subsets is generally not feasible, since fitting all the possible models can be prohibitively time-intensive [14, 15]. Recognizing these limitations, several methods have been proposed to reduce the model space, although this reduction might lead to a lack of interpretability and usability [7]. Some authors have proposed rules for shrinkage methods, suggesting that one could safely evaluate problems when there are at most 10 to 15 times more variables than observations, while others have established criteria to delete variables, for example, *Least Absolute Shrinkage and Selection Operator* (LASSO) [11, 16–18].

In this paper, we present FARMS, as acronym for forward and all subsets regression for model selection, a new algorithm for model selection based on Liebminger et al.'s proposal [19]. The unique feature of this algorithm is that it combines forward variable selection and all subsets regression with improved prediction performance. Importantly, it also requires less computational time than comparable approaches. The motivation to develop this algorithm was the urgent need to define immune parameters of controlled HIV infection that could guide HIV vaccine development. The special challenges with such analyses are the extreme outbred population structure of the human host organism, the extensive number of T cell responses an individual can mount in response to HIV infection, and viral parameters, especially viral genome sequence polymorphisms.

HIV disease status is commonly assessed by determining an individual's CD4 T cell count and his/her steady state viral load, also referred to as viral set point. It is this viral set point that largely predicts how fast an individual will develop AIDS. However, the mechanisms that determine the in vivo viral set point remain poorly defined and likely include a multitude of factors, including viral replicative fitness, viral sequence diversity and host genetics as well as quantitative and qualitative determinants of the cellular and humoral host immune response [20]. The cellular immune response to HIV, mediated by CD4+ and CD8+ T cells, can be assessed by relatively straight forward in vitro analyses that determine the number of T cells in the body that can interact with short antigenic viral determinants, referred to as "T cell epitopes." To this end, an in vitro ELISPOT assay can be employed in which an individual's peripheral blood T cells are incubated with synthetic peptides (generally 15–20 amino acids in length) representing the viral proteome [21, 22]. In our analyses, there are 410 such partially overlapping peptides (OLP), numbered from 1 to 425, as there are some gaps in the not entirely continuous numbering [21]. Of importance, reactivity to these 410 OLP depends, among other factors, on the presence or absence of genes in the human leukocyte antigen (HLA) [23] locus on chromosome 6, particularly the HLA class I alleles [24]. These HLA class I alleles encode specialized molecules that present small viral peptides on the surface of infected cells to host's CD8 T cells, also referred to as cytotoxic T lymphocytes (CTL). The classical HLA class I alleles include HLA-A, HLA-B, and HLA-C alleles, of which each individual encodes 2 versions that are inherited from the parents. There are close to 10,000 different allelic variants for the HLA-A, HLA-B, and HLA-C loci together (http://hla.alleles.org/nomenclature/stats.html). For each CD8+ T cell immune response detected in in vitro assays one of the 6 HLA class I molecules "restricts" a T cell response; that is, the T cell can react with viral antigen in the context of this class I molecule. For effective antiviral immunity and vaccine design alike, it is thought important that many HLA class I alleles can be involved in the immune defense both on the individual level and the population level as this ensures a broad antiviral T cell reactivity. As a consequence, large population screenings in HIV-infected individuals will provide a multitude of genetic, clinical, and experimental parameters. When intending to define best models of OLP reactivity and host HLA genetics that are associated with relative HIV control, this considerable number of parameters poses a formidable challenge for the necessary in-depth analysis.

2. Material and Methods

2.1. Datasets. The datasets included patients infected with HIV from two geographic areas: 631 patients from Durban (South Africa) and 236 patients from Lima (Peru) form two different cohorts with population-specific host genetics and immune reactivity data [25]. These two datasets were analyzed by means of FARMS aiming at assessing the predictive value of immune and genetic parameters for in vivo HIV viral loads in these untreated subjects. Protocols were approved in Lima by the IMPACTA Human Research Committee and in Durban by the Ethical Committee of the Nelson R. Mandela School of Medicine at the University of KwaZulu-Natal. All subjects provided written informed consent. We gathered information about HIV viral load, human leukocyte antigen [23] polymorphisms, and the reactivity of each of 410 individual overlapping peptides (OLP) spanning the entire viral proteome. The main characteristics of this cohort are summarized in Table 1. Although the HIV viral load of the patient, given as number of viral copies per mL, is a continuous variable, we note that there exists a lower detection limit depending on the technology used. In the present case, the lower detection limit was 50 copies/mL and, for the purpose of the analyses, the values were set as 49 copies/mL. The Box-Cox transformation was used to normalize the data: the lambda parameter was estimated (lambda = 0.061) and log10 transformation was applied accordingly. After logarithmic transformation, the data normality was assessed graphically by histograms, density plot and Q-Q plot as well as by the Shapiro-Wilk test.

Table 1: Data description. Summary of cohort characteristics either joined or split by country (Lima and Durban) showing viral load and its transformation log10 distribution. Below, number of HLA alleles and OLP present in the database and frequency of the top 5 parameters for each category.

	Viral load			Log10 (Viral load)		
	Global	Lima	Durban	Global	Lima	Durban
Median	37800	37240	37900	4.577	4.571	4.579
IQR	(8715; 131500)	(13310; 109000)	(7075; 138500)	(3.94; 5.119)	(4.124; 5.038)	(3.85; 5.141)
	HLA			OLP		
	Global	Lima	Durban	Global	Lima	Durban
Total number	73	62	66	406	391	371
1st	C∗07	A∗02	B∗15	76	76	78
	($N = 310$; 6.31%)	($N = 168$; 12.80%)	($N = 233$; 6.48%)	($N = 293$; 2.77%)	($N = 76$; 2.09%)	($N = 234$; 3.37%)
2nd	A∗02	B∗35	C∗07	78	84	76
	($N = 294$; 5.99%)	($N = 99$; 7.54%)	($N = 222$; 6.17%)	($N = 284$; 2.69%)	($N = 63$; 1.74%)	($N = 217$; 3.12%)
3rd	B∗15	C∗04	A∗30	84	81	84
	($N = 278$; 5.66%)	($N = 88$; 6.70%)	($N = 205$; 5.70%)	($N = 262$; 2.48%)	($N = 61$; 1.68%)	($N = 199$; 2.86%)
4th	A∗30	C∗07	C∗06	25	85	25
	($N = 225$; 4.58%)	($N = 88$; 6.70%)	($N = 194$; 5.39%)	($N = 192$; 1.82%)	($N = 53$; 1.46%)	($N = 178$; 2.56%)
5th	C∗04	B∗39	A∗68	41	78	41
	($N = 217$; 4.42%)	($N = 59$; 4.49%)	($N = 162$; 4.50%)	($N = 190$; 1.80%)	($N = 50$; 1.38%)	($N = 151$; 2.17%)

Information on the 2 HLA-A, 2 HLA-B, and the 2 HLA-C alleles that every individual possesses was collected for both cohorts. The HLA genes in the Peruvian cohort were determined using a method that was sufficiently sensitive to discriminate between different subtypes (i.e., variants) of alleles at a 4-digit precision, while the Durban data consisted of two-digit typing. New binary variables were generated, in both Lima and Durban datasets, as follows: (i) for each allelic variant on each of the three HLA loci (HLA-A, HLA-B, and HLA-C) a binary variable was created with value 1 if the subject had that respective variant and 0 otherwise; (ii) for each OLP a binary variable was set with value 1 if the subject has responded to this OLP and 0 otherwise.

2.2. FARMS Algorithm. FARMS is an algorithm that works by iterating several steps according to a specified criterion. Liebminger et al. proposed in 2007 [19] an algorithm for model selection that combined forward variable selection and all subsets regression. In their work, they compared their proposed algorithm with a genetic algorithm and the standard stepwise method applying all of them to three different datasets. They described how their algorithm selected models with improved prediction performance in addition to requiring less computational time. We have modified this algorithm in order to include additional features and lend it more flexibility according to our specific needs when analyzing HLA genetic data and human immune reactivity data against HIV. The FARMS algorithm is available on the GRASS web page at http://grass.upc.edu/software/farms_code/view.

The basic FARMS algorithm is illustrated in Figure 1 by means of a toy example consisting of a dataset of 10 variables.

Let P be the set of all potential explanatory variables of the outcome of interest. Define two subsets within P:

(a) P_s: the set of variables included in an initial model. This set of variables is the starting point for the exploration of P.

(b) P_F: the set of variables to be forced into a final model, not including those already in P_s.

The algorithm iterates over the following steps:

(1) Start by building model M_B with the variables P_s and P_F.

(2) The subset of variables not included in M_B is divided into user-defined fixed-sized groups of variables: P_a.

(3) For each P_a, build a new model based on fitting the variables P_a, P_s, and P_F and select the best subset of variables according to a predefined criterion C_1.

(4) Select the best model M_{BS} among all the best subsets obtained in step (3) following the criterion C_1.

(5) Compare models M_B and M_{BS}. While M_{BS} is better than M_B, define as P_s the explanatory variables included in M_{BS}, go to step (2) and start all over or until a user-defined maximum time of computation is reached. Otherwise go to step (6).

(6) M_B is declared the best model and the algorithm finishes.

The user can choose to specify

(i) the criterion C_1 to select a subset of variables among a group of them,

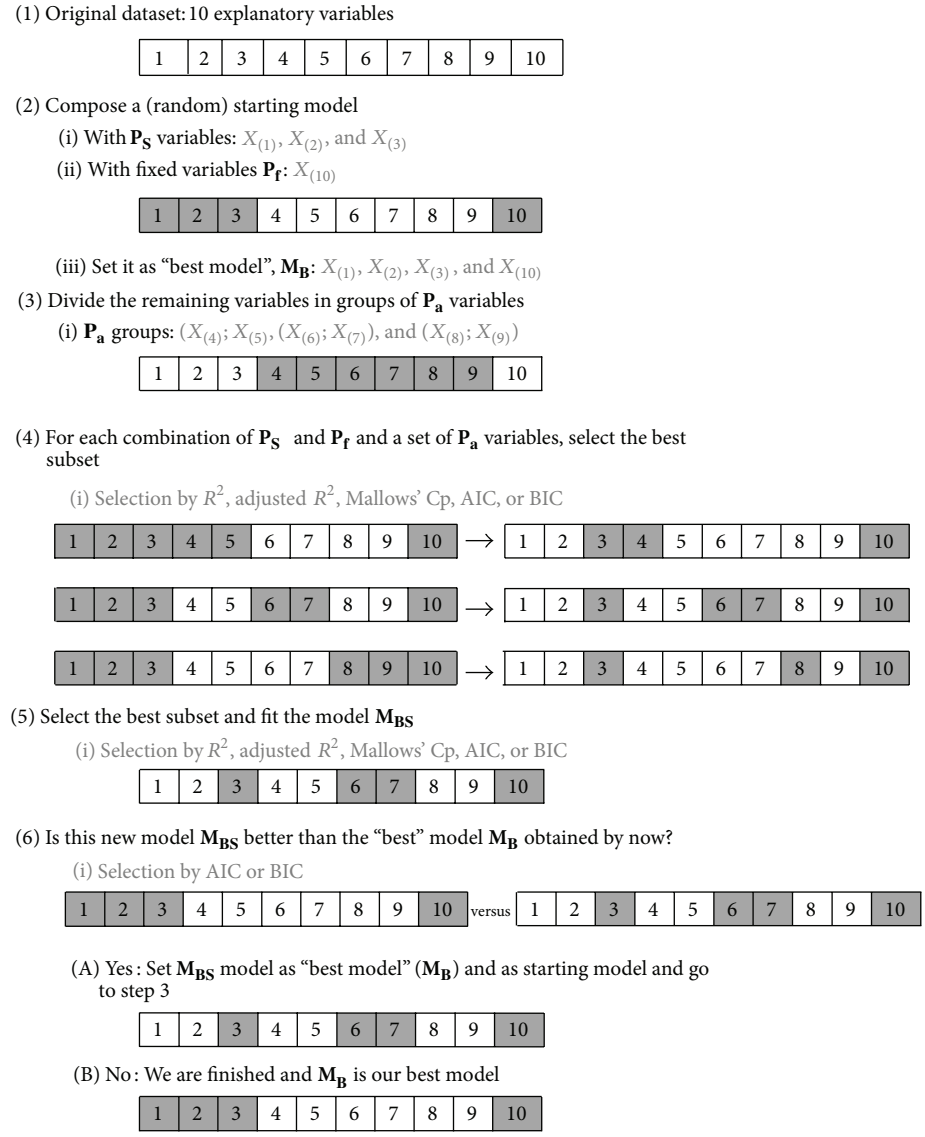


Figure 1: Illustration of FARMS algorithm. On a dataset of 10 covariates, variable "10" is forced to be always included. The starting model includes 3 variables (in addition to the forced-in ones) and adds 2 more variables in each iteration.

(ii) the criterion C_2 to select between two candidate models,

(iii) the "starting" variables P_S in the initial model: the user can specify the size of the set P_S as well as the initial variables. If this is not specified, FARMS has a default size value and by default chooses P_S randomly from all the available variables, that is, within P-P_F,

(iv) the variables forced into the final model P_F: the user can specify some variables that should always be included in the final selected model or leave this option blank and allow for the selection of variables among all the ones included in the original dataset (P),

(v) the total number of variables into the final model can be fixed as well. By default, this number is the total number of variables in the dataset; so the best model could, a priori, include all the available variables.

The selection of the best model can be made according to any of the following criteria: Mallows' Cp [26], Akaike's information criterion (AIC) [27], and the Schwarz or Bayesian information criterion (BIC) [28]. The best subset selection can be additionally based on the R-square, the adjusted R-square, and the residual sum of squares (RSS). All these criteria originated from very different points of view: Mallows' Cp measures the performance of the variables in terms of the standardized total mean squared error of prediction (MSE) providing a balance between the lack of fit and the complexity of the model for the observed data. The AIC selects the candidate model that minimizes the estimated information lost when the model is used to approximate

the process that generated the observed data (full reality) and it is a good option for prediction. The BIC, optimal for explaining the relationship between outcome and covariates, selects the model that maximizes the posterior probability applying a severe penalty term for the number of parameters in the model. Essentially, AIC and BIC will give similar answers with BIC leading to simpler models than AIC because it applies a larger penalty for complex models [29, 30]. Finally, the residual sum of squares, the coefficient of determination R-square, and the adjusted R-square are used [14] to measure the overall fit of a model, with the latter being the preferred alternative because it penalizes the R-square when extra variables are included in the model.

Data input and output formats to use FARMS are kept very user-friendly. In brief, for data input, any database in a format that R can read (i.e., an Excel, TXT, or CSV file) is suitable. Alternatively, data files generated by any statistical software that are compatible with R (such as SPSS or Stata) can be used. FARMS output is composed of 2 parts: first, an external txt file, written by R on the working directory, that adds a line on each algorithm iteration indicating the following information: the number of iteration, time since beginning of execution, value of the AIC and the BIC for the selected model at this point in the analysis, and for each variable on the dataset, an indication whether it is included in the model in that iteration (value = 1) or not included (value = 0) and second, an R object within the software environment including the information of the last model selected (information already included in the file). This object allows the user to easily fit the selected/best model within the R environment and continue the data analysis. An illustration of data input format and data output format is shown in Figure 2.

2.3. Univariate Data Analysis. Univariate analysis for the two cohorts (Lima and Durban) was performed over each HLA subtype present (Supplementary Table 1 in Supplementary Material available online at http://dx.doi.org/10.1155/2015/319797) and each reactive OLP (Supplementary Table 2) and related to HIV viral loads. The two cohorts were both analyzed separately and pooled into a single dataset. Specifically, for each HLA allele and for each OLP present in the database, we wondered whether people having the allele under analysis and/or responding to the OLP under analysis showed a significantly higher or lower viral load than the rest of the cohort. These questions were addressed by means of a bilateral Student's t-test to compare values of viral loads between the two groups defined as having or not having a specific HLA allele or making a response or not to an individual OLP. Results were reported as P values for a 95% confidence and corrected with q values setting the False Discovery Rate (FDR) to 10%.

2.4. Regression Methods. The following regression methods other than FARMS were applied to these two cohorts in order to find a relationship between the viral load and more than one HLA or OLP at a time: all subsets regression, forward selection regression, backward stepwise regression, and forward stepwise regression which is a combination of the previous two and which is later referred to as "Forward Stepwise". These regression methods, all of them criticized for not being able to reflect model uncertainty accurately enough, were included as benchmarks for FARMS performance. For the implementation of the regression methods, few further technicalities were taken into account:

(i) We randomly sorted HLAs and OLPs variables before applying FARMS or any of the other approaches to avoid any possible relation between the order of the variables in the dataset and either the time of execution or the obtained model.

(ii) For this methodological comparison we previously evaluated FARMS not only to assess its robustness but also to choose the best value for the parameters Ps and Pa in terms of time and optimal statistical properties in the model. Therefore, the number of variables in the starting model (*Ps*) and the number of variables to be added in each step (*Pa*) were set according to the best results.

2.5. FARMS Robustness. In order to assess how robustly FARMS performs, that is, how stable the final model remains as the number of variables included initially or added in each iteration of the algorithm was changed, we executed the algorithm by varying the values of the number of variables in the starting model and by changing the number of variables to add on each step, that is, the size of groups to divide the rest of variables. We let both parameters vary between 2 and 21 and ran FARMS for each one of the 400 possible combinations in two different scenarios with completely different variables, HLA alleles and OLP. It is important to remember that, on each execution, the starting model or starting variables were selected randomly as we did not specify which ones to use. As a by-product, we have used those executions to study the computational time needed and the number of iterations internally performed.

3. Results

3.1. Predicting HIV Viral Loads in terms of Response Rates to OLP and Presence of HLA Alleles. When we applied FARMS to the specific HIV-derived datasets presented here, individual OLP and OLP combinations were identified, which have a high predictability of viral load in chronically HIV infected subjects. Our FARMS based analyses also showed that OLP reactivity (i.e., the specificity of the T cell response to HIV) had a greater predictability than host genetics (HLA class I genes) [25]. From a vaccine point of view, this revelation is crucial as it indicates that individuals with poor HLA genetics still have a possibility to mount effective T cell responses to HIV and that vaccine immunogen design could profit from a focus on such relevant regions in the viral proteome.

3.2. FARMS Evaluation. The best model selected in all the 400 FARMS runs was always the same regardless of the number of variables included initially or added in each

	vresp	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
1	-0.6264538	0	0	0	1	0	1	1	0	0	1
2	0.1836433	0	1	1	0	1	1	0	0	0	0
3	-0.8356286	1	1	0	0	1	0	0	1	0	0
4	1.5952808	1	0	0	0	0	0	0	1	0	1
5	0.3295078	0	1	1	1	0	1	0	0	0	0
⋮											
N	-0.3053884	0	1	1	0	0	0	0	1	1	0

(a) Example of dataset input format

```
farms(df[1],df[2:11],addSize=2,startSize=3,
covarf=c(10),resname="OutputFileName")
```

(b) Example of FARMS function call

```
$usedTime
   [1] 0 0
$AIC
   [1] 269.9368 267.5311
$BIC
   [1] 285.5679 277.9518
$Sigma
   [1] 0.9015789 78.4611773
$RSq
   [1] 0.03316931 0.01763214
$AdjRSq
   [1] -0.007539346 -0.002622864
$MallowsCp
   [1] 0.000000 3.945667
$NumIter
   [1] 3
$models
          [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9][,10]
   models  1    0    0    1    0    0    1    1    0    0
     m     1    0    0    0    0    0    0    1    0    0
$Data
      X10  X1  X2  X3  X4  X5  X6  X7  X8  X9
   1   1   0   0   0   1   0   1   1   0   0
   2   0   0   1   1   0   1   1   0   0   0
   ⋮
   N   0   0   1   1   0   1   1   0   0   0
```

(c) Example of output from FARMS function within *R* environment

```
Time;AIC;BIC;Sigma;R^2;Adj R^2;MallowsCp;X10;X1;X2;X3;X4;X5;X6;X7;X8;X9;
0;269.93683771477;285.56785883069;0.901578919774;0.0331693140239;-0.0075393464381;1.238457762547;1;0;0;1;0;0;1;1;0;0
0;267.53108264467;277.95176338863;78.46117732028;0.0176321436344;-0.0026228637132;3.945667016824;1;0;0;0;0;0;0;1;0;0
```

(d) Example of TXT file output from FARMS function

Figure 2: Illustration of FARMS input and output. (a) FARMS function requires data as *R* data frame containing all the variables (response and explanatory variables). (b) Calling the FARMS function within *R* requires the indication of at least the response variable and the explanatory variables. In this illustration, which follows the explanation of Figure 1, we also indicate the number of variables to add in each iteration, the number of variables to compose the starting model, the columns containing the forced-in variables, and the name of the output file. (c) By default, FARMS function returns an *R* object containing information for each iteration (two iterations in this illustration) and the dataset as processed by the algorithm. (d) Optionally, FARMS function can produce a text file containing the same information as the *R* object output, adding a text line for each iteration, helping also to monitor the algorithm execution and to track the evolution of models until the final model is obtained.

Table 2: Comparison of results when basing variable selection on the FARMS algorithm or common strategies. The number of covariates included in the final model excludes the "forced in" covariates. FARMS parameters used in this case are number of adding covariates = 8, number of starting covariates = 10, and the selecting criteria for both best subset and best model was the BIC. (All runs were executed on an Intel Xeon x5680 machine with 6 CPU cores and 95 GB RAM memory under a Linux Suse 11.0 OS).

		Time (seconds)* Mean (IQR)	Number of vars.**	BIC	AIC	R^2	Adj. R^2
HLA	FARMS	**1.27 (1.00; 1.40)**	**9**	**2235.4**	**2183.03**	**11.67%**	**10.74%**
	All subsets	>1 month	—	—	—	—	—
	Forward selection[1]	3.84 (3.13; 3.52)	17	2259.4	2168.86	14.69%	12.98%
	Forward stepwise[1]	4.32 (3.51; 3.90)	17	2259.4	2168.86	14.69%	12.98%
	Forward selection[2]	2.01 (1.62; 1.88)	10	2235.45	2178.27	12.36%	11.33%
	Forward stepwise[2]	2.35 (1.89; 2.18)	9	2235.44	2183.03	11.67%	10.74%
	Forward selection[3]	1.99 (1.61; 1.88)	10	2235.45	2178.27	12.36%	11.33%
	Forward stepwise[3]	2.38 (1.92; 2.22)	9	2235.44	2183.03	11.67%	10.74%
	Backward stepwise[3]	13 s	10	2236.52	2174.58	12.93%	11.81%
OLP	FARMS	**33.4 (19.47; 37.95)**	**12**	**2224.8**	**2158.11**	**14.77%**	**13.57%**
	All subsets	>1 month	—	—	—	—	—
	Forward selection[1]	393.2 (324.40; 336.60)	79	2396.53	2010.56	38.40%	32.22%
	Forward stepwise[1]	545.9 (451.70; 469.70)	83	2415.46	2010.53	38.97%	32.51%
	Forward selection[2]	401.7 (329.50; 343.40)	80	2403.3	2012.56	38.40%	32.13%
	Forward stepwise[2]	462.4 (382.50; 401.50)	76	2385.18	2013.51	37.76%	31.77%
	Forward selection[3]	38.09 (31.34; 33.11)	12	2224.82	2158.11	14.77%	13.57%
	Forward stepwise[3]	38.31 (31.34; 33.43)	12	2224.82	2158.11	14.77%	13.57%
	Backward stepwise[3]	>12 hours	23	2232.63	2108.63	21.68%	19.45%

[1]Selection by AIC and base model with intercept-only.
[2]Selection by AIC and base model with forced-in variables.
[3]Selection by BIC and base model with forced-in variables.
*Time obtained after 100 executions for each scenario.
**Including forced-in variables.

iteration of the algorithm in either of the two different scenarios described at the end of Section 2. This homogeneity was only altered 11 times when using the OLP variables. In 389 runs, the final selected model contained a total of 12 variables with BIC = 2224.82. However for the remaining 11 combinations the final selected model included 15 variables, 5 of them different from the previous 12, and had a slightly smaller BIC (2218.85). In terms of prediction, the dominant model has a coefficient of determination close to 15% in contrast to 18% in the less frequent model.

As expected, these two varying parameters affected the number of iterations that FARMS needed to reach the best model and also impacted the computational time required. Figure 3 shows the evolution of time until the best model is selected for the HLA-only and the OLP-only models according to the number of variables in the starting model and the size of the adding subset. These results indicate that the number of covariates in the starting model does not have a strong effect on the total time of execution, although a large number of variables produce more variability: when the number of starting covariates is larger than 15, the time needed increases more than 50-fold. Compared to the starting size of variables, the number of variables to be added on each step has a much more profound effect on the total time of execution: as the number of variables to be added increases, the time also increases.

From these modulations and in light of the aims of the study (i.e., identifying a limited number of OLP that could be incorporated into an HIV vaccine sequence), we determined the optimum values of these two parameters: the number of variables for the starting model was set to 10, and the number of adding variables was set to 8.

3.3. FARMS Performance in Head-to-Head Comparisons with Other Methods. In order to compare the performance of FARMS with other established approaches, the data from Lima and Durban were used (Table 2). All subsets regression analysis exceeded one month of execution time for both scenarios (HLA and OLP variables), thus not providing us with a precise time to completion. The analysis using backward stepwise algorithm needed much more computational time, at least 3-fold longer, than the other stepwise approaches or FARMS and was thus only run once. The selected models using forward regression or forward stepwise took longer

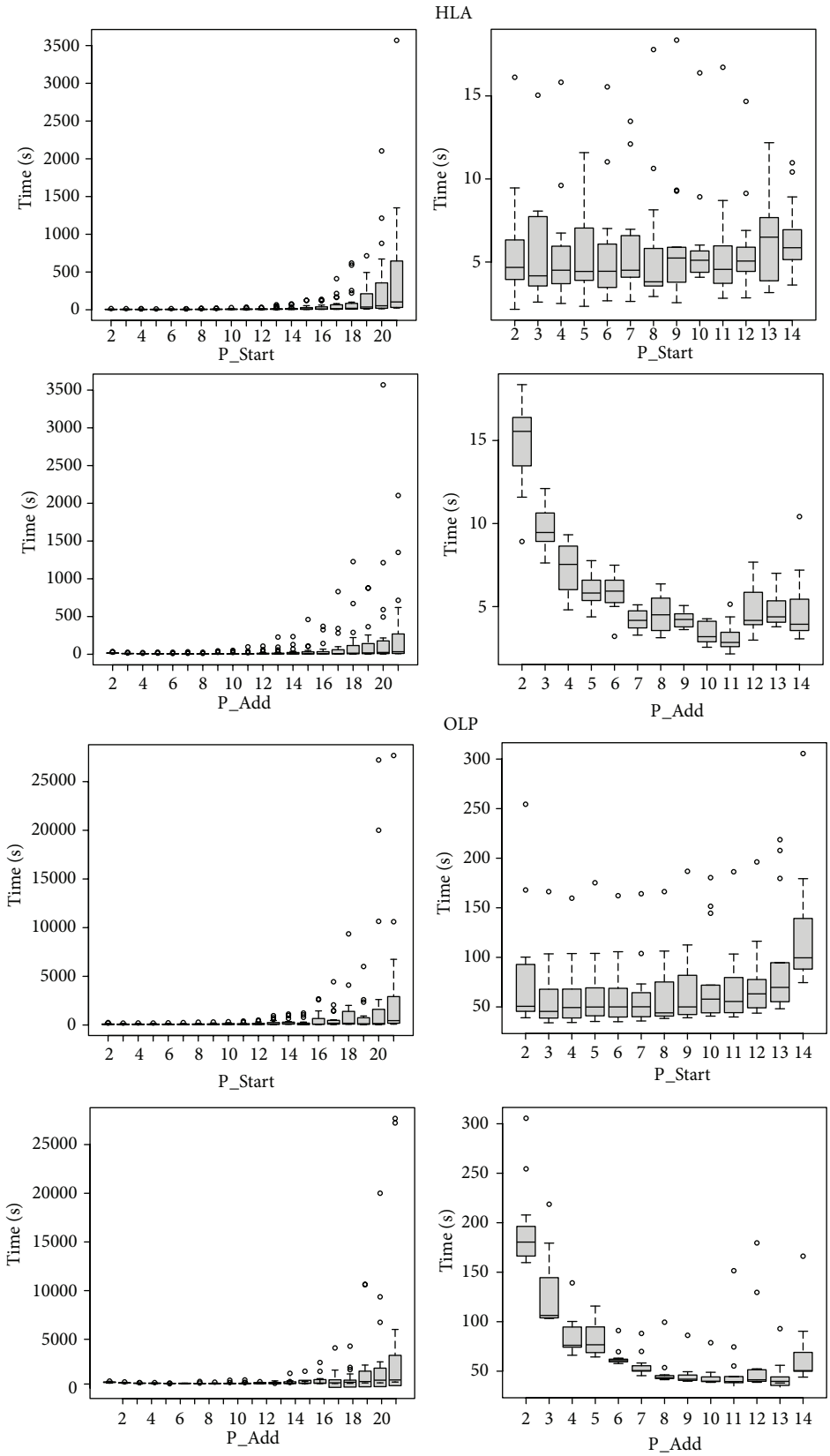


Figure 3: Evolution of the computing time (in seconds) needed to reach the final model according to FARMS parameters: starting number of covariates (P_Start) and number of covariates added in each step (P_Add). First two rows of figures refer to the HLA-only model; rows 3 and 4 correspond to OLP-only covariates model.

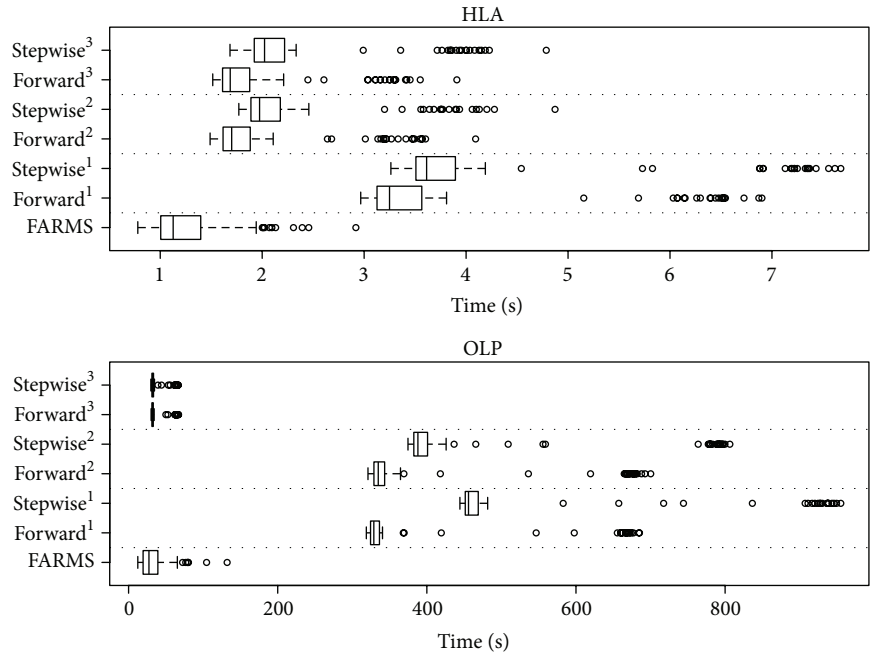


Figure 4: Comparison of computational time between FARMS, forward selection, and forward stepwise regression algorithms for the two possible scenarios, HLA-only and OLP-only. (1: selection by AIC and base model with intercept-only; 2: selection by AIC and base model with forced-in variables; 3: selection by BIC and base model with forced-in variables).

than FARMS when the selecting criterion was either the AIC or the BIC.

Models obtained with stepwise methods varied as we changed their parameters: selection by AIC or BIC and intercept-only baseline model or with forced-in variables. When setting the more appropriate parameters to approximate FARMS and stepwise methods (selection based on BIC and setting the model with forced-in variables as baseline model), forward selection provided a slightly worse model when using HLA variables; otherwise, the models obtained with FARMS and stepwise methods were the same and only differed in the time needed. These comparisons further confirm the validity of FARMS approach and demonstrate its favorable computational time.

To further investigate the time differences between FARMS and more comparable stepwise approaches, we ran all the methods in the two scenarios (HLA and OLP) 100 times and compared their execution time by Student's t-test: Figure 4 summarizes the time of execution needed by each approach in addition to the time used in each run. FARMS was significantly faster than stepwise approaches applied to HLA-only covariates (76 covariates): FARMS usually needed one second, whereas both forward regression and forward stepwise approaches needed up to 3 times longer to completion (F-test P value $< 2.2e - 16$). This difference increased and showed FARMS to be also significantly faster when the OLP-only (406 covariates) dataset was evaluated as FARMS completed the task roughly 10x faster than standard forward or stepwise approaches (F-test P value $< 2.2e - 16$) (Table 1). Thus, FARMS offers a novel, versatile, and flexible approach that is not dependent on excessive computational power and runs in widely available *R* packages.

4. Discussion

We here report the development and performance of a new algorithm for variable selection referred to as FARMS. Overall, this new algorithm, when applied to high-dimensional datasets, yields better results in less time compared to other regression approaches (stepwise or all subsets techniques).

Stepwise algorithms use as a starting model either the intercept-only or the full model and build on this model by introducing or excluding variables one by one with no reference to other variables until the best model is selected according to an information criterion. The method we propose here starts with a random model containing some variables but not a full model and improves it by selecting subsets of variables with different sizes, allowing the algorithm to include in a single step more than one variable.

The common information criterion used in model selection is the AIC, but selection by BIC has been suggested to represent a better option for our specific situation: BIC penalizes the number of covariates, helping our goal of finding best model with fewer covariates. In addition, the BIC criterion leads to the final model faster when the dataset contains a large number of covariates. FARMS also implements other criteria that facilitate obtaining better models according to their prediction capacity, such as R-square or adjusted R-square. These improvements in the FARMS implementation allow for making a better comparison of those methodologies

and enable the user to change the model selection according to the purpose of the data analysis.

Additional features that can be readily implemented in FARMS and can be very useful for statistical analyses include the evaluation of quadratic terms (the square of the value of each covariate), which can help to explain nonlinear relationship among the variables. In addition, FARMS offers the possibility of introducing the evaluation of the interaction terms, although for large datasets this may increase completion time considerably (but still less than needed by common approaches such as stepwise). Finally, the FARMS strategy can also be extended to other regression approaches, such as logistic regression, which is widely used in the biomedical and genetics field, as there are already packages in *R* that select best subsets of variables for logistic regression.

Conflict of Interests

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

Acknowledgments

This work was partially supported by Grant no. MTM2012-38067-C02-01 from the Ministerio de Economía y Competitividad del Gobierno de España, Grant no. 2014 SGR 464 from the Departament d'Economia i Coneixement de la Generalitat de Catalunya, the HIVACAT program, FIS (Grants nos. PS0900283 and FIPSE36-0737-09), and the European Community (CUT'HIVAC; Grant no. EC-7FP-241904). Christian Brander is an ICREA Senior Research Professor at IrsiCaixa.

References

[1] E. Núñez, E. W. Steyerberg, and J. Núñez, "Regression modeling strategies," *Revista Española de Cardiología (English Edition)*, vol. 64, no. 6, pp. 501–507, 2011.

[2] E. W. Steyerberg, B. van Calster, and M. J. Pencina, "Performance measures for prediction models and markers: evaluation of predictions and classifications," *Revista Espanola de Cardiologia*, vol. 64, no. 9, pp. 788–794, 2011.

[3] E. I. George, "The variable selection problem," *Journal of the American Statistical Association*, vol. 95, no. 452, pp. 1304–1308, 2000.

[4] R. R. Hocking, "The analysis and selection of variables in linear regression," *Biometrics*, vol. 32, no. 1, pp. 1–49, 1976.

[5] A. E. Goodenough, A. G. Hart, and R. Stafford, "Regression with empirical variable selection: description of a new method and application to ecological datasets," *PLoS ONE*, vol. 7, no. 3, Article ID e34338, 2012.

[6] K. Murray, S. Heritier, and S. Müller, "Graphical tools for model selection in generalized linear models," *Statistics in Medicine*, vol. 32, no. 25, pp. 4438–4451, 2013.

[7] W. Sauerbrei, P. Royston, and H. Binder, "Selection of important variables and determination of functional form for continuous predictors in multivariable model building," *Statistics in Medicine*, vol. 26, no. 30, pp. 5512–5528, 2007.

[8] F. G. Blanchet, P. Legendre, and D. Borcard, "Forward selection of explanatory variables," *Ecology*, vol. 89, no. 9, pp. 2623–2632, 2008.

[9] R. E. Wiegand, "Performance of using multiple stepwise algorithms for variable selection," *Statistics in Medicine*, vol. 29, no. 15, pp. 1647–1659, 2010.

[10] M. Hofmann, C. Gatu, and E. J. Kontoghiorghes, "Efficient algorithms for computing the best subset regression models for large-scale problems," *Computational Statistics & Data Analysis*, vol. 52, no. 1, pp. 16–29, 2007.

[11] M. G. G'Sell, T. Hastie, and R. Tibshirani, *False Variable Selection Rates in Regression*, 2013.

[12] H. S. Rabie and I. W. Saunders, "A simulation study to assess a variable selection method for selecting single nucleotide polymorphisms associated with disease," *Journal of Computational Biology*, vol. 19, no. 10, pp. 1151–1161, 2012.

[13] Q. He and D.-Y. Lin, "A variable selection method for genome-wide association studies," *Bioinformatics*, vol. 27, no. 1, pp. 1–8, 2011.

[14] B. C. Wallet, D. J. Marchette, J. L. Solka, and E. J. Wegman, "A genetic algorithm for best subset selection in linear regression," in *Proceedings of the 28th Symposium on the Interface*, Sydney, Australia, 1996.

[15] A. J. Miller, *Subset Selection in Regression*, Chapman and Hall/CRC, 2nd edition, 2002.

[16] R. B. O'Hara and M. J. Sillanpää, "A review of Bayesian variable selection methods: what, how and which," *Bayesian Analysis*, vol. 4, no. 1, pp. 85–118, 2009.

[17] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B: Methodological*, vol. 58, no. 1, pp. 267–288, 1996.

[18] R. B. Bendel and A. A. Afifi, "Comparison of stopping rules in forward 'Stepwise' regression," *Journal of the American Statistical Association*, vol. 72, no. 357, pp. 46–54, 1977.

[19] A. Liebminger, L. Seyfang, P. Filzmoser, and K. Varmuza, "A new variable selection method based on all subsets regression," in *Proceedings of the Scandinavian Symposium on Chemometrics*, Lappeenranta, Finland, June 2007.

[20] B. Mothe, J. Ibarrondo, A. Llano, and C. Brander, "Virological, immune and host genetics markers in the control of HIV infection," *Disease Markers*, vol. 27, no. 3, pp. 105–120, 2009.

[21] N. Frahm, T. Korber, C. M. Adams et al., "Consistent cytotoxic-T-lymphocyte targeting of immunodominant regions in human immunodeficiency virus across multiple ethnicities," *Journal of Virology*, vol. 78, no. 5, pp. 2187–2200, 2004.

[22] R. Draenert, M. Altfeld, C. Brander et al., "Comparison of overlapping peptide sets for detection of antiviral CD8 and CD4 T cell responses," *Journal of Immunological Methods*, vol. 275, no. 1-2, pp. 19–29, 2003.

[23] T. Wicker, E. Schlagenhauf, A. Graner, T. J. Close, B. Keller, and N. Stein, "454 sequencing put to the test using the complex genome of barley," *BMC Genomics*, vol. 7, article 275, 2006.

[24] J. G. Prado, J. Carrillo, J. Blanco-Heredia, and C. Brander, "Immune correlates of HIV control," *Current Medicinal Chemistry*, vol. 18, no. 26, pp. 3963–3970, 2011.

[25] B. Mothe, A. Llano, J. Ibarrondo et al., "Definition of the viral targets of protective HIV-1-specific T cell responses," *Journal of Translational Medicine*, vol. 9, article 208, 2011.

[26] C. L. Mallows, "Some comments on C_p," *Technometrics*, vol. 15, no. 4, pp. 661–675, 1973.

[27] H. Akaike, "Information theory and an extension of the maxiumum likelihood principle," in *Proceedings of the 2nd*

International Symposium on Information Theory, Budapest, Hungary, 1973.

[28] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[29] E. Gayawan and R. A. Ipinyomi, "A comparison of Akaike, Schwarz and R square criteria for model selection using some fertility models," *Australian Journal of Basic and Applied Sciences*, vol. 3, no. 4, pp. 3524–3530, 2009.

[30] J. Kuha, "AIC and BIC: comparisons of assumptions and performance," *Sociological Methods and Research*, vol. 33, no. 2, pp. 188–229, 2004.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 810514, 9 pages
http://dx.doi.org/10.1155/2015/810514

Research Article

Prediction of MicroRNA-Disease Associations Based on Social Network Analysis Methods

Quan Zou,[1,2] Jinjin Li,[1] Qingqi Hong,[3] Ziyu Lin,[1] Yun Wu,[4] Hua Shi,[4] and Ying Ju[1]

[1]*School of Information Science and Technology, Xiamen University, Xiamen 361005, China*
[2]*School of Computer Science and Technology, Tianjin University, Tianjin 300072, China*
[3]*Software School, Xiamen University, Xiamen 361005, China*
[4]*College of Computer and Information Engineering, Xiamen University of Technology, Xiamen 361024, China*

Correspondence should be addressed to Qingqi Hong; hongqq@gmail.com

Received 29 December 2014; Revised 9 March 2015; Accepted 16 March 2015

Academic Editor: Xiao Chang

MicroRNAs constitute an important class of noncoding, single-stranded, ~22 nucleotide long RNA molecules encoded by endogenous genes. They play an important role in regulating gene transcription and the regulation of normal development. MicroRNAs can be associated with disease; however, only a few microRNA-disease associations have been confirmed by traditional experimental approaches. We introduce two methods to predict microRNA-disease association. The first method, KATZ, focuses on integrating the social network analysis method with machine learning and is based on networks derived from known microRNA-disease associations, disease-disease associations, and microRNA-microRNA associations. The other method, CATAPULT, is a supervised machine learning method. We applied the two methods to 242 known microRNA-disease associations and evaluated their performance using leave-one-out cross-validation and 3-fold cross-validation. Experiments proved that our methods outperformed the state-of-the-art methods.

1. Introduction

MicroRNAs constitute a class of non-protein-coding small RNAs, 20 to 25 nucleotides long, that bind to the 3′ untranslated region of target mRNAs to regulate mRNA turnover and translation. There are many biological processes, which are regulated by microRNAs, such as development, differentiation, apoptosis, and diseases [1–3]. Many studies have found that microRNAs play an important role in cellular signaling networks [4], tissue development, [5–7] and cell growth [8]. They are also associated with various diseases [9, 10], including breast cancer [11, 12], lung cancer [13, 14], cardiomyopathy [15], and cell lymphoma [16]. If the microRNA abnormality causes the disease, the abnormal microRNA and the disease are associated by the causal relationship. And the microRNA-disease association is what we aim to predict. Predicting microRNA-disease associations has emerged as an important strategy in understanding disease mechanisms [17]. For example, dysregulation of microRNAs can affect apoptosis signaling pathways and cell cycle regulation in cancer [18].

The importance of microRNA-disease association prediction has been appreciated for some time [19]. However, most of the techniques that have been developed to achieve this suffer several inherent weaknesses; in particular, traditional experimental approaches are time-consuming and expensive. It is necessary to employ the bioinformatics analysis, which could make use of databases and the potential inferences. For bioinformatics approaches, it is important to measure the functional similarities among microRNAs in order to construct networks based on functional similarity [20–24]. The construction of functional similarity networks for genes encoding proteins has produced significant results [25–32]; however, the methods used to analyze protein-encoding genes are not always adaptable to enable use with microRNAs because the correlation between the functional similarities of genes and gene sequences or expression similarities may not exist for microRNAs [5, 6, 33, 34]. MicroRNAs directly

adjust the one-third of the human genes. The genes targeted by miRNAs identified are recognized from directed biological process. However, the previous published methods to find gene used bio-experiment or the characteristics of protein sequence. However, gene and miRNA identification is quite inefficient. Another issue is that there are not many validated associations between microRNAs and diseases. For studying microRNA-disease association, there are two well-known databases: the human microRNA-associated disease database (HMDD) and the miR2Disease database of differentially expressed MiRNAs in human cancers (dbDEMC). The data in HMDD and dbDEMC are manually collected and archived from publications [10, 21, 22, 35]. The last main challenge is that it is difficult to select negative samples as there are no verified negative microRNA-disease associations. It is the refore difficult to conduct biological experiments without such controls. Hence, it is necessary to develop effective computational methods to detect potential microRNA-disease associations.

To overcome the above challenges and to effectively predict associations, we explored the computational method KATZ [36] and the machine learning method CATAPULT [5, 6] to predict microRNA-disease associations. The two methods can succeed to overcome the challenges above. The highlight work is to discover unknown associations through known associations, including microRNA-microRNA associations, a small quantity of microRNA-disease associations, and disease-disease associations. Previous studies show that one or more mutations from the same functional module can give rise to diseases with overlapping clinical features [1, 37–39]. Biological experiments of human disease show that microRNAs causing similar diseases often interact with each other directly or indirectly [40–45]. Hence, we learn from the idea of social network. This is an integrated network composed of microRNA-microRNA association networks, known microRNA-disease association networks, and disease-disease association networks and is similar to social networks used to predict the relationship between two individuals [40, 46–49]. In this paper, we take full advantage of relationships among microRNAs and diseases to predict the association between microRNA and disease. Each predicted microRNA-disease association is denoted by a score. For each disease, we rank the microRNA on the basis of a score. For a disease, if a microRNA is ranked in the top k, the microRNA is expected to have a high probability of association with the disease [50, 51]. We show that KATZ and CATAPULT are superior to current methods by cross-validation. KATZ and CATAPULT are able to propose many potential associations, which is of great value for future studies.

2. Datasets

We used three types of data, microRNA-microRNA association, microRNA-disease association, and disease-disease association data. The microRNA-microRNA association dataset includes 271 microRNAs, and the association is denoted by a functional similarity score. The dataset was downloaded from http://www.webcitation.org/query.php [5, 6]. The disease-disease association dataset, including 5080 diseases, was downloaded from MimMiner [52], which provides a similarity score for each phenotype pair by text mining analysis of their phenotype descriptions in the Online Mendelian Inheritance in Man (OMIM) database [53]. The disease-disease similarity scores have been successfully used to predict or prioritize disease related genes [54, 55]. The microRNA-disease association dataset contained 271 microRNAs and 5080 diseases. Furthermore, there are 242 microRNA-disease associations. It means there are 242 nonzero elements in the matrix of microRNA-disease association. The microRNA-disease association dataset was downloaded from [56]. In addition, we verified that the 242 nonzero elements consisted of 99 microRNAs and 51 diseases. The details of the datasets are shown in Table 1.

Table 1: Distribution of the three datasets.

Dataset	Matrix	Similarity score >0
MicroRNA-microRNA association dataset	271 × 271	56289
Disease-disease association dataset	5080 × 5080	20285172
MicroRNA-disease association dataset	271 × 5080	242

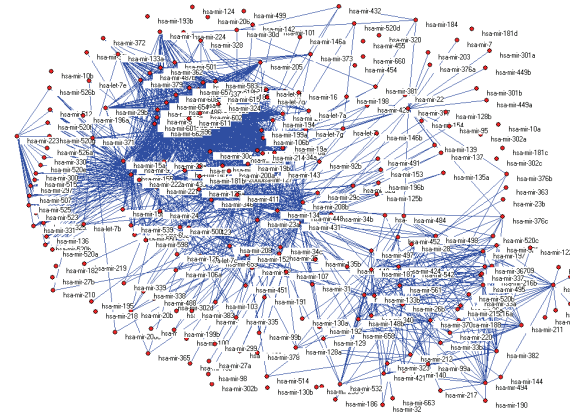


Figure 1: Bipartite graph of the microRNA-disease association network.

With the above datasets, we could construct a microRNA-microRNA network, a disease-disease network, and a microRNA-disease network using a bipartite graph. For example, Figure 1 denotes the bipartite graph of the microRNA-disease network. In the graph, the nodes denote microRNAs or diseases and the lines correspond to associations between microRNAs and diseases. If there is an association between a microRNA and a disease, there must be a line between the microRNA and the disease.

The degree distributions of microRNAs and diseases in the bipartite graph of the microRNA-disease association network are illustrated in Figure 2. The microRNA degree is defined as the number of diseases that connect with

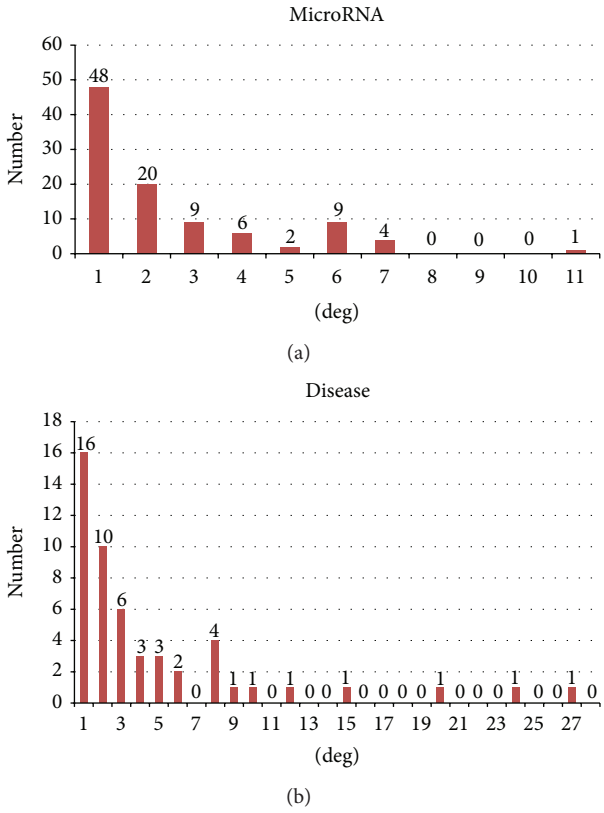


Figure 2: Degree distributions of microRNAs and diseases in the bipartite graph of the microRNA-disease association network.

Table 2: Statistical data for the bipartite graph of the microRNA-disease association network.

Title	Number
MicroRNAs	271
Diseases	5080
Known-associating microRNAs	99
Known-associating diseases	51
Known-associations	242
Average number of microRNA degrees	2.44
Average number of disease degrees	4.75

a microRNA. In the same way, the disease degree is defined as the number of microRNAs that connect with a disease. The node degree can show the activeness or status of the node (microRNA or disease) in the entire network.

We propose to compare our methods with the previously described microRNA-based similarity inference (MBSI), phenotype-based similarity inference (PBSI), and network-consistency-based inference (NetCBI) methods [55]. Hence, we used the same datasets as them and we present Table 2 to clearly describe the statistical data for the bipartite graph of the microRNA-disease association network. Table 2 illustrates that there are few known microRNA-disease associations for a disease. For example, it should have $271 * 5080$ microRNA-disease associations, but known associations are only 242.

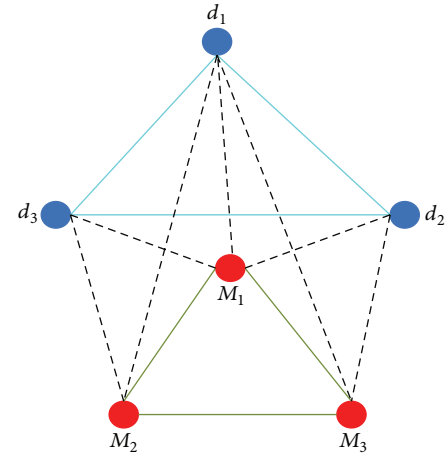


Figure 3: Unweighted, undirected graph.

3. Methods and Algorithm

We introduce two different computational methods, which were presented by [36] to predict microRNA-disease associations. The first method, KATZ [57], has been shown to be successful at predicting links in a social network. When KATZ is applied to predict microRNA-disease associations, it uses the functional similarity score to denote the associations. KATZ computes the similarity score based on walks of different lengths between the microRNA and disease nodes. The second method, CATAPULT, is a supervised learning method. For the supervised learning method, features must be offered that are derived from hybrid walks through the microRNA-disease association network. However, CATAPULT is a transformation of a general supervised learning method. For the problem of microRNA-disease association, there are only positive examples and unlabeled examples, which CATAPULT is able to overcome. Algorithm part will detailedly present KATZ and CATAPULT.

3.1. KATZ. KATZ is similar to classical approaches, such as random walk [58], Prince [59], and CIPHER [60]. The essence of these approaches is a ranking algorithm. For example, the KATZ method computes the functional similarity score for microRNA-disease node pairs based on the microRNA-disease association network and ranking the diseases for a microRNA on the basis of the functional similarity score [57]. KATZ was successfully applied to predict social associations based on a social network [60]. Predicting microRNA-disease associations on the basis of a microRNA-disease association network is equivalent to predicting associations in a social network. KATZ results show that it can also adapt to predict associations between microRNAs and diseases.

For the known associations between microRNAs and diseases, we constructed an unweighted, undirected graph and derived a corresponding adjacency matrix of the graph. To vividly describe the method, we illustrate a simple unweighted, undirected graph, in Figure 3. Suppose the corresponding adjacency matrix of Figure 1 is A; the adjacency

matrix A can be written with $A_{ij} = 1$, if microRNA node i and disease node j are connected, and $A_{ij} = 0$, if there is no line between microRNA node i and disease node j. However, there are not many direct lines linking microRNA and disease; therefore, it is difficult to denote the microRNA-disease association through the adjacency matrix A. Thus, we counted the number of walks of different lengths, which link microRNA node i and disease node j to signify the association between microRNA and disease. $(A^l)_{ij}$ denotes the number of walks of length l that link node i and node j.

Next, we integrated different walks of different length to obtain a comprehensive association measure. We introduced a nonnegative coefficient β_l, whose function is to control the contribution of different length walks. If $l1$ is larger than $l2$, β_{l1} is smaller than β_{l2}. Suppose microRNA node i and disease node j are not connected in the unweighted, undirected graph; then $A_{ij} = 0$ and the microRNA i and disease j association can be computed through

$$S(A)_{ij} = \sum_{l=1}^{k} \beta^l \left(A^l\right)_{ij}. \tag{1}$$

From formula (1), we can draw the conclusion that higher order paths contribute much less to microRNA-disease association. Formula (2) can process the entire unweighted, undirected graph:

$$S = \sum_{l=1}^{k} \beta_l A^l, \tag{2}$$

where if $l \to \infty$, $\beta_l \to 0$. In KATZ, if β_l is replaced by β^l, KATZ can be written as

$$S^{\text{katz}} = \sum_{l\geq 1} \beta^l A^l = (I - \beta A)^{-1} - I, \tag{3}$$

where β is chosen on the basis of $\beta < 1/\|A\|^2$. For the choice of value k, the sum over infinitely many path lengths is not necessarily considered. According to the experimental results, small values of k ($k = 3$ or $k = 4$) obtain good performance in the task of recommending linked nodes. We have carried out the experiments for the other values of k. When $k < 3$, the experimental results are worse. However, for $k > 4$, the results are no better than $k = 3$ or 4. In addition, when $k > 4$ or bigger, the experimental time is much longer.

To use KATZ, we need a microRNA-disease association adjacent matrix A, which is the adjacent matrix of the microRNA-disease association network and is denoted as follows:

$$A = \begin{bmatrix} G_{\text{MM}} & G_{\text{MD}} \\ G_{\text{MD}}^T & G_{\text{DD}} \end{bmatrix}, \tag{4}$$

where G_{MM} is the adjacent matrix of the microRNA-microRNA association network, G_{MD} is the adjacent matrix of the microRNA-disease association network, and G_{DD} is the adjacent matrix of the disease-disease association network. We substituted the adjacent matrix A into formula (3) to obtain the association score matrix of microRNAs and diseases.

Setting $k = 3$, the correlation score matrix $S^{\text{KATZ}}(A)$ denoting the association between microRNAs and diseases can be written as expression (5). Here we use KATZ with $k = 3$ to obtain the correlation score matrix. Consider

$$\begin{aligned} S^{\text{Katz}}(A) = {} & \beta G_{\text{MD}} + \beta^2 \left(G_{\text{MM}} G_{\text{MD}} + G_{\text{MD}} G_{\text{DD}}\right) \\ & + \beta^3 \left(G_{\text{MD}} G_{\text{MD}}^T G_{\text{MD}} + G_{\text{MM}}^2 G_{\text{MD}}\right. \\ & \left. + G_{\text{MM}} G_{\text{MD}} G_{\text{DD}} + G_{\text{MD}} G_{\text{DD}}^2\right). \end{aligned} \tag{5}$$

One of the advantages of KATZ is that it can study human microRNA-disease association and association for other species. In KATZ, this is achieved simply by changing the submatrix of adjacent matrix A, denoted as

$$\begin{aligned} G_{\text{MD}} &= \left[G_{\text{HS}} \;\; G_{\text{S}}\right], \\ G_{\text{DD}} &= \begin{bmatrix} D_{\text{PHS}} & 0 \\ 0 & D_{\text{PS}} \end{bmatrix}, \end{aligned} \tag{6}$$

where D_{PHS} and D_{PS} represent human disease and disease of other species, respectively. G_{HS} and G_{S} are microRNA-disease association of human and other species, respectively. When we conduct an experiment on human, set $D_{\text{PS}} = 0$ and $G_{\text{S}} = 0$.

3.2. CATAPULT. CATAPULT is a supervised learning method. General supervised learning methods need positive examples and negative examples. However, for microRNA-disease association, there is a lack of negative examples. Positive associations can be checked through existing methods, but there is not a method to prove negative associations. Because negative associations are seldom proven, we processed the problem by treating all nonpositive association node pairs as unlabeled because previous studies have shown that most unlabeled pairs have a negative association [55].

A study by Mordelet and Vert [61] used the bagging technique to obtain an aggregate classifier based on positive examples and unlabeled examples. CATAPULT uses a biased support vector machine (SVM) to classify microRNA-disease pairs. Hence, CATAPULT uses a bagging algorithm to train biased SVM. In CATAPULT, unlabeled samples are randomly selected from the set of all unlabeled examples and a classifier is used to train the selected unlabeled samples as negative examples and positive examples. The features of microRNA-disease pairs are obtained from hybrid walks through the heterogeneous network. To some extent, bagging could reduce the variance in the classifier. The variance is caused by randomly selecting negative examples. R is the set of randomly selected negative microRNA-disease pairs and $N-$ is the number of set R. T is the set of positive microRNA-disease pairs and $N+$ is the number of set T. U denotes all the unlabeled microRNA-disease pairs. The biased SVM means that we assign a penalty, $k-$, for false positives and a larger penalty, $k+$, for false negatives. Detail of the CATAPULT algorithm is displayed in the following part. To train a biased SVM, CATAPULT uses formula (7) based on the known

positive examples T and randomly selected negative examples R to obtain a biased SVM. ξ_i denotes the distance of example i from a boundary and SVM gives the example i corresponding penalty. $\langle \theta_t, \Phi(x) \rangle$ denotes the function score for iteration t, where θ_t is the normal to the hyper plane at the tth iteration and $\Phi(x)$ is the feature vector of example x. Besides, the feature vector of example x is the feature vector of the microRNA-disease pair. In our experiment, we assign 1 to $k-$ and 30 to $N-$ [36].

CATAPULT Algorithm.

INIT
For $t = 1, 2, 3, \ldots, N-$:

(1) Select the set R of size $N-$ from U as negative examples.

(2) Train a classifier based on positive examples T and negative examples

$$\begin{aligned} \min_{\theta' \in \Re^d} \quad & \frac{1}{2} \left\| \theta' \right\|^2 + k_- \sum_{i \in R} \xi_i + k_+ \sum_{i \in T} \xi_i \\ \text{subject to } & \xi_i \geq 0, \quad \forall i \in R \cup T, \\ & \left\langle \Phi(x_i), \theta' \right\rangle \geq 1 - \xi_i, \quad \forall i \in T, \\ & -\left\langle \Phi(x_i), \theta' \right\rangle \geq 1 - \xi_i, \quad \forall i \in R. \end{aligned} \tag{7}$$

(3) For any *update*:

$$\begin{aligned} nx &\longleftarrow nx + 1 \\ s(x) &\longleftarrow s(x) + \langle \theta_t, \Phi(x) \rangle \\ \text{return } s(x) &\longleftarrow \frac{s(x)}{n(x)}, \quad \forall x \in U. \end{aligned} \tag{8}$$

4. Implementation

4.1. Results. The KATZ and CATAPULT methods were applied to the 242 known microRNA-disease associations to infer potential microRNA-disease associations. First, we mainly verified microRNA-disease associations. The set of 242 known microRNA-disease associations is regarded as the "gold standard" data and was used to evaluate the performance of KATZ and CATAPULT methods in the leave-one-out and 3-fold cross-validation experiment and training dataset in the comprehensive prediction [62]. To compare our methods with MBSI, PBSI, and NetCBI, we carried out leave-one-out cross-validation on microRNA-disease associations using KATZ and CATAPULT methods. Furthermore, we carried out the 3-fold cross-validation to make sure that the outperformance of KATZ and CATAPULT is solid. For the leave-one-out cross-validation, each of the 242 known microRNA-disease associations is left out once in turn as the testing case. For the 3-fold cross-validation, the dataset containing 242 known microRNA-disease associations is divided into three parts, which is turned to act as testing. We ranked all microRNA-disease associations according to the scores obtained from KATZ and CATAPULT results.

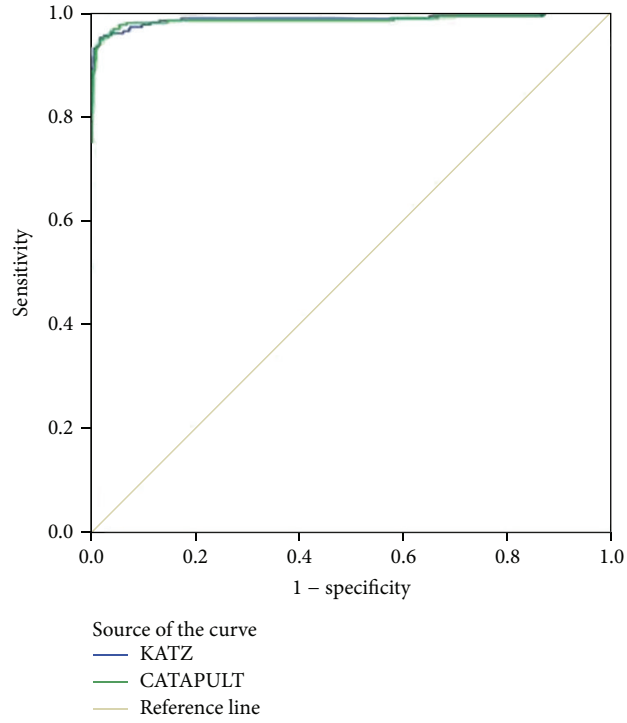


Figure 4: ROC curves of KATZ and CATAPULT methods by leave-one-out cross-validation.

We used a receiver operating characteristic (ROC) curve to evaluate the effect of the method. Varying the threshold plots a ROC curve, and the numeric representation of a ROC curve is the area under the curve (AUC). If we could not compare which method was best from the ROC curve, we could compare the AUC. In the experiment of leave-one-out cross-validation, KATZ and CATAPULT were tested on the 242 known microRNA-disease associations and AUC values 98.9% and 98.8% for KATZ and CATAPULT were achieved. Figure 4 is the corresponding ROC curve of KATZ and CATAPULT methods. This indicates that our methods have great potential to infer new microRNA-associations.

For the leave-one-out cross-validation, we carry out one loop for each known microRNA-disease association. In each loop, we hide a microRNA-disease association in the known association group and run KATZ and CATAPULT methods on the remaining associations repeating 242 times to ensure that each known microRNA-disease association is hidden exactly once. In each loop, we order the 5080 diseases for the microRNAs, which is the hidden association. We rule that if the disease that is the hidden association has the highest k value, then prediction is true. The principle behind this rule is that the method is better if it can predict the true microRNA-disease association with higher probability. Table 3 shows the distribution of diseases on the basis of the number of microRNAs. Figure 6 presents the result of prediction hidden microRNA-disease associations. The x-axis is the threshold

Table 3: Distribution of diseases on the basis of microRNAs.

Number of microRNAs	0	1	2	3	4	5	6	8	9	10	12	15	20	24	27
Number of diseases	5029	16	10	6	3	3	2	4	1	1	1	1	1	1	1

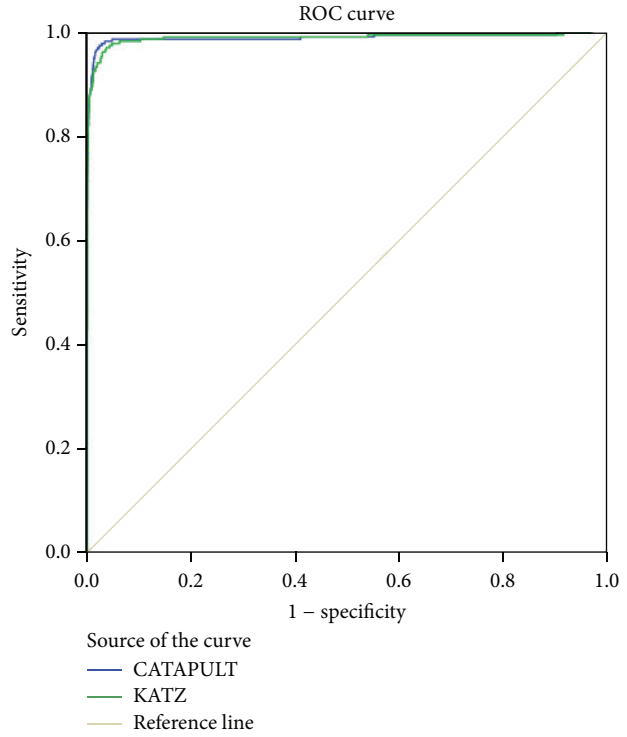


Figure 5: ROC curves of KATZ and CATAPULT methods by 3-fold cross-validation.

k and the y-axis is the amount of true prediction. Figure 6 shows the results for KATZ and CATAPULT.

In the experiment of 3-fold cross-validation, KATZ and CATAPULT were tested on the 242 known microRNA-disease associations and AUC values 98.4% and 98.3% for KATZ and CATAPULT were achieved. Figure 5 shows AUC values of KATZ and CATAPULT methods. The cross-validation results prove that the outperformance is solid.

4.2. Evaluation. To confirm the strength of our methods, we compared them with MBSI, PBSI, and NetCBI. MBSI and PBSI both work on the basis of recommendation. However, MBSI takes full advantage of microRNAs similarity. This means that if association between a microRNA and a disease has been validated, then other similar microRNAs would be recommended to the disease. The drawback of MBSI is that it overlooks disease-disease associations. In contrast, PBSI take full advantage of disease similarities but overlooks the microRNA-microRNA associations. NetCBI considers both associations. The basic idea of NetCBI is ranking. Suppose a microRNA and a disease are linked; if a microRNA is ranked top by querying the microRNAs and a disease is ranked top by querying the diseases, then it rules that associations exist between top-ranking microRNAs and top-ranking diseases.

Table 4: Comparison of different prediction methods based on AUC values.

Method	MBSI	PBSI	NetCBI	KATZ	CATAPULT
AUC	74.83%	54.02%	80.66%	98.9%	98.8%

Table 5: Top 10 newly predicted microRNA-disease associations by KATZ.

Rank	MicroRNA	OMIM disease ID	Disease	Source
1	hsa-let-7i	211980	Lung cancer	HMDD
2	hsa-let-7d	114480	Breast cancer	HMDD
3	hsa-mir-145	211980	Lung cancer	HMDD
4	hsa-mir-18a	114480	Breast cancer	HMDD
5	hsa-mir-145	114480	Breast cancer	HMDD
6	hsa-mir-106b	114480	Breast cancer	HMDD
7	hsa-let-7e	114480	Breast cancer	HMDD
8	hsa-let-7b	114480	Breast cancer	HMDD
9	hsa-mir-19a	114480	Breast cancer	HMDD
10	hsa-mir-125a	114480	Breast cancer	HMDD

We used leave-one-out cross-validation to compare our methods with previous methods based on the same datasets. Table 4 shows the comparative results and our methods are clearly better at predicting microRNA-disease associations than the other methods. The assessment criteria that we used were ROC and AUC. AUC and ROC are the measure of the standard classifier model which is good or bad. ROC presents the evaluation criteria in a visual form, and the AUC value is the area under the ROC curve. Our methods yield 98.9% and 98.8%, which are better than MBSI (74.83%), PBSI (54.02%), and NetCBI (80.66%).

We verify the top 10 predicted associations, which were not identified in our microRNA-disease association dataset. However, the latest online databases provide the evidence. The online databases that we referenced were OMIM, HMDD, and miR2Disease. Tables 5 and 6 show the prediction results by KATZ and CATAPULT. Each predicted association is confirmed by one of the three databases.

5. Conclusions

Identifying microRNA-disease associations is an important part of understanding disease mechanisms. Although experimental methods can identity microRNA-disease associations, they are time-consuming and expensive. Hence, efficient methods to identity microRNA-disease associations are desired.

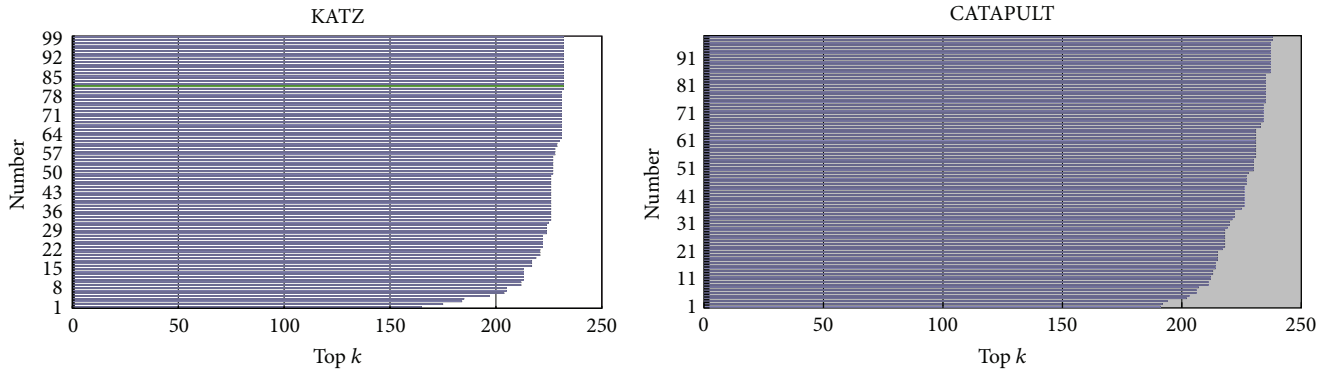


Figure 6: Recovery of microRNA-disease associations with respect to disease rank under leave-one-out cross-validation.

Table 6: Top 10 newly predicted microRNA-disease associations by CATAPULT.

Rank	MicroRNA	OMIM disease ID	Disease	Source
1	hsa-let-7a	176807	Prostate cancer	miR2Disease
2	hsa-mir-34a	114480	Breast cancer	HMDD
3	hsa-mir-21	211980	Lung cancer	HMDD
4	hsa-let-7c	114480	Breast cancer	HMDD
5	hsa-mir-19a	114480	Breast cancer	HMDD
6	hsa-let-7a	151400	Chronic lymphocytic leukemia	miR2Disease
7	hsa-mir-29b	114480	Breast cancer	miR2Disease
8	hsa-mir-146a	211980	Lung cancer	HMDD
9	hsa-mir-155	211980	Lung cancer	HMDD
10	hsa-let-7c	114550	Hepatocellular carcinoma	miR2Disease

We introduce KATZ and CATAPULT methods for predicting microRNA-disease associations. KATZ succeeds in processing social network links to achieve prediction, which is a different strategy to other methods, such as PBSI and MBSI. The KATZ method uses the entire heterogeneous network, including microRNA-microRNA association, microRNA-disease association, and disease-disease association networks. CATAPULT is a supervised learning method and uses a biased SVM. KATZ and CATAPULT significantly outperform other prediction microRNA-disease association methods, assessed by the leave-one-out and 3-fold cross-validation evaluation strategy. The potential microRNA-disease association predicted by KATZ and CATAPULT will facilitate biological experiments, which identify the true associations between microRNAs and diseases. The KATZ uses the simple measure on the heterogeneous network to predict the potential microRNA-disease associations. KATZ's performance is relatively poor on the sparse known associations.

Although our methods perform well, better methods would be proposed to predict microRNA-disease associations. There are many features of microRNAs and diseases that are not used to help predict microRNA-disease associations, such as gene ontology and the external manifestations of disease. With the use of more factors in prediction methods and the emergence of new relevant data, the prediction of microRNA-disease association will further advance. Ultimately this will help the medical treatment of disease.

Conflict of Interests

The authors declare that they have no conflict of interests.

Authors' Contribution

Quan Zou analyzed data and designed the project and coordinated it. Jinjin Li created the front end user interface and developed the web server. Yun Wu and Hua Shi were involved in drafting the paper. Ying Ju had given final approval of the version to be published. Qingqi Hong helped revise the paper and gave helpful suggestion. All authors read and approved the final paper.

Acknowledgments

The work was supported by the Natural Science Foundation of China (nos. 61370010, 61202011, and 61303004), the Natural Science Foundation of Fujian Province of China (no. 2014J01253, no. 2013J05103), and the Open Fund of Shanghai Key Laboratory of Intelligent Information Processing, China (no. IIPL-2014-004).

References

[1] P. Jiménez, F. Thomas, and C. Torras, "3D collision detection: A survey," *Computers & Graphics*, vol. 25, no. 2, pp. 269–285, 2001.

[2] T. Uziel, F. V. Karginov, S. Xie et al., "The miR-17~92 cluster collaborates with the Sonic Hedgehog pathway in medulloblastoma," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 8, pp. 2812–2817, 2009.

[3] H. Ding, P. Feng, W. Chen, and H. Lin, "Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis," *Molecular BioSystems*, vol. 10, no. 8, pp. 2229–2235, 2014.

[4] Q. Cui, Z. Yu, E. O. Purisima, and E. Wang, "Principles of microRNA regulation of a human cellular signaling network," *Molecular Systems Biology*, vol. 2, article 46, 2006.

[5] D. Wang, J. Wang, M. Lu, F. Song, and Q. Cui, "Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases," *Bioinformatics*, vol. 26, no. 13, pp. 1644–1650, 2010.

[6] Y. Wang, T. Cui, C. Zhang et al., "Global protein-protein interaction network in the human pathogen mycobacterium tuberculosis H37Rv," *Journal of Proteome Research*, vol. 9, no. 12, pp. 6665–6677, 2010.

[7] Y. Wang, L. Chen, B. Chen et al., "Mammalian ncRNA-disease repository: a global view of ncRNA-mediated disease network," *Cell Death & Disease*, vol. 4, article e765, 2013.

[8] A. Esquela-Kerscher and F. J. Slack, "Oncomirs—microRNAs with a role in cancer," *Nature Reviews Cancer*, vol. 6, no. 4, pp. 259–269, 2006.

[9] M. V. G. Latronico, D. Catalucci, and G. Condorelli, "Emerging role of microRNAs in cardiovascular biology," *Circulation Research*, vol. 101, no. 12, pp. 1225–1236, 2007.

[10] M. Lu, Q. Zhang, M. Deng et al., "An analysis of human microRNA and disease associations," *PLoS ONE*, vol. 3, no. 10, Article ID e3420, 2008.

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

[12] 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.

[13] A. Esquela-Kerscher, P. Trang, J. F. Wiggins et al., "The let-7 microRNA reduces tumor growth in mouse models of lung cancer," *Cell Cycle*, vol. 7, no. 6, pp. 759–764, 2008.

[14] Q. Wang, L. Wei, X. Guan, Y. Wu, Q. Zou, and Z. Ji, "Briefing in family characteristics of microRNAs and their applications in cancer research," *Biochimica et Biophysica Acta—Proteins and Proteomics*, vol. 1844, no. 1, pp. 191–197, 2014.

[15] B. Yang, H. Lin, J. Xiao et al., "The muscle-specific microRNA miR-1 regulates cardiac arrhythmogenic potential by targeting GJA1 and KCNJ2," *Nature Medicine*, vol. 13, no. 4, pp. 486–491, 2007.

[16] R. W. Chen, L. T. Bemis, C. M. Amato et al., "Truncation in CCND1 mRNA alters miR-16-1 regulation in mantle cell lymphoma," *Blood*, vol. 112, no. 3, pp. 822–829, 2008.

[17] J. Xu, C.-X. Li, Y.-S. Li et al., "MiRNA-miRNA synergistic network: construction via co-regulating functional modules and disease miRNA topological features," *Nucleic Acids Research*, vol. 39, no. 3, pp. 825–836, 2011.

[18] T.-H. Cheung, K.-N. M. Man, M.-Y. Yu et al., "Dysregulated microRNAs in the pathogenesis and progression of cervical neoplasm," *Cell Cycle*, vol. 11, no. 15, pp. 2876–2884, 2012.

[19] K. Han, P. Xuan, J. Ding, Z. J. Zhao, L. Hui, and Y. L. Zhong, "Prediction of disease-related microRNAs by incorporating functional similarity and common association information," *Genetics and Molecular Research*, vol. 13, no. 1, pp. 2009–2019, 2014.

[20] T. Huang and Y.-D. Cai, "An information-theoretic machine learning approach to expression QTL analysis," *PLoS ONE*, vol. 8, no. 6, Article ID e67899, 2013.

[21] Q. Zou, X. Li, W. Jiang, Z. Lin, G. Li, and K. Chen, "Survey of MapReduce frame operation in bioinformatics," *Briefings in Bioinformatics*, vol. 15, no. 4, pp. 637–647, 2014.

[22] Q. Zou, J. Li, C. Wang, and X. Zeng, "Approaches for recognizing disease genes based on network," *BioMed Research International*, vol. 2014, Article ID 416323, 10 pages, 2014.

[23] G. Yu, C.-L. Xiao, X. Bo et al., "A new method for measuring functional similarity of microRNAs," *Journal of Integrated OMICS*, vol. 1, no. 1, pp. 49–54, 2010.

[24] B. Liu, J. Xu, S. Fan, R. Xu, J. Zhou, and X. Wang, "PseDNA-Pro: DNA-binding protein identification by combining Chou's PseAAC and physicochemical distance transformation," *Molecular Informatics*, vol. 34, no. 1, pp. 8–17, 2015.

[25] B. Liu, X. Wang, Q. Zou, Q. Dong, and Q. Chen, "Protein remote homology detection by combining chou's pseudo amino acid composition and profile-based protein representation," *Molecular Informatics*, vol. 32, no. 9-10, pp. 775–782, 2013.

[26] Z. Du, L. Li, C.-F. Chen, P. S. Yu, and J. Z. Wang, "G-SESAME: web tools for GO-term-based gene similarity analysis and knowledge discovery," *Nucleic Acids Research*, vol. 37, no. 2, pp. W345–W349, 2009.

[27] S. G. Lee, J. U. Hur, and Y. S. Kim, "A graph-theoretic modeling on GO space for biological interpretation of gene clusters," *Bioinformatics*, vol. 20, no. 3, pp. 381–388, 2004.

[28] K. Li, D. Wu, X. Chen et al., "Current and emerging biomarkers of cell death in human disease," *BioMed Research International*, vol. 2014, Article ID 690103, 10 pages, 2014.

[29] J. Lin, C. M. Gan, X. Zhang et al., "A multidimensional analysis of genes mutated in breast and colorectal cancers," *Genome Research*, vol. 17, no. 9, pp. 1304–1318, 2007.

[30] P. W. Lord, R. D. Stevens, A. Brass, and C. A. Goble, "Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation," *Bioinformatics*, vol. 19, no. 10, pp. 1275–1283, 2003.

[31] C. Pesquita, D. Faria, H. Bastos, A. E. N. Ferreira, A. O. Falcão, and F. M. Couto, "Metrics for GO based protein semantic similarity: a systematic evaluation," *BMC Bioinformatics*, vol. 9, article S4, 2008.

[32] J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu, and C.-F. Chen, "A new method to measure the semantic similarity of GO terms," *Bioinformatics*, vol. 23, no. 10, pp. 1274–1281, 2007.

[33] W. Li, E. Deng, H. Ding, W. Chen, and H. Lin, "iORI-PseKNC: a predictor for identifying origin of replication with pseudo k-tuple nucleotide composition," *Chemometrics and Intelligent Laboratory Systems*, vol. 141, pp. 100–106, 2015.

[34] Y. Li, L. Zhuang, Y. Wang et al., "Connect the dots: a systems level approach for analyzing the miRNA-mediated cell death network," *Autophagy*, vol. 9, no. 3, pp. 436–439, 2013.

[35] Q. Jiang, Y. Wang, Y. Hao et al., "miR2Disease: a manually curated database for microRNA deregulation in human disease," *Nucleic Acids Research*, vol. 37, no. 1, pp. D98–D104, 2009.

[36] U. M. Singh-Blom, N. Natarajan, A. Tewari, J. O. Woods, I. S. Dhillon, and E. M. Marcotte, "Prediction and validation of

gene-disease associations using methods inspired by social network analyses," *PLoS ONE*, vol. 8, no. 5, Article ID e58977, 2013.

[37] B. Grisart, W. Coppieters, F. Farnir et al., "Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition," *Genome Research*, vol. 12, no. 2, pp. 222–231, 2002.

[38] A. M. Krichevsky, K. S. King, C. P. Donahue, K. Khrapko, and K. S. Kosik, "A microRNA array reveals extensive regulation of microRNAs during brain development," *RNA*, vol. 9, no. 10, pp. 1274–1281, 2003.

[39] G. Thaller, C. Kühn, A. Winter et al., "DGAT1, a new positional and functional candidate gene for intramuscular fat deposition in cattle," *Animal Genetics*, vol. 34, no. 5, pp. 354–357, 2003.

[40] L. Cheng, Z.-G. Hou, Y. Lin, M. Tan, W. C. Zhang, and F.-X. Wu, "Recurrent neural network for non-smooth convex optimization problems with application to the identification of genetic regulatory networks," *IEEE Transactions on Neural Networks*, vol. 22, no. 5, pp. 714–726, 2011.

[41] A. Clop, F. Marcq, H. Takeda et al., "A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep," *Nature Genetics*, vol. 38, no. 7, pp. 813–818, 2006.

[42] J. Lim, T. Hao, C. Shaw et al., "A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration," *Cell*, vol. 125, no. 4, pp. 801–814, 2006.

[43] H. Lin, E. Deng, H. Ding, W. Chen, and K. Chou, "iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition," *Nucleic Acids Research*, vol. 42, no. 21, pp. 12961–12972, 2014.

[44] M. Oti and H. G. Brunner, "The modular nature of genetic diseases," *Clinical Genetics*, vol. 71, no. 1, pp. 1–11, 2007.

[45] L. D. Wood, D. W. Parsons, S. Jones et al., "The genomic landscapes of human breast and colorectal cancers," *Science*, vol. 318, no. 5853, pp. 1108–1113, 2007.

[46] D. Mrozek, B. Maysiak-Mrozek, and A. Sinik, "Search GenBank: interactive orchestration and ad-hoc choreography of Web services in the exploration of the biomedical resources of the National Center For Biotechnology Information," *BMC Bioinformatics*, vol. 14, article 73, 9 pages, 2013.

[47] B. Liu, J. Xu, Q. Zou, R. Xu, X. Wang, and Q. Chen, "Using distances between Top-n-gram and residue pairs for protein remote homology detection," *BMC Bioinformatics*, vol. 15, supplement 2, article S3, pp. 133–139, 2014.

[48] J.-B. Pan, S.-C. Hu, H. Wang, Q. Zou, and Z.-L. Ji, "PaGeFinder: quantitative identification of spatiotemporal pattern genes," *Bioinformatics*, vol. 28, no. 11, pp. 1544–1545, 2012.

[49] L. Wei, M. Liao, Y. Gao, R. Ji, Z. He, and Q. Zou, "Improved and promising identification of human microRNAs by incorporating a high-quality negative set," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 11, no. 1, pp. 192–201, 2014.

[50] Z. G. Hou, L. Cheng, M. Tan, and X. Wang, "Distributed adaptive coordinated control of multi-manipulator systems using neural networks," in *Robot Intelligence*, pp. 49–69, Springer, London, UK, 2010.

[51] B. Liu, X. Wang, Q. Chen, Q. Dong, and X. Lan, "Using amino acid physicochemical distance transformation for fast protein remote homology detection," *PLoS ONE*, vol. 7, no. 9, Article ID e46633, 2012.

[52] M. A. van Driel, J. Bruggeman, G. Vriend, H. G. Brunner, and J. A. M. Leunissen, "A text-mining analysis of the human phenome," *European Journal of Human Genetics*, vol. 14, no. 5, pp. 535–542, 2006.

[53] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick, "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders," *Nucleic Acids Research*, vol. 33, no. supplement 1, pp. D514–D517, 2005.

[54] U. Ala, R. M. Piro, E. Grassi et al., "Prediction of human disease genes by human-mouse conserved coexpression analysis," *PLoS Computational Biology*, vol. 4, no. 3, Article ID e1000043, 2008.

[55] H. Chen and Z. Zhang, "Similarity-based methods for potential human microRNA-disease association prediction," *BMC Medical Genomics*, vol. 6, article 12, pp. 215–221, 2013.

[56] Q. Jiang, Y. Hao, G. Wang et al., "Prioritization of disease microRNAs through a human phenome-microRNAome network," *BMC Systems Biology*, vol. 4, article S2, 2010.

[57] L. Katz, "A new status index derived from sociometric analysis," *Psychometrika*, vol. 18, no. 1, pp. 39–43, 1953.

[58] Y. Li and J. C. Patra, "Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network," *Bioinformatics*, vol. 26, no. 9, Article ID btq108, pp. 1219–1224, 2010.

[59] O. Vanunu, O. Magger, E. Ruppin, T. Shlomi, and R. Sharan, "Associating genes and protein complexes with disease via network propagation," *PLoS Computational Biology*, vol. 6, no. 1, Article ID 1000641, 2010.

[60] X. Wu, R. Jiang, M. Q. Zhang, and S. Li, "Network-based global inference of human disease genes," *Molecular Systems Biology*, vol. 4, article 189, 2008.

[61] F. Mordelet and J.-P. Vert, "A bagging SVM to learn from positive and unlabeled examples," *Pattern Recognition Letters*, vol. 37, no. 1, pp. 201–209, 2014.

[62] H. Chen and Z. Zhang, "Prediction of associations between OMIM diseases and MicroRNAs by random walk on OMIM disease similarity network," *The Scientific World Journal*, vol. 2013, Article ID 204658, 6 pages, 2013.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 573956, 11 pages
http://dx.doi.org/10.1155/2015/573956

Research Article

Low-Rank and Sparse Matrix Decomposition for Genetic Interaction Data

Yishu Wang,[1] Dejie Yang,[2] and Minghua Deng[1,3,4]

[1]*Center for Quantitative Biology, Peking University, Beijing 100871, China*
[2]*Institute of Computing Technology, Chinese Academy of Science, Beijing 100190, China*
[3]*School of Mathematical Sciences, Peking University, Beijing 100871, China*
[4]*Center for Statistical Sciences, Peking University, Beijing 100871, China*

Correspondence should be addressed to Minghua Deng; dengmh@math.pku.edu.cn

Received 8 January 2015; Accepted 13 March 2015

Academic Editor: Junwen Wang

Background. Epistatic miniarray profile (EMAP) studies have enabled the mapping of large-scale genetic interaction networks and generated large amounts of data in model organisms. One approach to analyze EMAP data is to identify gene modules with densely interacting genes. In addition, genetic interaction score (*S* score) reflects the degree of synergizing or mitigating effect of two mutants, which is also informative. Statistical approaches that exploit both modularity and the pairwise interactions may provide more insight into the underlying biology. However, the high missing rate in EMAP data hinders the development of such approaches. To address the above problem, we adopted the matrix decomposition methodology "low-rank and sparse decomposition" (LRSDec) to decompose EMAP data matrix into low-rank part and sparse part. *Results*. LRSDec has been demonstrated as an effective technique for analyzing EMAP data. We applied a synthetic dataset and an EMAP dataset studying RNA-related processes in *Saccharomyces cerevisiae*. Global views of the genetic cross talk between different RNA-related protein complexes and processes have been structured, and novel functions of genes have been predicted.

1. Introduction

Genetic interactions, which represent the degree to which the presence of one mutation modulates the phenotype of a second mutation, could be measured systematically and quantitatively in recent years [1, 2]. Genetic interactions can reveal functional relationships between genes and pathways. Furthermore, genetic networks measured via high-throughput technologies could reveal the schematic wiring of biological processes and predict novel functions of genes [3]. Recently, several high-throughput technologies have been developed to identify genetic interactions at genome scale, including Synthetic Genetic Array (SGA) [4], Diploid-Based Synthetic Lethality Analysis on Microarrays (dSLAM) [5], and epistatic miniarray profile (EMAP) [6]. In particular, EMAP systematically construct double deletion strains by crossing query strains with a library of test strains and identify genetic interactions by measuring a growth phenotype. An *S* score was calculated based on statistical methods for each pair of genes, while negative *S* scores represent synthetic sick/lethal and positive *S* scores indicate alleviating interactions [6].

Consequently, for each pair of genes, there are two different measures of relationship in EMAP platform. First, the genetic interaction score (*S* score) represents the degree of synergizing or mitigating effects of the two mutations in combination. Second, the similarity (typically measured as a correlation) of their genetic interaction profiles represents the congruency of the phenotypes of the two mutations across a wide variety of genetic backgrounds. So there are two important aspects in exploiting EMAP data. On the one hand, cellular functions and processes are carried out in series of interacting events, so genes participating in the same biological process tend to interact with each other. Therefore, algorithms that detect gene modules composed of densely interacting genes are of great interest. Within these modules, genes tend to have similar genetic interaction profiles; thus the submatrix for these genes tends to have a low-rank

structure. On the other hand, the cross talks between modules are usually indicated by gene pairs with high S scores (so that the genetic interaction is significant). Removing them results in better low-rank structure. Evocatively, these gene pairs are likely shadows over the low-rank matrix and connect different rank areas. These cross talks reveal the relationships of different biological process or protein complexes. Meanwhile, gene pairs exhibiting high absolute value of S scores may encode proteins that are physically associated or be enriched in protein-protein interactions [7–9]. So the investigation of S score is equally important. However, the current methodologies in genetic interaction networks analysis did not efficiently address these two important issues simultaneously.

In order to identify modules and between-module cross talks in genetic interaction networks, we employ the "low-rank and sparse decomposition" (LRSDec) to decompose EMAP data matrix into a low-rank part and a sparse part. We propose that the low-rank structure accounts for gene modules, in which genes have high correlations, and the sparsity matrix captures the significant S scores. In particular, entries in sparse matrix found by LRSDec correspond to two sources of biologically meaningful interactions, within-module interactions and between-module links. In this paper, we focus our discussion of the sparse matrix on the results of between-module links, while the results of within-module interactions can be found in the Supplementary Material available online at http://dx.doi.org/10.1155/2015/573956 (Supplementary Data 1).

Low-rank and sparse of matrix structures have been profoundly studied in matrix completion and compressed sensing [10, 11]. The robust principal component analysis (RPCA) [12] proved that the low-rank and the sparse components of a matrix can be exactly recovered if it has a unique and precise "low-rank + sparse" decomposition. RPCA offers a blind separation of low-rank data and sparse noises, which assumed $\mathbf{X} = \mathbf{L} + \mathbf{S}$ ($\mathbf{S}$ is the sparse noise), and exactly decomposes $\mathbf{X}$ into $\mathbf{L}$ and $\mathbf{S}$ without predefined rank($\mathbf{L}$) and card($\mathbf{S}$). Another successful matrix decomposition method GoDec studied the approximated "low-rank + sparse" decomposition of a matrix $\mathbf{X}$ by estimating the low-rank part $\mathbf{L}$ and the sparse part $\mathbf{S}$ from $\mathbf{X}$, allowing noise, that is, $\mathbf{X} = \mathbf{L} + \mathbf{S} + \mathbf{e}$, and constrained the rank range of $\mathbf{L}$ and the cardinality range of S [13]. It has been stated that GoDec has outperformed other algorithms before.

In this paper, we modified the GoDec matrix decomposition method and developed "low-rank and sparse decomposition" (LRSDec) to estimate the low-rank part $\mathbf{L}$ and the sparse part $\mathbf{S}$ of $\mathbf{X}$. LRSDec minimizes the nuclear norm of $\mathbf{L}$ and predefines the cardinality range of $\mathbf{S}$, while considering the additive noise $\mathbf{e}$. Different from GoDec, which directly constrains the rank range of $\mathbf{L}$, LRSDec minimizes its responding convex polytopes, that is, the nuclear norm of $\mathbf{L}$. It has been proven that the nuclear norm outperforms the rank-restricted estimator [14]. Furthermore, if, in presence of missing data, LRSDec could impute the missing entries while decomposing, with no need for data pretreatment, while GoDec could not accomplish decomposition and imputation simultaneously, then we stated the convergence properties of our algorithm and proved that, given the two regularization parameters, the objective value of LRSDec monotonically decreases. By applying both methods to a synthetic dataset, we demonstrated the superiority of LRSDec over GoDec. Finally, we analyzed a genetic interaction dataset (EMAP) using our algorithm and identified many biologically meaningful modules and cross talks between them.

2. Model

Let $\mathbf{X}$ be an $m \times n$ matrix that represents a genetic interaction dataset, where m is the number of query genes and n is the number of library genes. We propose to decompose $\mathbf{X}$ as

$$\mathbf{X} = \mathbf{L} + \mathbf{S} + \mathbf{e}, \tag{1}$$

where $\mathbf{L} \in \mathbb{R}^{m \times n}$ denotes the low-rank part and $\mathbf{S} \in \mathbb{R}^{m \times n}$ denotes the sparse part, and $\mathbf{e}$ is the noise. Here, we introduce $\mathbf{L} \in \mathbb{R}^{m \times n}$ to account for modules, in which genes are highly correlated. These modules correspond to protein complexes, pathways, and biological pathways, in which genes tend to share similar genetic interaction profiles [15]. $\mathbf{S} \in \mathbb{R}^{m \times n}$ is introduced to account for significant S scores, which are either gene pairs in the same module that have genetic interactions or cross talks among different functional modules.

Based on the assumptions above, we propose to solve the following optimization problem:

$$\begin{aligned} &\text{minimize } \operatorname{rank}(\mathbf{L}), \quad \text{minimize } \operatorname{card}(\mathbf{S}) \\ &\text{subject to } \sum_{(i,j)} \left(X_{ij} - L_{ij} - S_{ij}\right)^2 \leqslant \delta, \end{aligned} \tag{2}$$

where $\delta \geq 0$ is a regularization parameter that controls the error tolerance, and card($\mathbf{S}$) denote the number of nonzero entries in matrix $\mathbf{S}$.

To make the minimization problem tractable, we relax the rank operator on $\mathbf{L}$ with the nuclear norm, which has been proven to be an effective convex surrogate of the rank operator [14]

$$\begin{aligned} &\text{minimize } \|\mathbf{L}\|_*, \quad \text{minimize } \operatorname{card}(\mathbf{S}) \\ &\text{subject to } \sum_{(i,j)} \left(X_{ij} - L_{ij} - S_{ij}\right)^2 \leqslant \delta, \end{aligned} \tag{3}$$

where $\|\mathbf{L}\|_*$ is the nuclear norm of $\mathbf{L}$ ($\|\mathbf{L}\|_* = \sum_{i=1}^{r} \sigma_i$, where $\sigma_1, \ldots, \sigma_r$ are the singular values of $\mathbf{L}$ and r is the rank of $\mathbf{L}$).

However, missing data is commonly encountered in EMAP data, confounding techniques such as cluster analysis and matrix factorization. Here, we extend our basic model (3) to handle EMAP data with missing values by imputing missing entries in the matrix simultaneously when estimating low-rank matrix $\mathbf{L}$ and sparse matrix $\mathbf{S}$. Suppose that we only observe a subset of $\mathbf{X}$, indexed by Ω, and the missing entries are indexed by $\Omega^{\perp}$. In order to find a low-rank matrix $\mathbf{L}$ and

a sparse matrix $\mathbf{S}$ based on the observed data, we propose to solve the following optimization problem:

$$\begin{aligned} &\text{minimize } \|\mathbf{L}\|_*, \quad \text{minimize card}(\mathbf{S}) \\ &\text{subject to } \sum_{(i,j)\in\Omega} \left(X_{ij} - L_{ij} - S_{ij}\right)^2 \leqslant \delta. \end{aligned} \tag{4}$$

3. Algorithm

Similar to GoDec, the optimization problem of (3) can be solved by alternatively optimizing the following two subproblems until convergence:

$$\mathbf{L}_t = \arg\min_{\|\mathbf{L}\|_*} \left\|\mathbf{X} - \mathbf{L} - \mathbf{S}_{t-1}\right\|_F^2, \tag{5a}$$

$$\mathbf{S}_t = \arg\min_{\text{card}(\mathbf{S})\leq k} \left\|\mathbf{X} - \mathbf{L}_t - \mathbf{S}\right\|_F^2. \tag{5b}$$

In each iteration, we optimize the objective function by alternatively updating $\mathbf{L}$ and $\mathbf{S}$. Firstly, the subproblem (5a) can be solved by [14]. For fixed $\mathbf{S}$, the solution of (5a) is

$$\mathbf{L}_t = \mathbf{T}_\lambda\left(\mathbf{X} - \mathbf{S}_{t-1}\right). \tag{6}$$

Here, $\lambda \geq 0$ is a regularization parameter controlling the nuclear norm of estimated value $\mathbf{L}_t$, where

$$\begin{aligned} \mathbf{T}_\lambda(\mathbf{W}) &= \mathbf{U}\mathbf{D}_\lambda\mathbf{V}', \\ \text{with } \mathbf{D}_\lambda &= \text{diag}\left[(d_1 - \lambda)_+, \ldots, (d_r - \lambda)_+\right]. \end{aligned} \tag{7}$$

$\mathbf{U}\mathbf{D}\mathbf{V}'$ is the *Singular Value Decomposition* (SVD) of $\mathbf{W}$ and here $t_+ = \max(t, 0)$. The notation $\mathbf{T}_\lambda(\mathbf{W})$ refers to *soft-thresholding* [14].

Next, the subproblem (5b) in (3) could be updated via entry-wise hard thresholding of $\mathbf{X} - \mathbf{L}_t$ for fixed $\mathbf{L}_t$. Before giving the solution, we define an orthogonal projection operator $\mathscr{P}$. Suppose there is a subset of dataset $\mathbf{W}$, indexed by Ω; then the matrix $\mathbf{W}$ can be projected onto the linear space of matrices supported by Ω:

$$P_\Omega(\mathbf{W})_{(i,j)} = \begin{cases} W_{ij} & \text{if } (i,j) \in \Omega \\ 0 & \text{if } (i,j) \notin \Omega. \end{cases} \tag{8}$$

And $\mathscr{P}_{\Omega^\perp}$ is its complementary projection; that is, $\mathscr{P}_\Omega(\mathbf{W}) + \mathscr{P}_{\Omega^\perp}(\mathbf{W}) = \mathbf{W}$.

Then the solution of (5b) could be given as follows:

$$\mathbf{S} = \mathscr{P}_\Theta\left(\mathbf{X} - \mathbf{L}_t\right), \tag{9}$$

where $\mathscr{P}$ is the orthogonal projection operator as defined above, Θ is the nonzero subset of the first k largest entries of $|(\mathbf{X} - \mathbf{L}_t)|$. Then, the matrix $(\mathbf{X} - \mathbf{L}_t)$ can be projected onto the linear space of matrices supported by Θ:

$$\mathscr{P}_\Theta\left(\mathbf{X} - \mathbf{L}_t\right)_{(i,j)} = \begin{cases} \left(\mathbf{X} - \mathbf{L}_t\right)_{ij} & \text{if } (i,j) \in \Theta \\ 0 & \text{if } (i,j) \notin \Theta. \end{cases} \tag{10}$$

So far we have developed the algorithm for solving problem (3). As for problem (4), due to the existence of missing values, we took the optimization on the observed data, Ω. We updated $\mathbf{L}_t$ and $\mathbf{S}_t$ of the following optimization subproblems, respectively:

$$\mathbf{L}_t = \arg\min_{\|\mathbf{L}\|_*} \left\|\mathscr{P}_\Omega\left(\mathbf{X} - \mathbf{L} - \mathbf{S}_{t-1}\right)\right\|_F^2, \tag{11a}$$

$$\mathbf{S}_t = \arg\min_{\text{card}(\mathbf{S})\leq k} \left\|\mathscr{P}_\Omega\left(\mathbf{X} - \mathbf{L}_t - \mathbf{S}\right)\right\|_F^2. \tag{11b}$$

The term $\|\mathscr{P}_\Omega(\mathbf{X} - \mathbf{L} - \mathbf{S})\|_F^2$ is the sum of squared errors on the observed entries indexed by Ω.

The subproblem (11a) can be solved by updating $\mathbf{L}$ with an arbitrary initialization using [14]

$$\mathbf{L}_t \longleftarrow \mathbf{T}_\lambda\left(\mathscr{P}_\Omega\left(\mathbf{X} - \mathbf{S}_{t-1}\right) + \mathscr{P}_{\Omega^\perp}(\mathbf{L})\right). \tag{12}$$

The solution of subproblem (11b) is

$$\mathbf{S}_t = \mathscr{P}_\Theta\left(\mathscr{P}_\Omega\left(\mathbf{X} - \mathbf{L}_t\right)\right), \tag{13}$$

where Θ is the nonzero subset of the first k largest entries of $|\mathscr{P}_\Omega(\mathbf{X} - \mathbf{L}_t)|$.

Now we have the following algorithm.

Algorithm 1 (LRSDec). (i) Input: $\mathbf{X} \in \mathbb{R}^{m\times n}$. Initialize $\mathbf{S} \leftarrow \mathbf{0}$.
(ii) Iterate until convergence:
(a) $\mathbf{L}$-step: iteratively update $\mathbf{L}$ using (12).
(b) $\mathbf{S}$-step: Solve $\mathbf{S}$ using (13).
(iii) Output: $\mathbf{L}$, $\mathbf{S}$.

The convergence analysis of our algorithms is provided in the Supplementary Material.

4. Parameter Tuning

We have two parameters that need to be tuned in our models: λ and k. Here, we propose a 10-fold cross validation strategy to select them. The idea is as follows: let Ω be the index of observed entries of $\mathbf{X}$. We randomly partition Ω into 10 equal size subsets and choose training entries Ω_1 and testing entries Ω_2: $\Omega_1 \cup \Omega_2 = \Omega$ and $\Omega_1 \cap \Omega_2 = \emptyset$, $|\Omega_1| = 0.9 * |\Omega|$, $|\Omega_2| = 0.1 * |\Omega|$. We may solve problem (15) on a grid of (λ, k) values on the training data:

$$\mathbf{L}_t = \arg\min_{\|\mathbf{L}\|_*} \left\|\mathscr{P}_{\Omega_1}\left(\mathbf{X} - \mathbf{L} - \mathbf{S}_{t-1}\right)\right\|_F^2, \tag{14a}$$

$$\mathbf{S}_t = \arg\min_{\text{card}(\mathbf{S})\leq k} \left\|\mathscr{P}_{\Omega_1}\left(\mathbf{X} - \mathbf{L}_t - \mathbf{S}\right)\right\|_F^2. \tag{14b}$$

Then we evaluate the prediction error (15) on the testing data:

$$\text{Err}(\lambda, k) = \frac{1}{2}\left\|\mathscr{P}_{\Omega_2}\left(\mathbf{X} - \mathbf{L}(\lambda, k) - \mathbf{S}(\lambda, k)\right)\right\|_F^2. \tag{15}$$

The cross validation process is repeated for 10 times. Then we can find the optimal parameter (λ^*, k^*), which minimizes the mean of the prediction error.

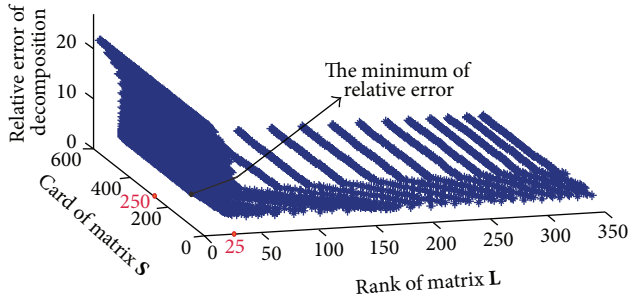


FIGURE 1: Results of parameter tuning on synthetic data. x-axis: different cardinality corresponding to parameter k. y-axis: different rank corresponding to parameter λ. z-axis: the relative error: $\|\mathbf{X} - \widehat{\mathbf{X}}\|_F^2 / \|\mathbf{X}\|_F^2$.

5. Results

5.1. Synthetic Data. We simulated a synthetic data and then applied LRSDec algorithm and GoDec algorithm to it. Specifically, low-rank part, sparse part, and noises are generated as follows.

(i) Low-rank part: the covariance matrix Σ is generated by $\mathbf{H}\mathbf{H}^T$, where $\mathbf{H} \in \mathbb{R}^{m \times K}$ and $\mathbf{H}_{i,j} \sim \mathcal{N}(0, 1)$. Here K is the number of hidden modules. The random entries $\mathbf{L}_j$ are drawn from $\mathcal{N}(\mathbf{0}, \tau\Sigma)$. Let $\mathbf{L} = [\mathbf{L}_1, \ldots, \mathbf{L}_n]$.

(ii) Sparse part: the non-zero entries in sparse matrix are generated from the tail of Gaussian distribution $\mathcal{N}(1, 2)$, whose upper quantile is $\alpha = 0.01$. We randomly selected 70% of them to assign the opposite sign. This is consistent with EMAP datasets, in which negative genetic interactions are much more prevalent than the positive ones.

(iii) $\mathbf{e} = 10^{-2} * \mathbf{F}$, wherein $\mathbf{F}$ is a standard Gaussian matrix.

A low-rank matrix L with rank 25 and sparse matrix with cardinality 250 were generated, respectively. Now we have

$$\mathbf{X} = \mathbf{L} + \mathbf{S} + \mathbf{e}. \tag{16}$$

The first step is parameter training, and the result is showed in Figure 1. Minimal prediction error was achieved when $\lambda = 250$ and $k = 25$, which coincides with the rank and cardinality of the synthetic data. This demonstrated the effectiveness of cross validation procedure.

Next, we compared the performance of LRSDec algorithm and GoDec algorithm by comparing their prediction error. The relative error is defined as

$$\frac{\left\|\mathbf{W} - \widehat{\mathbf{W}}\right\|_F^2}{\|\mathbf{W}\|_F^2}, \tag{17}$$

where $\mathbf{W}$ is the original matrix and $\widehat{\mathbf{W}}$ is an estimate/approximation. As both algorithms are influenced seriously by the parameters, we compared the relative error of the two algorithms by given different parameters. To make the comparison simple, we only changed one parameter with another parameter fixed (Figure 2). One can see that both algorithms reach the smallest relative if adopting the correct two parameters, and LRSDec outperforms GoDec. In the Supplementary Material, we also compare the performance of both algorithms under different noise setting, and the trends are the same.

5.2. Application to EMAP Data from Yeast. We also applied our method to EMAP data interrogating RNA processing in *S. cerevisiae*, which consists of 552 genes involved in RNA-related processes [9]. This genetic map contains about 152,000 pairwise genetic interaction measurements with about 29% missing entries in data matrix. We applied our method to this EMAP data, denoted as $\mathbf{X}$, to obtain two matrices, a low-rank matrix $\mathbf{L}$ and a sparse matrix $\mathbf{S}$. $\widehat{\mathbf{X}} = \mathbf{L} + \mathbf{S}$ is the new complete data matrix with imputed missing entries. To exploit the quantitative information from EMAP data, we first subjected the entire low-rank matrix $\mathbf{L}$ to hierarchical clustering, an approach that groups genes with similar patterns of genetic interactions. It should be noted that using low-rank matrix $\mathbf{L}$ in cluster improved the performance of hierarchical clustering in detecting genetic interaction modules [16]. According to the clustering result, we reordered rows and columns of matrix $\widehat{\mathbf{X}}$, so that the protein complexes and biological processes showed in [9] could be found (Figure S1).

To help identify more modules of cellular functions and processes and reveal the relationships between them, we further analyzed the matrices $\mathbf{L}$ and $\mathbf{S}$. Figure 3 is a flowchart of our strategy 1 to detect modules and cross talks between them through low-rank matrix $\mathbf{L}$. In this paper, we define module as a cluster from hierarchical clustering (HC) that passes through GO-enriched filtering (Supplementary Section 3). The details are as follows. Firstly, we clustered the row genes of matrix $\mathbf{L}$ using hierarchical clustering (HC). Here, we adopted the average-linkage hierarchical clustering algorithm in which the distance of gene A and gene B is defined as $1 - |\text{cor}(A, B)|$, where $|\text{cor}(A, B)|$ is the absolute value of the correlation of genetic interaction profile of gene A and gene B. A cutoff needs to be applied for HC to cut the hierarchical clustering tree. We used the Jaccard index (Supplementary Section 4) to determine how well the predicted gene sets correspond to benchmark (theoretical) gene sets [17]. Here, GO functional gene sets are used as benchmark (theoretical) gene sets. The cutoff at which HC achieved the highest Jaccard index is used to cut the hierarchical tree. We calculated the Jaccard index of every "height" cutoff in hierarchical clustering from 0.2 to 0.95 by 0.05 interval. This step resulted in the best Jaccard index with height = 0.7 and 84 clusters. Now we got the clusters of row genes, in which genes act in a consistent manner across the entire column genes. Then we filtered the clustering results by a hypergeometric test that calculates the significance of enrichment of GO items, and the p value was set to 0.01. The clusters enriched in GO functional categories are defined as row modules. Secondly, for each row module, we exploited modules of column genes based on this row gene module in matrix $\widehat{\mathbf{X}}$ by clustering the column genes of this submatrix of $\widehat{\mathbf{X}}$. Thirdly, we screened column clusters whose interactions with the row modules are

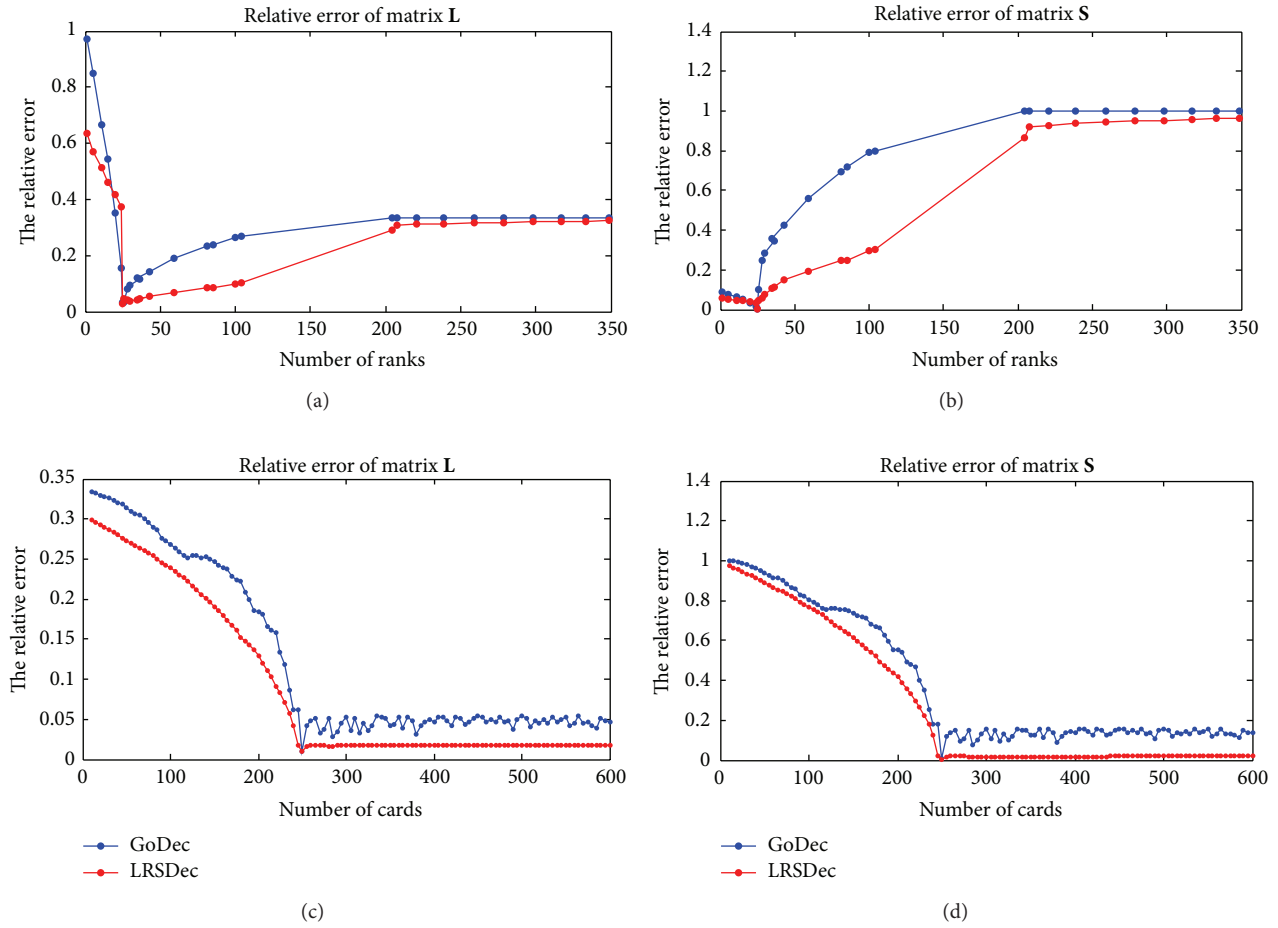


Figure 2: Performances of LRSDec and GoDec in low-rank and sparse decomposition tasks on synthetic data under different parameters. ((a)-(b)) Fixed parameter card, different parameter rank; ((c)-(d)) fixed parameter rank, different parameter card.

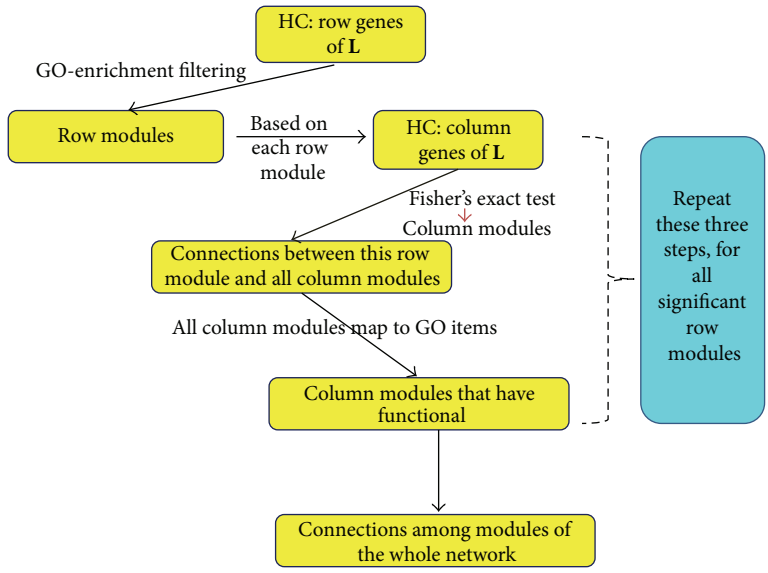


Figure 3: Flowchart of our strategy 1 for detecting modules and cross talks between them in genetic interaction network by low-rank matrix **L**.

Table 1: Fisher's exact test table. (Gene set A)$^{\perp}$ denotes the complementary set of gene set A. #{AB} denotes the number of connections between gene set A and gene set B. #{$AB^{\perp}$} denotes the number of connections between gene set A and the complementary set of gene set B.

	Gene set B	(Gene set B)$^{\perp}$
Gene set A	#{AB}	#{$AB^{\perp}$}
(Gene set A)$^{\perp}$	#{$A^{\perp}B$}	#{$A^{\perp}B^{\perp}$}

Table 2: Clustering results.

(a) GO slim as benchmark gene set

# Clusters	Low-rank matrix		Original matrix	
	JC-index	# Enriched[@]	JC-index	# Enriched[@]
200	0.063	190	0.022	46
150	0.070	138	0.032	44
100	0.084	90	0.050	50
50	0.088	30	0.067	24

(b) GO BP FAT as benchmark gene set

# Clusters	Low-rank matrix		Original matrix	
	JC-index	# Enriched[@]	JC-index	# Enriched[@]
200	0.137	183	0.044	47
150	0.147	142	0.058	50
100	0.155	96	0.091	52
50	0.131	34	0.078	26

@: hypergeometric test applied to test the enrichment of gene sets. Significance level: FDR <= 0.05. # Cluster: the number of clusters to cut off the hierarchical clustering tree. # Enriched: the number of modules predicted by hierarchical clustering enriched in the GO iterms.

significantly enriched by Fisher's exact test (Table 1, p value = 0.05). Here, we defined the positive genetic interactions as those gene pairs with genetic interaction scores $S \geq 2.0$ and negative as $S \leq -2.5$ [9]. The reduced gene sets of column genes were defined as column modules (corresponding to the row module). Next we identified the enriched GO functional categories of these column modules by mapping them to GO items (hypergeometric test). Finally, repeating these steps for all row modules, we identified the modules and intermodule genetic cross talk of the whole genetic interaction network (Figures 4–6), where red and green represent a statistically significant enrichment of positive and negative interactions. The cross talks constructed in these figures are the S scores among genes in the original matrix. In the following, we will discuss many of the interesting connections that have been reported previously.

The low-rank matrix found more functional modules than the original matrix (Table 2). In Table 2, we cut the dendrogram at different heights and compared the clusters obtained from the low-rank matrix and that of the original matrix. Jaccard index [17] is used to determine how well the predicted clusters recaptured the benchmark gene sets ((a) GO slim and (b) GO BP FAT). The definition of Jaccard index can be found in the Supplementary Material. In the ideal situation where predicted clusters perfectly match the benchmark gene sets, the Jaccard index is 1. The larger the Jaccard index, the better the predictions. The clusters obtained from clustering of the low-rank matrix are more enriched with GO functional categories at varying cutoffs (Table 2).

Figure 4 gives an overview of the relationships among biological processes when GO slim (downloaded from SGD [18], a broad overview of all of the top GO categories) is used as GO items. We found that several sets of genes that have been known to function in the same biochemical processes contain predominantly positive or negative interactions, which was also observed in [9]. For example, genes classed as involved in RNA splicing and transcription are significantly enriched in negative genetic interactions (Figure 4, green nodes). In contrast, the module involved in protein folding has strong positive interactions (red node). In addition, Figure 4 also suggested that not all modules have consistent pattern of interactions (yellow node), which is reasonable in biological processes. Finally, several connections have been previously discussed. For example, there is good evidence for functional interactions between splicing and transcription in [19] and functional interactions between splicing and translation in [20]. Furthermore, [21] reported the cooperative relationship between protein folding and chromosome organization.

Actually, if we classify the GO functional categories to more fine items (GO BP FAT, downloaded from DAVID http://david.abcc.ncifcrf.gov/), we can get a more comprehensive network (Figure 5). Many of the interaction results in Figure 5 have been reported before. For instance, genes involved in tubulin complex assembly process have negative genetic interactions with genes involved in tRNA wobble uridine modification process, supported by SGD, while the negative genetic interactions between RNA splicing process and tRNA metabolic process could also be found. Actually, the genetic interaction between RNA splicing and chromatin modification has been studied in [22]. And the balance of the interactions between the processing of ribonucleoprotein assembly of intronic noncoding RNAs and the splicing process regulating the levels of ncRNA and host mRNA can be found in [23]. Tubulin functionally relating to roles of the elongator complex in tRNA wobble uridine modification is supported by [24]. Moreover, epistasis and chromatin immunoprecipitation experiments indicating that the loss of Rrp6 (regulation of transcription) function is paralleled by the recruitment of Hda1 (histone deacetylase) have been reported by [25]. Finally, cotranscriptional recruitment of the mRNA export factor Yra1 realized by direct interaction with the 3′ end processing factor Pcf11 was in [26].

In an effort to gain insight into the functional organization of RNA-related complexes, we used the GO CC FAT (downloaded from DAVID) as the GO functional categories to create a map that highlights strong genetic trends both within and between these complexes. This result could be found in the Supplementary Material.

Now let us turn to the analysis of the sparse matrix **S**. The sparse matrix gives two distinct measures to exploit genetic information. First, extreme S scores indicate cofunctional membership more efficiently [27]. Second, some S scores

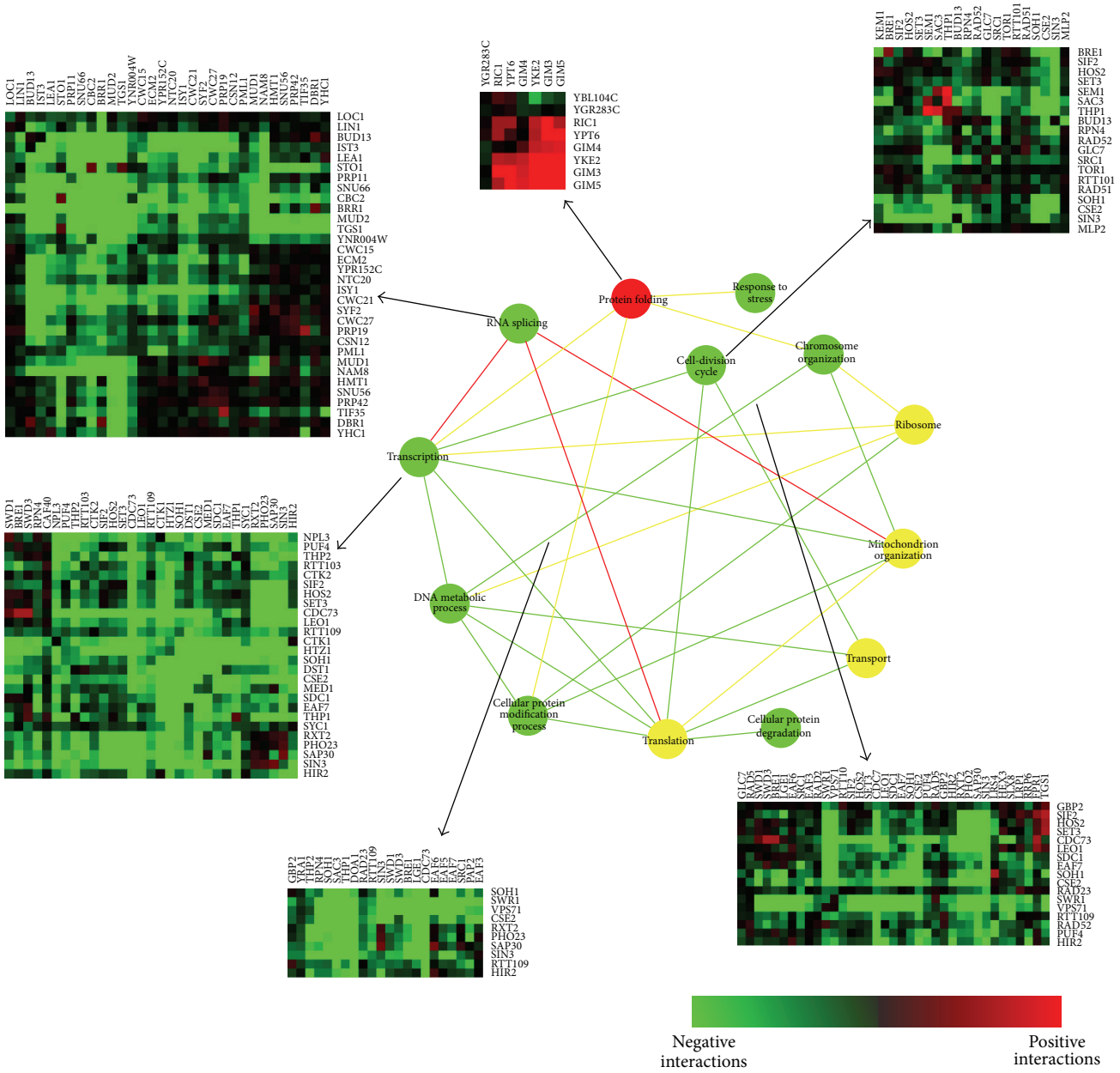


Figure 4: Global view of the genetic cross talks between different RNA-related processes (GO slim items). Green and red represent a statistically significant enrichment of negative (genetic interaction score $[S] \leq -2.5$) and positive (genetic interaction score $[S] > 2.0$) interactions, respectively, whereas yellow corresponds to cases where there are roughly equal numbers of positive and negative genetic interactions. Nodes (balls) correspond to distinct functional processes; edges (lines) represent how the processes are genetically connected. The square heat maps represent scores of interactions within one process, and the rectangle heat maps represent scores of interactions between two processes.

indicate the significant genetic interactions between genes in different gene sets. Following this clue, by analyzing the matrix **S**, we found much evidence of genes involved in the same functional modules and many cross talks between functional modules (Figure 6). Actually, many of them support the network in Figures 4-5. The information of gene pairs with extreme S scores and the involved modules could be found in the Supplementary Material (Supplementary Data 2).

Strategy 2 of sparse matrix analysis is similar to strategy 1 (Figure 3). First, for every row module (the same definition as that in Figure 3), cluster the column genes based on their genetic spectrums across genes in this row module. Then select the column gene sets, in which there are genes belonging to nonzero entries of sparse matrix to be defined as column modules. Finally, map these column modules to GO items, identifying their enriched functional categories (hypergeometric test). Similarly, for all row modules, repeat

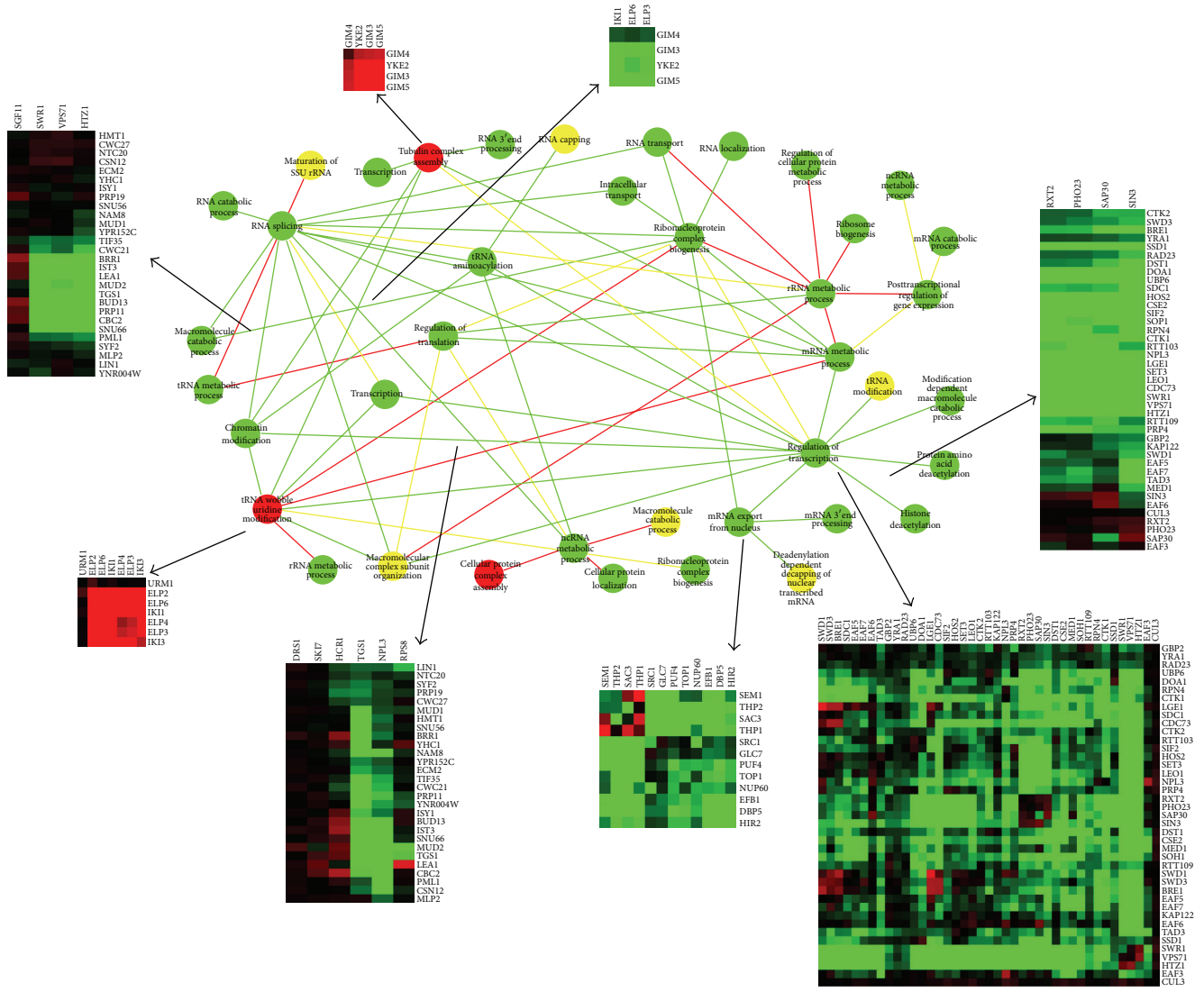


FIGURE 5: Global view of the genetic cross talks between different RNA-related processes (GO BP FAT). Green and red represent a statistically significant enrichment of negative (genetic interaction score $[S] \leq -2.5$) and positive (genetic interaction score $[S] > 2.0$) interactions, respectively, whereas yellow corresponds to cases where there are roughly equal numbers of positive and negative genetic interactions. Nodes (balls) correspond to distinct functional processes; edges (lines) represent how the processes are genetically connected. The square heat maps represent scores of interactions within one process, and the rectangle heat maps represent scores of interactions between two processes.

the above steps. Now we got the information of connections between different functional modules (Figure 6).

Several interesting connections become evident when the data is analyzed in this way. For example, there are negative genetic interactions between SRC1 and POM152 [28] and also physical interactions between them [28]. We have found the predominantly negative interactions between RNA transport and maturation of SSU-rRNA (Figure 6(a)). Also we found negative interactions between RNA transport and RNA localization (Figure 6(b)), where the negative interaction between EFB1 and DBP5 revealed in the sparse matrix probably reflects the cross talk between them. Another striking finding is the obviously negative interactions between protein folding and mRNA 3′ end process (Figure 6(d)). PAN3 has negative interactions with GIM4, which has been stated in [9, 29, 30], with GIM5, which has been stated in [29], and with YKE2, which has been stated in [29, 30]. NAB2 was clustered together with PAN3 but showed no obvious genetic interactions with protein folding genes in the original dataset, but in fact it has physical interactions with GIM3, GIM4, and GIM5 [31]. Furthermore, we found cross talk between protein folding and regulation of transcription. Although genes involved in regulation of transcription present low S score between each other, they are enriched in the same GO functional item: regulation of transcription (p value = 0.017813). More results could be found in Supplementary Material.

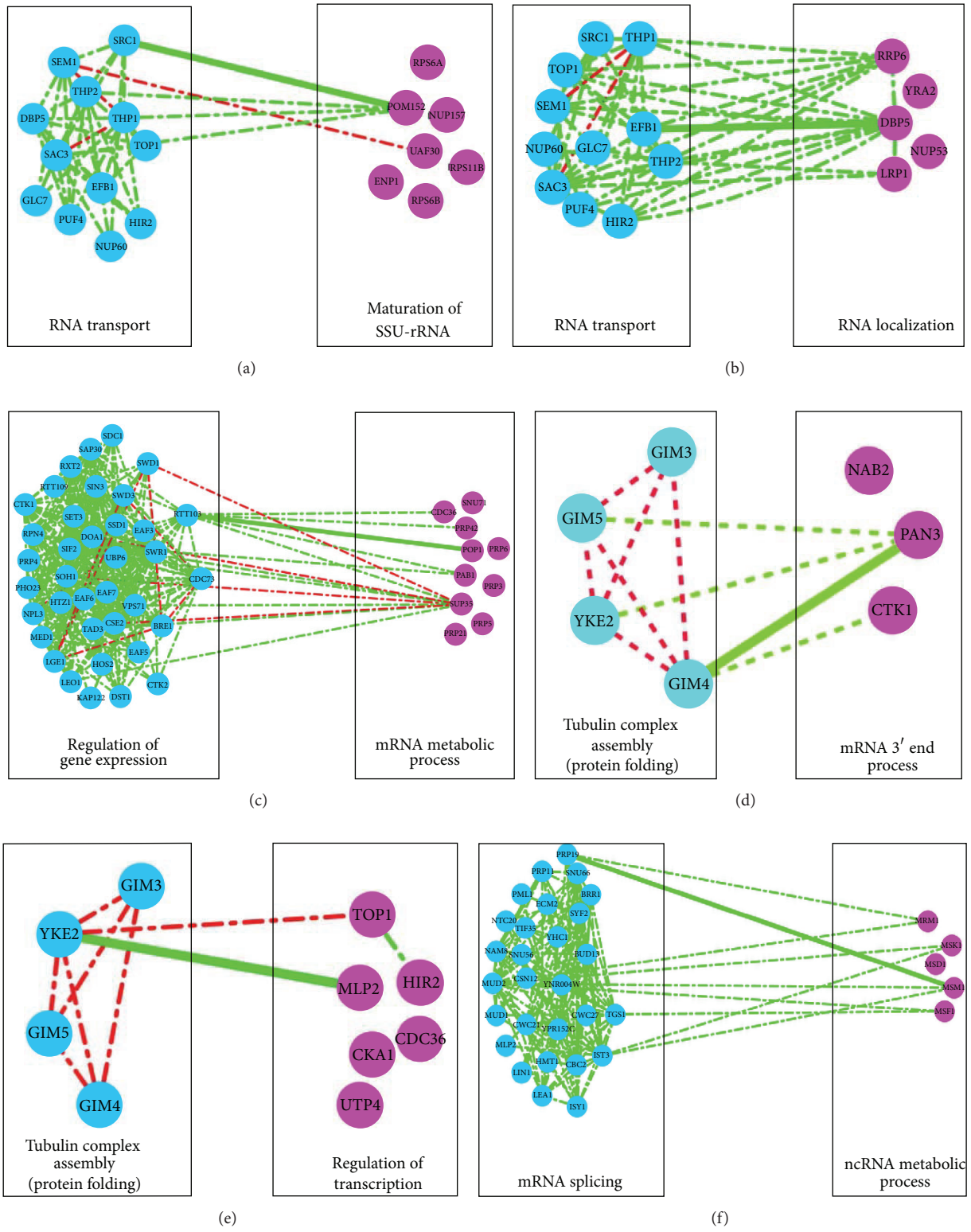


FIGURE 6: Functional associations between functional modules identified by sparse information. Blue and purplish red represent genes (nodes) involved in different modules. Edges (lines) represent how the processes are genetically connected, where green and red represent a statistically significant enrichment of negative (genetic interaction score [S] ≤ -2.5) and positive (genetic interaction score [S] > 2.0) interactions. The emphasized (thicker) lines are the significant S scores in sparse matrix.

6. Conclusion

In this paper, we have introduced a method named "LRSDec" to identify gene modules and cross talks between them in the genetic interaction network. LRSDec is based on low-rank approximation with regularization parameters and nearly optimal error bounds. We developed LRSDec to estimate the low-rank part **L** and the sparse part **S** of the original matrix **X**. In the synthetic data, LRSDec performed better than another matrix decomposition algorithm "GoDec," which has been shown to be other existing decomposition algorithms [13]. Then we applied our algorithm to a genetic interaction dataset to identify modules and cross talks between them. After the decomposition, subsequent analysis revealed many novel and biologically meaningful connections. Moreover, LRSDec could impute missing data while decomposing, which could not be accomplished by other decomposition algorithms. Actually, LRSDec will not be limited by the yeast genetic interaction data. As long as the dataset has internal low-rank structure and some sparse information, we can use the LRSDec algorithm to decompose the data matrix into addition of two matrixes and then analyze them separately. This algorithm could be used widely in the field of genetic interaction data analysis, image processing, and so on. We also had a try on the genetic interaction data of *C. elegans* in the Supplementary Material.

Conflict of Interests

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

Acknowledgments

This work is supported by the National Natural Science Foundation of China (nos. 31171262, 31471246, and 31428012) and the National Key Basic Research Project of China (no. 2015CB910303).

References

[1] R. A. Fisher, "XV.—the correlation between relatives on the supposition of mendelian inheritance," *Transactions of the Royal Society of Edinburgh*, vol. 52, no. 2, pp. 399–433, 1919.

[2] S. R. Collins, K. M. Miller, N. L. Maas et al., "Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map," *Nature*, vol. 446, no. 7137, pp. 806–810, 2007.

[3] R. Mani, R. P. St. Onge, J. L. Hartman IV, G. Giaever, and F. P. Roth, "Defining genetic interaction," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 9, pp. 3461–3466, 2008.

[4] A. H. Y. Tong, M. Evangelista, A. B. Parsons et al., "Systematic genetic analysis with ordered arrays of yeast deletion mutants," *Science*, vol. 294, no. 5550, pp. 2364–2368, 2001.

[5] X. Pan, D. S. Yuan, D. Xiang et al., "A robust toolkit for functional profiling of the yeast genome," *Molecular Cell*, vol. 16, no. 3, pp. 487–496, 2004.

[6] S. R. Collins, M. Schuldiner, N. J. Krogan, and J. S. Weissman, "A strategy for extracting and analyzing large-scale quantitative epistatic interaction data," *Genome Biology*, vol. 7, no. 7, article R63, 2006.

[7] S. R. Collins, P. Kemmeren, X.-C. Zhao et al., "Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*," *Molecular & Cellular Proteomics*, vol. 6, no. 3, pp. 439–450, 2007.

[8] A. Roguev, S. Bandyopadhyay, M. Zofall et al., "Conservation and rewiring of functional modules revealed by an epistasis map in fission yeast," *Science*, vol. 322, no. 5900, pp. 405–410, 2008.

[9] G. M. Wilmes, M. Bergkessel, S. Bandyopadhyay et al., "A genetic interaction map of rna-processing factors reveals links between sem1/dss1-containing complexes and mrna export and splicing," *Molecular Cell*, vol. 32, no. 5, pp. 735–746, 2008.

[10] R. H. Keshavan and S. Oh, "A gradient descent algorithm on the grassman manifold for matrix completion," http://arxiv.org/abs/0910.5260.

[11] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[12] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *JACM—Journal of the ACM*, vol. 58, no. 3, article 11, 2011.

[13] T. Zhou and D. Tao, "GoDec: randomized low-rank & sparse matrix decomposition in noisy case," in *Proceedings of the 28th International Conference on Machine Learning (ICML '11)*, pp. 33–40, July 2011.

[14] R. Mazumder, T. Hastie, and R. Tibshirani, "Spectral regularization algorithms for learning large incomplete matrices," *The Journal of Machine Learning Research*, vol. 11, pp. 2287–2322, 2010.

[15] A. H. Y. Tong, G. Lesage, G. D. Bader et al., "Global mapping of the yeast genetic interaction network," *Science*, vol. 303, no. 5659, pp. 808–813, 2004.

[16] Y. Wang, L. Wang, D. Yang, and M. Deng, "Imputing missing values for genetic interaction data," *Methods*, vol. 67, no. 3, pp. 269–277, 2014.

[17] J. Song and M. Singh, "How and when should interactome-derived clusters be used to predict functional modules and protein function?" *Bioinformatics*, vol. 25, no. 23, pp. 3143–3150, 2009.

[18] M. J. Cherry, C. Ball, S. Weng et al., "Genetic and physical maps of Saccharomyces cerevisiae," *Nature*, vol. 387, no. 6632, supplement, pp. 67–73, 1997.

[19] K. T. Chathoth, J. D. Barrass, S. Webb, and J. D. Beggs, "A splicing-dependent transcriptional checkpoint associated with prespliceosome formation," *Molecular Cell*, vol. 53, no. 5, pp. 779–790, 2014.

[20] R. J. Szczesny, M. A. Wojcik, L. S. Borowski et al., "Yeast and human mitochondrial helicases," *Biochimica et Biophysica Acta—Gene Regulatory Mechanisms*, vol. 1829, no. 8, pp. 842–853, 2013.

[21] K. Khan, U. Karthikeyan, Y. Li, J. Yan, and K. Muniyappa, "Single-molecule DNA analysis reveals that yeast Hop1 protein promotes DNA folding and synapsis: Implications for condensation of meiotic chromosomes," *ACS Nano*, vol. 6, no. 12, pp. 10658–10666, 2012.

[22] E. A. Moehle, C. J. Ryan, N. J. Krogan, T. L. Kress, and C. Guthrie, "The yeast sr-like protein npl3 links chromatin modification to mrna processing," *PLoS Genetics*, vol. 8, no. 11, Article ID e1003101, 2012.

[23] J. W. S. Brown, D. F. Marshall, and M. Echeverria, "Intronic noncoding RNAs and splicing," *Trends in Plant Science*, vol. 13, no. 7, pp. 335–342, 2008.

[24] R. Zabel, C. Bär, C. Mehlgarten, and R. Schaffrath, "Yeast α-tubulin suppressor Ats1/Kti13 relates to the Elongator complex and interacts with Elongator partner protein Kti11," *Molecular Microbiology*, vol. 69, no. 1, pp. 175–187, 2008.

[25] J. Camblong, N. Iglesias, C. Fickentscher, G. Dieppois, and F. Stutz, "Antisense RNA stabilization induces transcriptional gene silencing via histone deacetylation in *S. cerevisiae*," *Cell*, vol. 131, no. 4, pp. 706–717, 2007.

[26] S. A. Johnson, G. Cubberley, and D. L. Bentley, "Cotranscriptional recruitment of the mrna export factor yra1 by direct interaction with the 3′ end processing factor pcf11," *Molecular Cell*, vol. 33, no. 2, pp. 215–226, 2009.

[27] L. Wang, L. Hou, M. Qian, F. Li, and M. Deng, "Integrating multiple types of data to predict novel cell cycle-related genes," *BMC Systems Biology*, vol. 5, supplement 1, article S9, 2011.

[28] W. T. Yewdell, P. Colombi, T. Makhnevych, and C. P. Lusk, "Lumenal interactions in nuclear pore complex assembly and stability," *Molecular Biology of the Cell*, vol. 22, no. 8, pp. 1375–1388, 2011.

[29] M. Costanzo, A. Baryshnikova, J. Bellay et al., "The genetic landscape of a cell," *Science*, vol. 327, no. 5964, pp. 425–431, 2010.

[30] S. J. Dixon, Y. Fedyshyn, J. L. Y. Koh et al., "Significant conservation of synthetic lethal genetic interaction networks between distantly related eukaryotes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 43, pp. 16653–16658, 2008.

[31] J. Batisse, C. Batisse, A. Budd, B. Böttcher, and E. Hurt, "Purification of nuclear poly(A)-binding protein Nab2 reveals association with the yeast transcriptome and a messenger ribonucleoprotein core structure," *The Journal of Biological Chemistry*, vol. 284, no. 50, pp. 34911–34917, 2009.