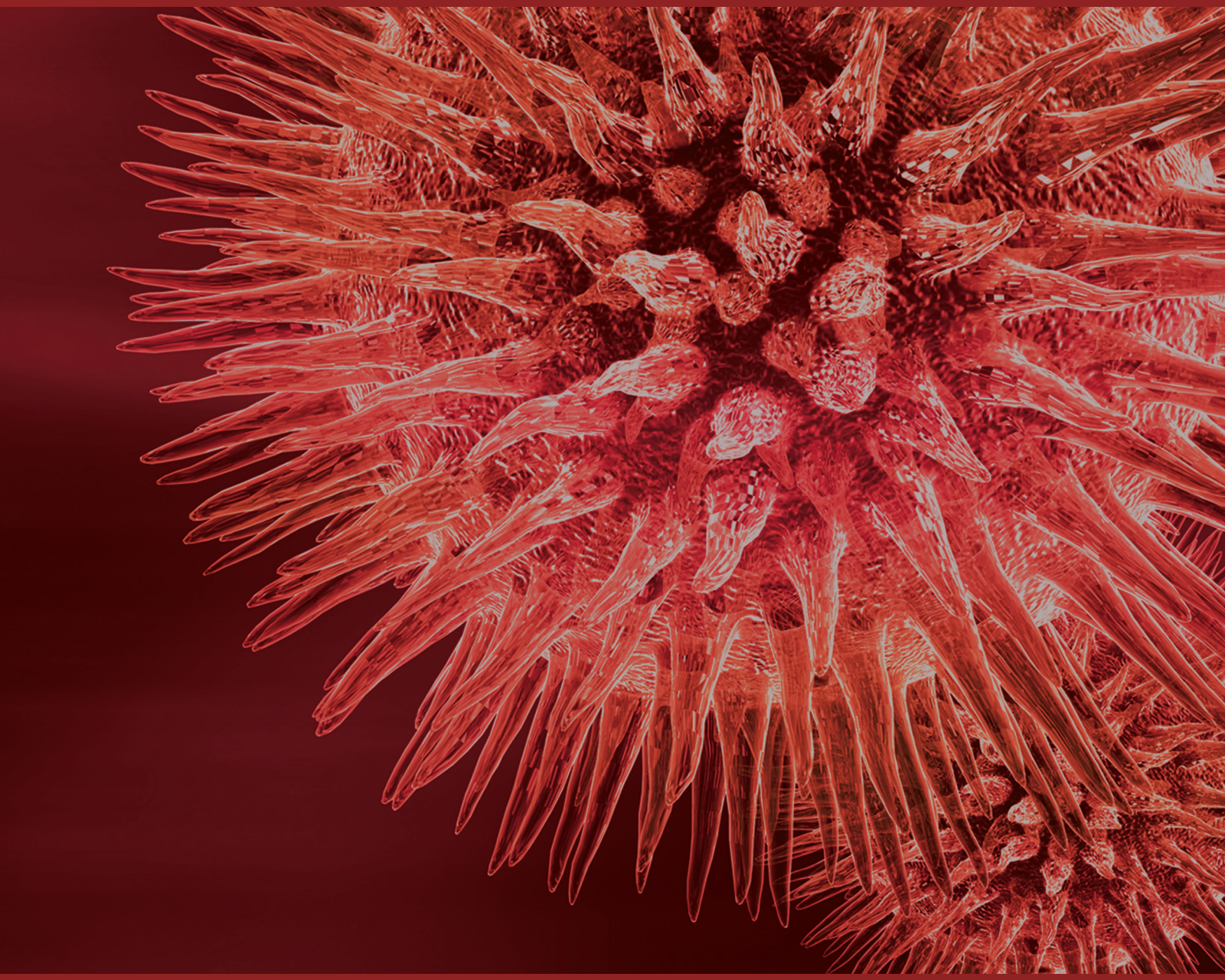
BioMed Research International

Novel Bioinformatics Approaches for Analysis of High-Throughput Biological Data

Guest Editors: Julia Tzu-Ya Weng, Li-Ching Wu, Wen-Chi Chang, Tzu-Hao Chang, Tatsuya Akutsu, and Tzong-Yi Lee



Novel Bioinformatics Approaches for Analysis of High-Throughput Biological Data

BioMed Research International

Novel Bioinformatics Approaches for Analysis of High-Throughput Biological Data

Guest Editors: Julia Tzu-Ya Weng, Li-Ching Wu, Wen-Chi Chang, Tzu-Hao Chang, Tatsuya Akutsu, and Tzong-Yi Lee

Contents

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 814092, 3 pages
http://dx.doi.org/10.1155/2014/814092

Editorial

Novel Bioinformatics Approaches for Analysis of High-Throughput Biological Data

Julia Tzu-Ya Weng,[1,2] Li-Ching Wu,[3] Wen-Chi Chang,[4] Tzu-Hao Chang,[5] Tatsuya Akutsu,[6] and Tzong-Yi Lee[1,2]

[1]*Department of Computer Science and Engineering, Yuan Ze University, Taoyuan 320, Taiwan*
[2]*Innovation Center for Big Data and Digital Convergence, Yuan Ze University, Taoyuan 320, Taiwan*
[3]*Institute of Systems Biology and Bioinformatics, National Central University, Taoyuan 320, Taiwan*
[4]*Institute of Tropical Plant Sciences, National Cheng Kung University, Tainan 701, Taiwan*
[5]*Graduate Institute of Biomedical Informatics, Taipei Medical University, Taipei 110, Taiwan*
[6]*Bioinformatics Center, Institute for Chemical Research, Kyoto University, Kyoto 611-0011, Japan*

Correspondence should be addressed to Tzong-Yi Lee; francis@saturn.yzu.edu.tw

Received 2 October 2014; Accepted 2 October 2014; Published 28 December 2014

1. Introduction

With the advent of high-throughput technologies, molecular biology is experiencing a surge in both growth and scope. As the amount of experimental data increases, the demand for the development of ways to analyze these results also increases. For example, the next-generation sequencing (NGS) technology has generated various sequencing data. Mass spectrometry- (MS-) based experiments are also widely applied in proteomics studies. Rapidly advancing technologies have offered us the opportunities to examine the genome, transcriptome, and proteome in comprehensive ways. Yet, extracting meaningful information from this vast sea of data and approaching biological problems from systems biology perspective have become the Holy Grail in bioinformatics. The main focus of this special issue is novelty: new ideas, original research findings, and practical applications that intend to answer biological questions through high-throughput technologies. The papers in this special issue present methods and experiments that demonstrate novel platforms and systems and new bioinformatics tools and models, as well as new data-analytical methods for high-throughput biological data.

In this special issue, U. Rosani et al. attempted to unravel the genome of *Mytilus galloprovincialis*, the Mediterranean mussel, through a target capture and high-throughput massive sequencing approach to reduce whole genome sequencing cost and effort. However, inferences from sequencing data rely heavily on careful experimental design, as well as efficient detection and removal of artifacts. While analyzing restriction-based reduced representation genomic data, D. C. Ilut et al. demonstrated that, by setting an optimal clustering threshold, false homozygosity or heterozygosity can be effectively minimized.

With the advancement of genomic researches, the number of sequences processed in comparative methods has grown immensely. E. A. Marucci et al. developed a parallel algorithm for multiple sequence similarities calculation using the k-mers counting method. Their tests showed that the algorithm provides a very good scalability and a nearly linear speedup. In "*A de novo genome assembly algorithm for repeats and nonrepeats*," Z. Dai et al. proposed a new genome assembly algorithm called the sliding window assembler (SWA), which assembles repeats and nonrepeats by adopting a new overlapping extension strategy to extend each seed and implementing a compensational mechanism for low coverage datasets. Results of their analysis on three datasets support the practicability and efficiency of SWA as a promising algorithm for NGS data.

High-throughput technology holds great promises for the efficient investigation of transcriptomes, but the enormous amount of gene expression data demands effective analytical

tools. By combining three gene-set analytical methods in one R statistical package, C.-Y. Chien et al. presented MAVTgsa, offering a systematic pipeline for the identification of significant gene-set modules from a set of gene expression data. Often, genes are coexpressed and coregulated or interact together to orchestrate a series of biological processes. To decipher the complex genetic networks associated with different cellular functions, M. Huerta et al. proposed to study the expression dependence between not only coexpressed genes but also sets of coexpressed genes.

In an attempt to predict the survival time in patients with oral squamous cell carcinoma, O. Hamidi et al. demonstrated that the three sparse variable selection techniques, when applied on gene expression microarray data, were able to yield better prediction results. For bladder cancer, Y.-H. Wong et al. proposed a statistical method based on carcinogenesis relevance values (CRVs) to identify 152 and 50 significant proteins and subsequently generated novel protein-protein interaction (PPI) network markers for early and late stage bladder cancer. Their findings not only provide new clues specific to cancer but also offer cancer researchers new directions for targeted cancer therapy.

In metagenomics, C.-M. Chiu et al. developed a pipeline for the systematic analysis of the association between gut flora and obesity through high-throughput sequencing and bioinformatics approaches. Eighty-one stool samples were collected and the V4 region of 16S rRNA genes was selected for metagenomics analysis. The results demonstrate that bacterial communities in the gut could be clustered into the N-like (normal) group and OB-like (obese) group. Remarkably, most of the normal samples were clustered in the N-like group, and the OB-like group was enriched with case samples, indicating that bacterial communities in the gut were highly associated with obesity. The results provide new insights into the correlation of gut flora with the rising trend in obesity.

In order to explore the molecular mechanism of flounder sex determination and development, Z. Fan et al. applied RNA-seq technology to investigate the transcriptomes of flounder gonads, obtaining 22,253,217 and 19,777,841 qualified reads from the ovary and testes, respectively. These reads were jointly assembled into 97,233 contigs. Among them, 2,193 contigs were identified to be differentially expressed in the ovary and 887 in the testes. Following annotation, several sex-related biological pathways including ovarian steroidogenesis and estrogen signaling pathways were revealed in the flounder for the first time.

Several bioinformatics tools are now being employed to analyze high-throughput expression data. In an attempt to study the molecular changes as a result of radiation exposure, K.-F. Lee et al. designed a set of expression microarray experiments studying the changes in gene and microRNA expression in peripheral mononuclear blood cells treated with varying doses of radiation. Combined with the existing tools for biochip analysis, K.-F. Lee et al. identified the various pathways associated with the exposure to differing doses of radiation and the potential gene-microRNA interactions that regulate these pathway changes.

The rapid increase in microRNA NGS data demands the development of comprehensive and customized tools for data analysis. In "*Large-scale investigation of human TF-miRNA relations based on coexpression profiles*," C.-H. Chien et al. developed a computational strategy to investigate the transcription factors of human miRNA genes on a global scale. The proposed method helps enhance our understanding of the transcriptional regulatory mechanisms of miRNAs. On the other hand, in "*miRSeq: a user-friendly standalone toolkit for sequencing quality evaluation and miRNA profiling*," C.-T. Pan et al. introduced a new tool for NGS data alignment that not only is easy to implement but also offers various methods for evaluating sequencing quality and provides profiles for up to 105 species for users to compare with. These studies demonstrate that customizability, easy access, and user-friendliness are crucial to high-throughput data analysis.

In the analysis of protein catalytic sites, C.-S. Yu et al. found that the side chain of catalytic residues usually points to the center of the catalytic site. The results demonstrate that the proposed method (EXIA2) could outperform the existing methods on several benchmark datasets that include over 1,200 enzyme structures. In "*High-throughput functional screening of steroid substrates with wild-type and chimeric P450 enzymes*," P. Urban et al. identified the structural features of steroid-based substrates catalyzed by CYP1A enzymes containing wild-type and synthetic variants. The results are interesting and may help extend the scope of knowledge surrounding the structural properties of enzymes recognizing and metabolizing exo- and endogenous substrates including drugs. In "*Bioinformatic prediction of WSSV-host protein-protein interaction*," Z. Sun et al. used bioinformatics methods to identify possible protein-protein interactions between white spot syndrome virus (WSSV) and its shrimp host. Their findings provide certain insights to readers in the relevant fields. In "*MPINet: metabolite pathway identification via coupling of global metabolite network structure and metabolomic profile*," F. Li et al. demonstrate a network-based metabolite pathway identification method, which identifies novel pathways related to disease.

In clinical medicine, length of stay (LOS) in the intensive care unit (ICU) of spontaneous intracerebral hemorrhage (sICH) patients is one of the most important issues. C.-L. Chan et al. showed that the threshold of a prolonged ICU stay is a good indicator of hospital utilization in ICH patients. This indicator can be improved using quality control methods such as complications prevention and efficiency of ICU bed management. Patients' stay in ICUs and in hospitals will be shorter if integrated care systems are established. In "*Ultrasonographic fetal growth charts: an informatic approach by quantitative analysis of the impact of ethnicity on diagnoses based on a preliminary report on Salentinian population*," A. Tinelli et al. provide customized fetal growth charts to formulate an accurate fetal assessment and to avoid unnecessary obstetric interventions. The fetal growth assessment is crucial to the health of newborns. Results suggest a careful reexamination for the appropriateness of continued use of currently adopted reference growth curves to classify neonates.

This special issue presents novel applications or methodologies of biomedical or bioinformatics analysis. The selected

articles show the importance of integrating diverse ideas and multidisciplinary knowledge to answer complex biological questions through high-throughput technologies.

Acknowledgments

We would like to thank the authors for their scientific contribution to this special issue. Also the anonymous reviewers are greatly appreciated for their critical comments that improve the quality of the papers published in this special issue.

Julia Tzu-Ya Weng
Li-Ching Wu
Wen-Chi Chang
Tzu-Hao Chang
Tatsuya Akutsu
Tzong-Yi Lee

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 159078, 23 pages
http://dx.doi.org/10.1155/2014/159078

Research Article

Evolution of Network Biomarkers from Early to Late Stage Bladder Cancer Samples

Yung-Hao Wong, Cheng-Wei Li, and Bor-Sen Chen

Lab of Control and Systems Biology, Department of Electrical Engineering, National Tsing Hua University, Hsinchu 30013, Taiwan

Correspondence should be addressed to Bor-Sen Chen; bschen@ee.nthu.edu.tw

Received 11 April 2014; Revised 9 July 2014; Accepted 10 July 2014; Published 18 September 2014

Academic Editor: Tzu-Hao Chang

We use a systems biology approach to construct protein-protein interaction networks (PPINs) for early and late stage bladder cancer. By comparing the networks of these two stages, we find that both networks showed very significantly different mechanisms. To obtain the differential network structures between cancer and noncancer PPINs, we constructed cancer PPIN and noncancer PPIN network structures for the two bladder cancer stages using microarray data from cancer cells and their adjacent noncancer cells, respectively. With their carcinogenesis relevance values (CRVs), we identified 152 and 50 significant proteins and their PPI networks (network markers) for early and late stage bladder cancer by statistical assessment. To investigate the evolution of network biomarkers in the carcinogenesis process, primary pathway analysis showed that the significant pathways of early stage bladder cancer are related to ordinary cancer mechanisms, while the ribosome pathway and spliceosome pathway are most important for late stage bladder cancer. Their only intersection is the ubiquitin mediated proteolysis pathway in the whole stage of bladder cancer. The evolution of network biomarkers from early to late stage can reveal the carcinogenesis of bladder cancer. The findings in this study are new clues specific to this study and give us a direction for targeted cancer therapy, and it should be validated in vivo or in vitro in the future.

1. Introduction

Cancer is the leading cause of death worldwide and its etiology occurs at the DNA, RNA, or protein level. It is a very complex disease involving cascades of spatial and temporal changes in the genetic network and metabolic pathways [1]. Various research studies have revealed that cancers are caused by multiple factors and intertwined events. Thus, in cancer therapy, it is important to dissect the diverse molecular mechanisms of cancer to identify potential cancers. Bladder cancer is amongst the 10 most common carcinomas in the USA, with 72,570 newly diagnosed cases, and it was the cause of 15,120 deaths in 2013 [2]. In particular, Kaufman et al. pointed out to it as the second most common form of cancer in 2008 [3]. In this study, we compared the early and late stages of bladder cancer to reveal additional mechanisms of bladder cancer development [4].

Biomarker discovery of various cancers is one of the key topic areas of cancer research. It can aid investigations into carcinogenesis and novel drug designs for cancer therapy. Several bioinformatics methods have been developed and applied to compare normal tissue with cancerous tissue to determine what cancer driving genes can act as cancer biomarkers [5–12].

Genes and proteins function cooperatively to regulate common biological cell processes by coregulating each other [13]. Generally, molecular regulation and interaction proceed with time and vary in different tissues. There must exist great differences in these variations between cancer and normal tissue. Proteins mutually interact with each other in the cell, and they form the PPI networks (PPINs). Currently, a lot of the research has focused on the relationship between PPINs and cancer development. For example, analysis of the cancer-related PPINs of apoptosis has unraveled the molecular mechanisms of cancer, which has helped to identify potential novel drug targets [14]. Our previous work [14] had successfully identified the network markers of lung cancer. In this study, we modified our previous method and applied the

novel concept to study the evolution of network markers from early to late stage bladder cancer.

Based on their PPI information and the gene expression profiles from cancer and surrounding normal samples, two PPI networks with quantitative protein association abilities for each cancer stage (early stage and late stage) and the surrounding noncancerous tissue are constructed, respectively. For each stage, the network structure and protein association abilities of the cancer and noncancer PPI networks are then compared to obtain sets of significant proteins which play important roles in the carcinogenesis process of bladder cancer.

Recently, PPI targets seem to have become a paradigm for the drug discovery of cancer therapy and precision medicine [15]. Unlike conventional drug design focusing on the inhibition of a single protein, usually an enzyme or receptor, small-molecule inhibition of direct PPIs that mediate many important biological processes is an emerging and challenging concept in drug design, especially for cancer. Extensive biological and clinical investigations have led to the identification of PPI hubs and nodes that have been critical for the acquisition and maintenance of characteristics for cell transformation in cancer. Such cancer-enabling PPIs will become promising therapeutic targets in anticancer strategies as the technologies in PPI modulator discovery and validating agents in the clinical setting advance in the future [15].

Therefore, future research directed at PPI target discovery, PPI interface characterization, and PPI-focused chemical libraries are expected to accelerate the development of the next generation of PPI-based anticancer agents. However, the PPI networks of cancer are very complex and quite differ between early and late stage cancer. In such circumstances, we will focus on the PPI network markers with their significant carcinogenesis relevance value (CRV) to exploit the important targets and their PPI interface for early and late stage cancer characterization. Then, we will not only gain insight into the crucial common pathways involved in bladder carcinogenesis, but we will also obtain a highly promising PPI target for bladder cancers. If we are then able to develop various combined anticancer strategies to target PPIs in the early and late stage network markers in the future, it may provide emerging opportunities for anticancer therapeutic approaches.

Chen et al. developed a dynamical network biomarker (DNB) that can serve as a general early warning signal to indicate an imminent bifurcation or sudden deterioration before the critical transition occurs; that means it can identify predisease state by time series microarray data. We use different approach from their methods by sample microarray data from bladder cancer patients of different stages. Our approach could also be extended to predict some similar results as their research. That is, in this study, we simply divided the cancer into early and late stages, but there are more stages of cancer, such as stages I, II, III, and IV. If we could observe the time evolution of the cancer biomarkers at these more different stages, we could also predict the predisease state by comparing it with these cancer biomarkers at different stages [16–18].

2. Materials and Methods

2.1. Overview of the Bladder Cancer Network Markers Construction Process. A flowchart representing the construction of network biomarkers for early and late stage bladder cancer is shown in Figure 1. We combined two data sources: (1) microarray data of bladder cancer and noncancer samples from the GEO database, while the cancer samples were divided into two groups: early stage and late stage bladder cancer. (2) The PPI database was required to construct the PPINs for bladder cancer. This data was used for PPI pool selection and the selected PPIs and the microarray data were then used for PPI network (PPIN) construction. Through regression modeling and the maximum likelihood parameter estimation method, a cancer PPIN (CPPIN) and a noncancer PPIN (NPPIN) was then obtained. The two constructed cancer and noncancer PPINs were compared to obtain the sets of significant proteins for bladder cancer based on the carcinogenesis relevance value (CRV) for each protein and the statistical assessment. The significant proteins and PPIs within these proteins were used to construct network markers at early and late stage bladder cancer.

2.2. Data Selection and Preprocessing. The microarray gene expression dataset of bladder cancer was obtained from the NCBI gene expression omnibus (GEO) [19]. In this study, we chose GSE13507 [20] and its corresponding platform GPL6012 as our research object. The same dataset contained the early and late stage bladder cancer and noncancer samples. We only used the data derived from nonprocessed primary biopsies to avoid the discrepancies in gene expression that are intrinsic to cell culture and fixation. Therefore, the dataset utilized contained primary tumor samples of both stages from patients and adjacent nontumor tissue samples from the same cancer patients, which could be considered as control samples. To describe the extent of a patient's cancer, the cancers were classified into four stages according to their degree of invasion and migration using the TNM staging system, as defined by the American Joint Committee on Cancer (AJCC) and the International Union against Cancer (UICC). We then divided the cancer samples into two groups. In general, stages I and II described early stage cancers that have higher curability rates with medical treatment, while stages III and IV described the late stages. However, there were no corresponding noncancer samples in the surrounding area for each stage and we had only one group of surrounding noncancer samples (Table 1). We built CPPIN and NPPIN for both early and late stage bladder cancer in this study. We obtained 37 and 106 samples for the early and late stage cancer, respectively, and 58 noncancer samples. To avoid overfitting in network construction, the maximum degree of the proteins in the PPI network should be less than the cancer/noncancer sample number [14]. In this dataset, we had a greater number of cancer and noncancer samples to overcome the sample size restriction on the size of the network. Prior to further analysis, the gene expression value, h_{ij}, was normalized to z-transformed scores, g_{ij}, for each gene, i, and then the normalized expression value resulting

Table 1: Descriptive information on datasets extracted from the GEO database used in this study.

Cancer	GEO accession number	Early stage	Late stage	Adjacent normal	Platform
Bladder cancer	GSE13507	106	37	58	GPL6102

Cases are grouped by cancer and surrounding normal tissues came from human patients of early stage and late bladder cancer.

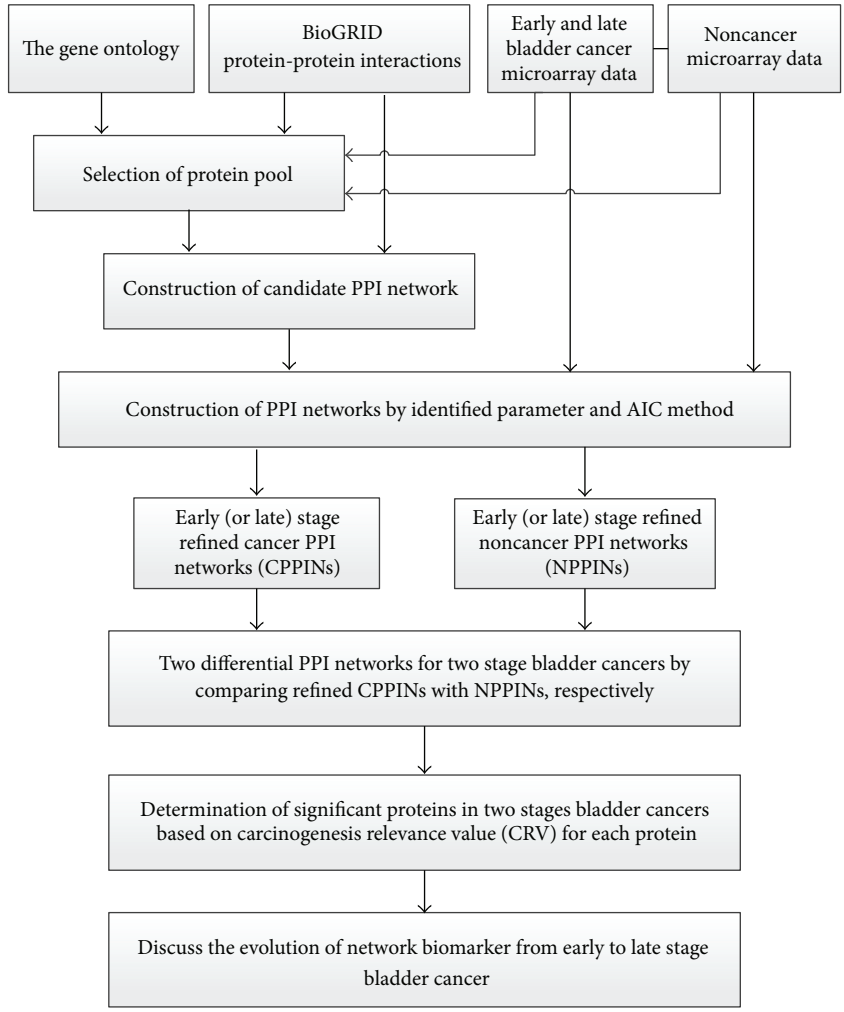


Figure 1: The flowchart of constructing both stages of network marker of bladder cancer and the investigation of the carcinogenesis mechanisms. We integrate microarray data, GO database, and PPI information to construct the PPI network. These data are used for pool selection, and then the selected proteins and the microarray data are used for the contribution of protein-protein interaction network (PPIN) by maximum likelihood estimation and model order detection method, resulting in bladder cancer PPIN (CPPIN) and noncancer PPIN (NPPIN) of early and late stage. The two constructed PPINs can be used for the determination of significant proteins of tumorigenesis by the difference between two PPI matrices of two constructed PPINs. With the help of the differential PPI matrix (network) between CPPIN and NPPIN, carcinogenesis relevance value (CRV) is computed for each protein, and significant proteins in carcinogenesis are determined based on P value the CRVs of these proteins in the differential PPI matrix between CPPIN and NPPIN. These significant proteins are obtained for early and late stage bladder cancers.

had a mean $\mu_i = 0$ and standard deviation $\sigma_i = 1$ over sample j [11, 14].

The PPI data for *Homo sapiens* were extracted from the Biological General Repository for Interaction Database (BioGRID, downloaded in October 2012). BioGRID is an open-access archive of genetic and protein interactions that are curated from the primary biomedical literature of all major model organisms. As of September 2012, BioGRID houses more than 500,000 manually annotated interactions from more than 30 model organisms [21]. The above two databases were mined for bladder cancer and noncancer PPI networks using their corresponding microarray data. These early and late stage bladder cancer and noncancer PPI networks were then compared to obtain network markers.

2.3. Selection of Protein Pool and Identification of the Protein-Protein Interaction Networks (PPINs) for Cancerous and Noncancerous Cells. To integrate gene expression with PPI data

to construct the corresponding CPPINs and NPPINs, we set up a protein pool containing differentially expressed proteins. The gene expression values were reasonably assumed to correlate with protein expression levels. We used one-way analysis of variance (ANOVA) to analyze the expression of each protein and select for proteins with differential expression levels. This method allowed determination of significant differences between cancer and noncancer datasets. The null hypothesis (Ho) was based on the assumption that the mean protein expression levels of cancer and noncancer sets are the same. Bonferroni adjustment [22], a type of multiple testing, was used to detect and correct proteins with discrepancy. Proteins with a P value of less than 0.01 were included in the protein pool. However, if the proteins in the protein pool did not have PPI information, they were eliminated. In addition, proteins that were not already in the protein pool were included if their PPI information could determine that they had a tight relationship with proteins already in the pool. As a result, the protein pool contained proteins that had certain differences in expression levels and proteins that had tight relationships with the aforementioned proteins. In this case, the protein pool in bladder cancer consisted of 2,245 proteins in the early stage and 1,101 proteins in the late stage.

On the strength of the significant pool and PPI information, candidate PPI networks for early and late stage bladder cancer were constructed for bladder cancer and noncancer by linking the proteins that interacted with each other. In other words, the proteins that had PPI information through the pool were linked together, resulting in candidate PPI networks.

As the candidate PPIN included all possible PPIs under various environments, different organisms, and experimental conditions, the candidate PPIN needed to be further confirmed by microarray data to identify appropriate PPIs according to the biological processes that are relevant to cancer. To remove false positive PPIs from each candidate PPIN for different biological conditions, we used both a PPI model and a model order detection method to prune each candidate PPIN using the corresponding microarray data to approach the actual PPIN. Here, the PPIs of a target protein i in the candidate PPIN can be depicted by the following protein association model:

$$x_i[n] = \sum_{j=1}^{M_i} \alpha_{ij} x_j[n] + \omega_i[n], \tag{1}$$

where $x_i[n]$ represents the expression levels of the target protein i for the sample n; $x_j[n]$ represents the expression level of the jth protein interacting with the target protein i for the sample n; α_{ij} denotes the association interaction ability between the target protein i and its jth interactive protein; M_i represents the number of proteins interacting with the target protein i; and $\omega_i[n]$ represents the stochastic noise due to other factors or model uncertainty. The biological meaning of (1) is that the expression levels of the target protein i are associated with the expression levels of the proteins interacting with it. Consequently, a protein association (interaction) model for each protein in the protein pool can be built as (1).

After constructing (1) for the PPI model of each protein in the candidate PPIN, we used the maximum likelihood estimation method [23] to identify the association parameters in (1) by microarray data as follows (see Supplementary Materials S.1 available online at http://dx.doi.org/10.1155/2014/159078):

$$x_i(n) = \sum_{j=1}^{M_i} \widehat{\alpha}_{ij} x_j(n) + w_i(n), \tag{2}$$

where $\widehat{\alpha}_{ij}$ is identified using microarray data in accordance with the maximum likelihood estimation method (see Supplementary Materials).

Once the association parameters for all proteins in the candidate PPI network were identified for each protein, the significant protein associations were determined using the interaction model order detection method based on the estimated association abilities. The Akaike information criterion (AIC) [23] and Student's t-test [24] were employed for both model order selection and significance determination of the protein associations in $\widehat{\alpha}_{ij}$ (see Supplementary Materials S.2).

2.4. Determination of Significant Proteins and Their Network Structures in the Carcinogenesis of Four Types of Cancers. After P values were determined using the AIC order detection and Student's t-test, spurious false positive PPIs $\widehat{\alpha}_{ij}$ in (2) were pruned away and only the significant PPIs that remained were refined as follows:

$$x_i(n) = \sum_{j=1}^{M_i'} \widehat{\alpha}_{ij} x_j(n) + w_i'(n), \quad i = 1, 2 \ldots M, \tag{3}$$

where $M_i' \leq M_i$ denotes the number of significant PPIs of PPIN, with the target protein i. In other words, a number of $M_i - M_i'$ (or false positives) are pruned in the PPIs of target protein i. One protein by one protein (i.e., $i = 1, 2, \ldots, M$ for all proteins in the refined PPIN in (3)) results in the following refined PPIN:

$$X(n) = AX(n) + w(n)$$

$$X(n) = \begin{bmatrix} x_1(n) \\ x_2(n) \\ \vdots \\ x_M(n) \end{bmatrix}, \qquad A = \begin{bmatrix} \widehat{\alpha}_{11} & \cdots & \widehat{\alpha}_{1M} \\ \vdots & \ddots & \vdots \\ \widehat{\alpha}_{M1} & \cdots & \widehat{\alpha}_{MM} \end{bmatrix},$$

$$w(n) = \begin{bmatrix} w_1'(n) \\ w_2'(n) \\ \vdots \\ w_M'(n) \end{bmatrix}, \tag{4}$$

where the interaction matrix A denotes the PPIs.

If there is no PPI between proteins i and j or it is pruned away by AIC order detection due to insignificance in the refined PPIN then $\widehat{\alpha}_{ij} = 0$. In general, $\widehat{\alpha}_{ij} = \widehat{\alpha}_{ji}$, but if this is

not the case, the larger one will be chosen as $\widehat{\alpha}_{ij} = \widehat{\alpha}_{ji}$ to avoid the situation where $\widehat{\alpha}_{ij} \neq \widehat{\alpha}_{ji}$. The above PPIN construction method was employed to construct the refined PPINs for each stage of bladder cancer (early and late) and noncancer cells. The interaction matrices A of the refined PPINs in (4) for cancer and noncancer cells of both the early and late stages of bladder cancer were constructed, respectively, as follows:

$$A_C^k = \begin{bmatrix} \widehat{\alpha}_{11,C}^k & \cdots & \widehat{\alpha}_{1M,C}^k \\ \vdots & \ddots & \vdots \\ \widehat{\alpha}_{M1,C}^k & \cdots & \widehat{\alpha}_{MM,C}^k \end{bmatrix},$$
$$A_N^k = \begin{bmatrix} \widehat{\alpha}_{11,N}^k & \cdots & \widehat{\alpha}_{1M,N}^k \\ \vdots & \ddots & \vdots \\ \widehat{\alpha}_{M1,N}^k & \cdots & \widehat{\alpha}_{MM,N}^k \end{bmatrix}, \tag{5}$$

where k = early and late stage bladder cancer; A_C^k and A_N^k denote the interaction matrices of refined PPIN of the kth cancer and noncancer, respectively; M is the number of proteins in the refined PPIN. Therefore, the protein association model for CPPIN and NPPIN in the kth stage bladder cancer and noncancer can be represented by the following equations according to (4) and (5):

$$x_C^k(n) = A_C^k x_C(n) + w_C^k(n),$$
$$x_N^k(n) = A_N^k x_N(n) + w_N^k(n), \tag{6}$$

where k = early and late stage bladder cancer;

$$x_C^k(n) = \left[x_{1C}^k \;\; x_{2C}^k \;\; \cdots \;\; x_{MC}^k\right]^T,$$
$$x_N^k(n) = \left[x_{1N}^k \;\; x_{2N}^k \;\; \cdots \;\; x_{MN}^k\right]^T \tag{7}$$

denote the vectors of expression levels; and $w_C^K(n)$ and $w_N^K(n)$ indicate the noise vectors of PPINs in the kth cancer and noncancer cells, respectively.

The different matrix $A_C^k - A_N^k$ of the differential PPI network between CPPIN and NPPIN in the kth cancer is defined as follows:

$$D^k = \begin{bmatrix} d_{11}^k & \cdots & d_{1M}^k \\ \vdots & \ddots & \vdots \\ d_{M1}^k & \cdots & d_{MM}^k \end{bmatrix}$$
$$= \begin{bmatrix} \widehat{\alpha}_{11,C}^k - \widehat{\alpha}_{11,N}^k & \cdots & \widehat{\alpha}_{1M,C}^k - \widehat{\alpha}_{1M,N}^k \\ \vdots & \ddots & \vdots \\ \widehat{\alpha}_{M1,C}^k - \widehat{\alpha}_{M1,N}^k & \cdots & \widehat{\alpha}_{MM,C}^k - \widehat{\alpha}_{MM,N}^k \end{bmatrix}, \tag{8}$$

where k = early and late stage bladder cancer; d_{ij}^k denotes the protein association ability difference between CPPIN and NPPIN in the kth stage bladder cancer; and the matrix D^k indicates the difference in network structure between CPPIN and NPPIN in the kth stage bladder cancer. In order to investigate carcinogenesis from the difference matrix D^k between CPPIN and NPPIN of the kth stage bladder cancer in (8), a score, which we named the carcinogenesis relevance value (CRV), was presented to quantify the correlation of each protein in D^k with the significance of carcinogenesis as follows [14]:

$$\mathrm{CRV}^k = \begin{bmatrix} \mathrm{CRV}_1^k \\ \vdots \\ \mathrm{CRV}_i^k \\ \vdots \\ \mathrm{CRV}_M^k \end{bmatrix}, \tag{9}$$

where $\mathrm{CRV}_i^k = \sum_{j=1}^{M} |d_{ij}^k|$, and k = early and late stage bladder cancer.

The CRV_i^k in (9) quantifies the differential extent of protein associations of the ith protein (the absolute sum of the ith row of D^k in (8)) and the CRV^k can differentiate CPPIN from NPPIN in the kth stage bladder cancer. In other words, the CRV_i^k in (9) could represent the network structure difference of the ith protein between the cancer and noncancer networks in the kth stage bladder cancer.

In order to investigate what proteins are more likely involved in the kth stage bladder cancer, we needed to calculate the corresponding empirical P value to determine the statistical significance of CRV_i^k. To determine the observed P value of each CRV_i^k, we repeatedly permuted the network structure of the candidate PPIN of the kth stage bladder cancer as a random network of the kth stage bladder cancer. Each protein in the random network of the kth stage bladder cancer will have its own CRV to generate a distribution of CRV_i^k for k = early and late stage bladder cancer. Although there was random disarrangement of the network structure, the linkages of each protein were maintained. In other words, the proteins with which a particular protein interacted were permuted without changing the total number of protein interactions. This procedure was repeated 100,000 times and the corresponding P value was calculated as the fraction of random network structure in which the CRV_i^k is at least as large as the CRV of the real network structure. According to the distributions of the CRV_i^k of the random networks, the CRV_i^k in (9) with a P value of less than or equal to 0.01 was regarded as a significant CRV and the corresponding protein was determined to be a significant protein in the carcinogenesis of the kth stage bladder cancer: a protein with a P value greater than 0.01 was removed from the list of significant proteins in carcinogenesis (in other words, if the P value of CRV_i^k was greater than 0.01, then the ith protein was removed from the CRV_i^k in (9) and the remainder in the CRV^k with P values of CRVs less than 0.01 were considered significant proteins of the kth stage bladder cancer).

Based on the P value of the CRVs for all proteins ($i = 1, 2, \ldots, M$) and the two stages of bladder cancer (k = early and late stage bladder cancer), we generated two lists of significant proteins for each of the two stages according to the CRV and the statistical assessment of each significant protein in CRV^k in (9). We found 152 significant proteins

in early stage bladder cancer and 50 significant proteins in late stage bladder cancer. These proteins showed significant changes between the CPPIN and NPPIN in the carcinogenic process according to their corresponding stage of cancer and we suspected that these changes might play important roles in the carcinogenesis process of bladder cancer. These findings warrant further investigation.

The intersections of these significant proteins in the early and late stages of bladder cancer and their PPIs are known as the core network markers appearing in all stages of bladder cancer. In contrast, the unique significant proteins and their PPIs in each stage of bladder cancers are known as the specific network markers for each stage of cancer. We found that there were 18 significant proteins that could be classified as a core network marker in the whole carcinogenesis process of bladder cancer. We also found 134 significant proteins in the specific network marker of early stage bladder cancer and 32 significant proteins in the specific network marker of late stage bladder cancer.

2.5. Pathway Analysis. Much valuable cellular information can be found in the known pathways, which are useful for describing most "normal" biological phenomena. All of these known pathways are the result of repeated testing and verification and the entire pathway network has given definitions for most links. Therefore, the proteins we identified to be significant in the above network markers were mapped onto the known pathway networks (e.g., the KEGG or PANTHER pathway) to investigate significant pathways with the network marker and to explore the relationships between these pathways and the carcinogenesis of bladder cancer. This approach supports the view that systems biology can help identify significant network biomarkers in both normal and cancerous pathways to their roles in the pathogenesis of cancer.

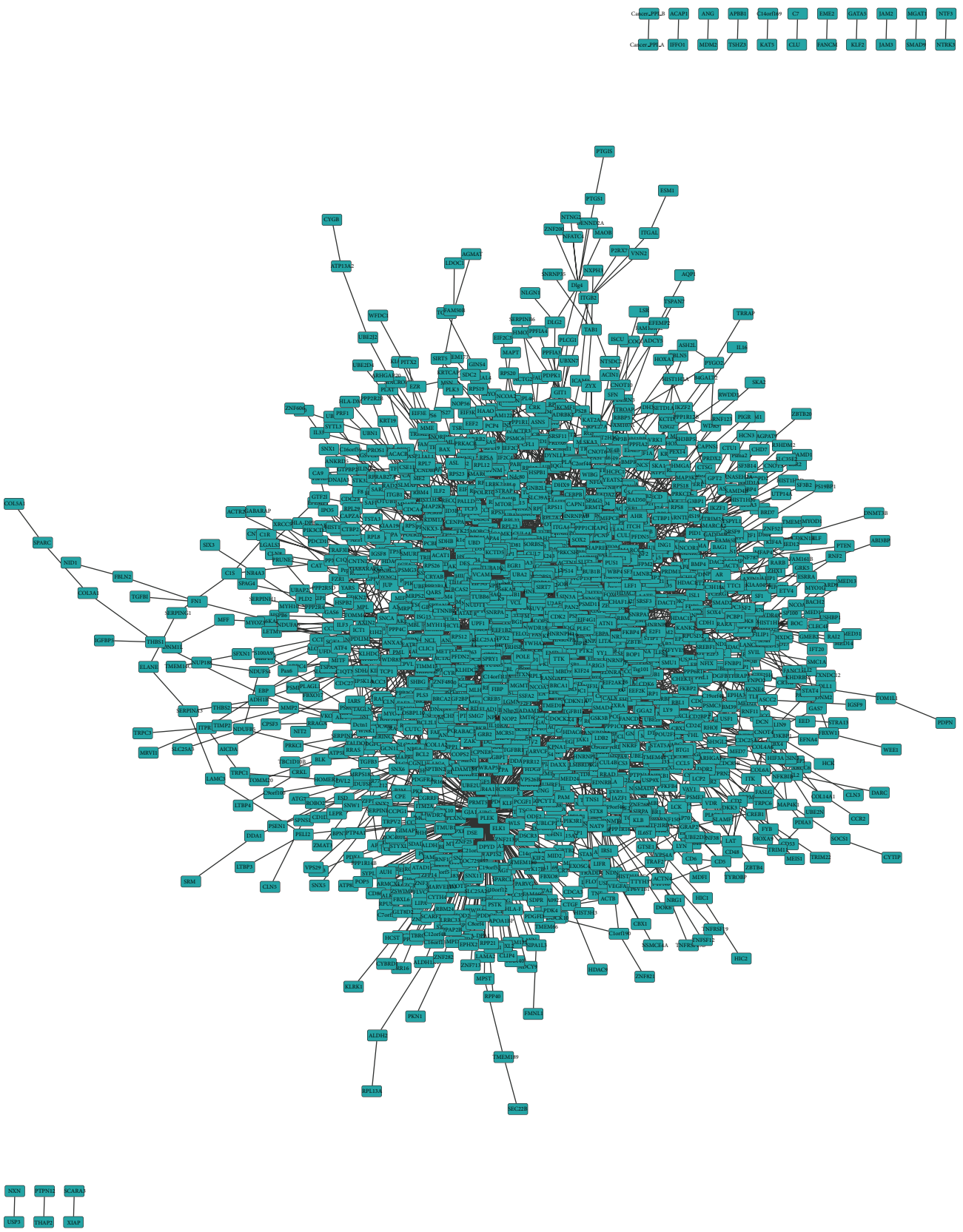
Together with comprehensive pathway databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG), we used a series of bioinformatics pathway analysis tools to identify biologically relevant pathway networks [25]. KEGG includes manually curated biological pathways that cover three main categories: systems information (e.g., human diseases and drugs), genomics information (e.g., gene catalogs and sequence similarities), and chemical information (e.g., metabolites and biochemical reactions). At present, KEGG contains 134,511 distinct pathways generated from 391 original reference pathways [26]. Therefore, to investigate the pathways involved in carcinogenesis, the bioinformatics database DAVID [27, 28], which generates automatic outputs of the results from KEGG pathway analysis [27], was used for the pathway analysis of significant proteins identified in network markers to determine their roles in the pathogenesis of early and late stage bladder cancer. Our methodology does not contain the pathway analysis and gene set enrichment analysis. To complete our research results, we used the NOA software to do the pathway analysis and gene set enrichment analysis on biological processes, cellular components, and molecular functions [19, 29].

2.6. The Contribution of Protein Interaction Network Will Affect the Results of Biomarkers and the Evolution of Network Biomarkers. Our cancer PPI model is constructed from the differential expression of cancer and noncancer microarray data and data mining of PPI information from BioGRID database. So, the early and late stage bladder cancer CPPINs (cancer PPI networks) and NPPINs (noncancer PPI networks) are the results of our systems biology model using the original microarray data and PPI databases. There are three key factors that will affect the final results.

(i) The effect of different microarray data: we know that the microarray data has the shortage of irreproducible. That means even in the same case the microarray data does not promise to produce the same result as the previous ones. Also, for the same cancers, patients of different ethnics, different age, or different sex will give the different microarray data. This is the first factor to affect the final results.

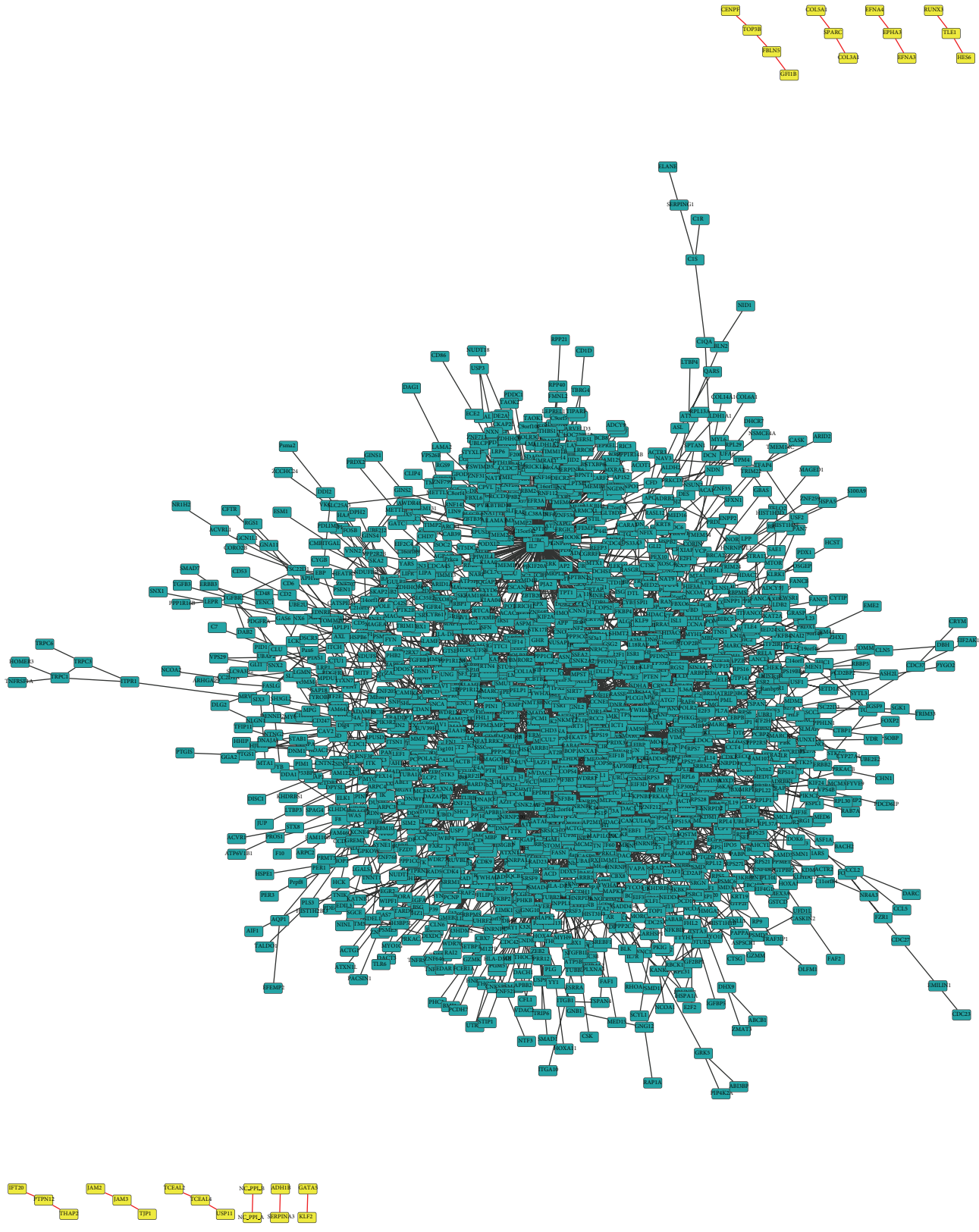
(ii) The effect of different original PPI databases: we know that PPI databases, such as BioGRID and MIPS, are constructed from putative and validated by wet-lab experiments. Due to the advances of many high-throughput experimental skills, the original PPI databases are evolved with time growing. The new updated original PPI databases are the second factor to affect the final results.

(iii) The effect of systems biology model: microarray data, PPI databases, and PPI interaction model in (1) are employed to construct the PPI networks of normal and cancer cells by the maximum likelihood parameter estimation method (see Supplementary Material S.1). The AIC system order detection method (Supplementary Materials S.2) is employed to prune the false positive PPIs to obtain the real PPI networks of normal and cancer cells; that is, we use the so-called reverse engineering method to construct PPI networks of normal and cancer cells. Then the differential PPI network between cancer PPI network and normal PPI network is obtained in (8) to investigate PPI variations of each protein in the differential PPI network due to the carcinogenesis. Finally, the carcinogenesis value (CRV) based on PPI variations is also proposed to evaluate the significance of carcinogenesis for each protein of differential PPI network. Proteins with significant CRV (P value $<$ 0.01) are considered as significant proteins of the cancer. The significant proteins in Table 3 are these significant proteins of early and late stage bladder cancers, and these proteins and their PPIs construct the interaction network in Figure 2. Finally, from the early to late stage bladder cancer network markers, we investigate the mechanism of carcinogenesis process with the help of databases (e.g., GO database, DAVID, and KEGG pathway database) and try to find multiple network target therapy of cancer. Unlike the conventional theoretical methods, which always give a single mathematical model for cancer network for a more detailed theoretical analysis, this study



(A) CPPIN for early stage bladder cancer

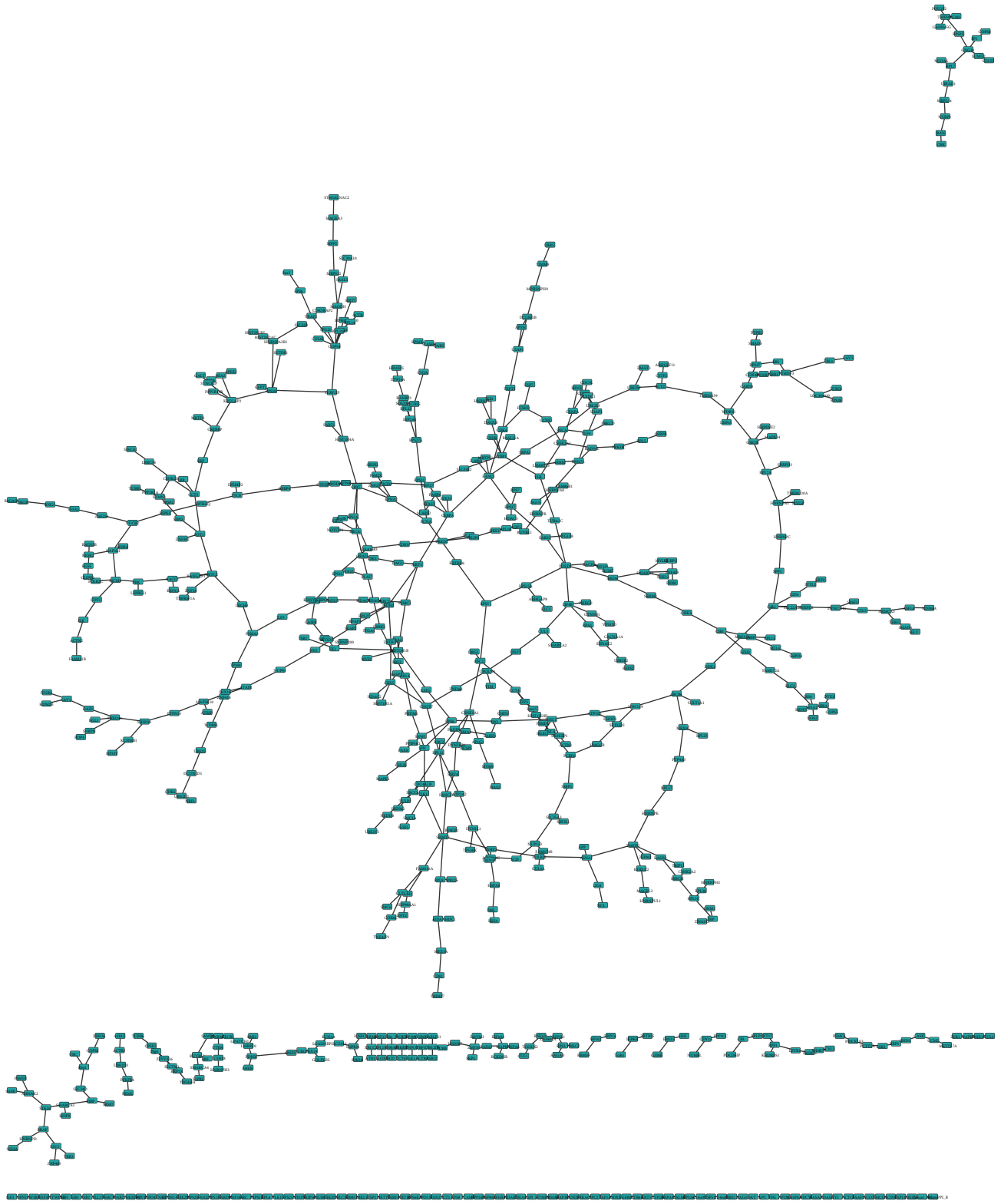
Figure 2: Continued.



(B) NPPIN for early stage bladder cancer

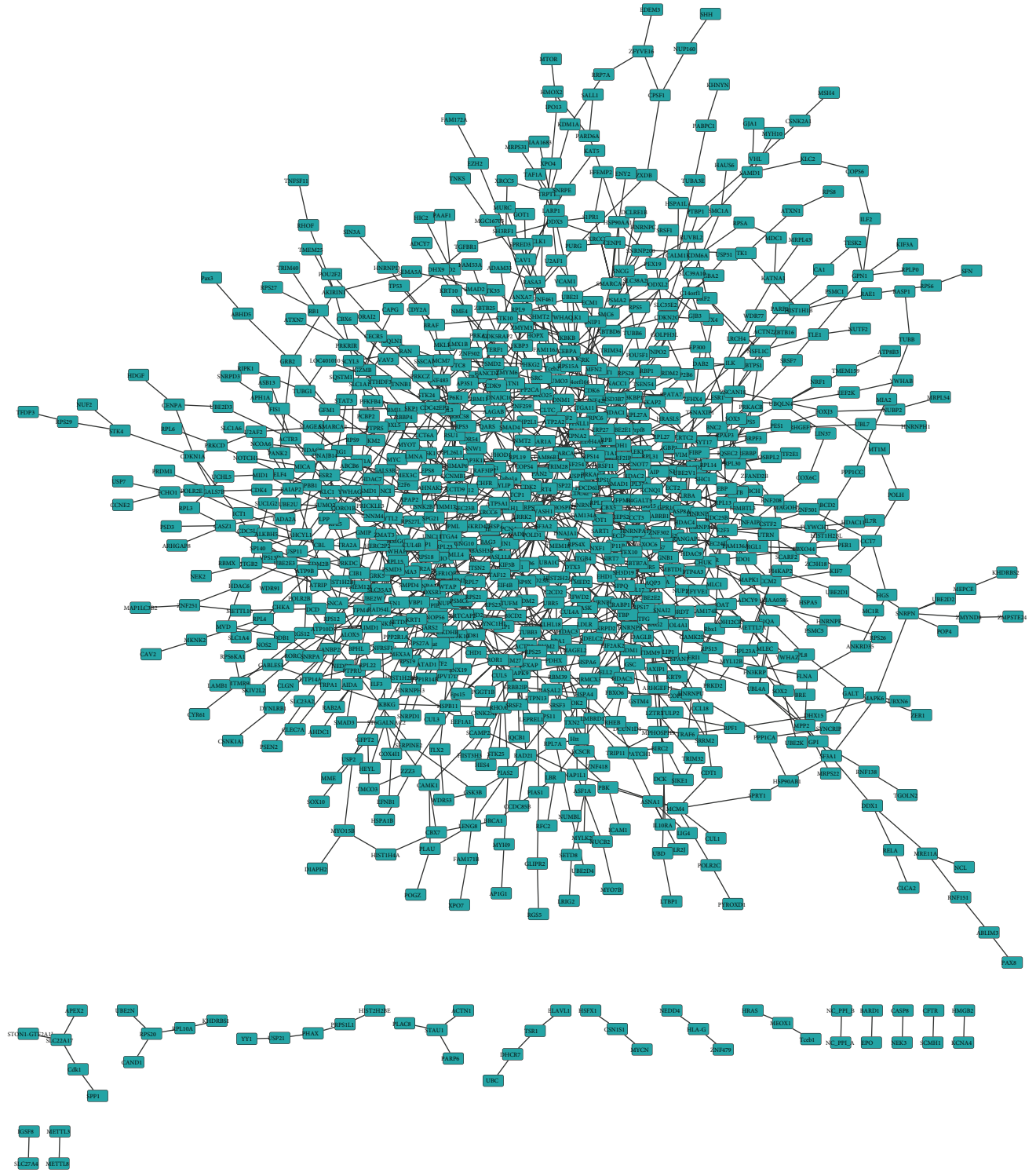
(a)

Figure 2: Continued.



(A) CPPIN for late stage bladder cancer

Figure 2: Continued.



(B) CPPIN for late stage bladder cancer

(b)

Figure 2: The constructed cancer PPIN (CPPIN) and noncancer PPIN (NPPIN) for early and late stage bladder cancer. The protein association numbers of CPPIN and NPPIN with respect to early and late bladder cancers are listed below (CPPIN/NPPIN): early stage bladder cancer (3388/3151) and late stage bladder cancer (634/1185). The figures are created using Cytoscape.

Table 2: The 18 identified significant proteins of core network marker in both early and late stage bladder cancers.

Common network marker of early and late stage bladder cancer				
Protein	CRV-early	*P* value-early	CRV-late	*P* value-late
UBC	29.91709	$<1e-5$	158.5321	$<1e-5$
CUL3	27.96694	$<1e-5$	13.0117	$<1e-5$
CUL5	14.97713	$<1e-5$	4.834916	0.002872
RPL22	10.47367	$<1e-5$	8.110447	$<1e-5$
SUMO2	8.391421	$<1e-5$	10.34113	$<1e-5$
APP	6.933807	$<1e-5$	11.47363	$<1e-5$
SH3KBP1	6.765387	$<1e-5$	4.447911	0.00619
PTBP1	6.740458	$<1e-5$	4.511016	0.005506
ELAVL1	6.635085	$<1e-5$	10.77056	$<1e-5$
SIRT7	5.55441	$3.13E-05$	7.656515	$<1e-5$
MYC	4.6109	0.00072	13.0423	$<1e-5$
HSP90AA1	4.60136	0.00072	6.513345	0.000123
COPS5	3.898548	0.003163	6.601779	$9.22E-05$
ESR1	3.873735	0.003383	5.6189	0.000614
BRCA1	3.788256	0.00415	11.5863	$<1e-5$
TERF2IP	3.680287	0.00487	5.202998	0.00149
SOX2	3.534024	0.007219	5.495338	0.00076
CUL1	3.521039	0.00736	13.2669	$<1e-5$

is to introduce a systems biology approach to cancer network markers based on real microarray data through the so-called reverse engineering, theoretical statistical method and data mining method in combination with big databases. These are the novelty and significance of our paper. Although we described the novelty of our systems biology model, we have validated our results by literature surveying in the research. In the future, our results will be validated by other researchers' wet-lab experiments, and we will modify our mathematical model again and again. This is the third key factor to affect the results. Although not directly, it will also have the influence on protein interaction network.

We also know that the biosystems are evolved with time. It is obvious that the early stage and late stage patients have very different symptoms; they are the key features for us to classify early and late stage bladder cancers. Since the two stage bladder cancer patients have great different symptoms, it is undoubted that the microarray data of these two stage patients will show to be quite different. As described above, the protein expression from microarray data is one of the key factors of our systems biology model to give the final CPPINs and NPPINs. And the CPPINs and NPPINs give the final network biomarkers from our systems biology model. So, the most important thing for the network biomarkers evolving is due to the evolution of microarray data at both stages of bladder cancer, which is inherent in the exhibition of cancer-related genes due to DNA mutations in the carcinogenesis process.

3. Results and Discussion

3.1. Time Evolution of the Network Biomarker from Early to Late Stage Bladder Cancer. In the first instance, we built the CPPIN and NPPIN for early and late stage bladder cancer (Figure 2). From the differential networks between CPPIN and NPPIN of early stage and late stage bladder cancer, we then calculated the CRV of each protein in the network structure. Screening in accordance with the *P* value of CRV, we determined the significant proteins of network markers for the two stages of bladder cancer. In the following, we will discuss the significant proteins identified in both stages and their intersection to reveal the carcinogenesis mechanisms from early to late stage bladder cancer.

3.2. Network Marker of Early and Late Stage Bladder Cancer. After *P* value (0.01) screening, we found that there were 152 and 50 significant proteins for early and late stage bladder cancer, respectively. In addition, their corresponding CRV values ranged between 4.1 and 158.5 and 3.4–29.9, respectively. These significant proteins and their PPIs were used to construct the network markers at early and late stage bladder cancer. The intersection network marker of both stages was a core feature that contained 18 significant proteins in carcinogenesis. We listed the 18 significant proteins and their corresponding CRV and *P* value in both stages of bladder cancer (Table 2). From this, we separately identified the 10 most significant proteins in early and late stage bladder cancer (Table 3). The full list of the 152 and 50 significant proteins for the two stages of bladder cancer is detailed in supplementary tables (Tables S1 and S2).

3.3. Pathway Analysis of Early Stage Bladder Cancer. We analyzed the pathway of early stage bladder cancer using the DAVID database. Our initial observation revealed that several cancer pathways were hit by the 152 key proteins, including 11 genes in hsa05200: pathways in cancer (Figure 3(a)), 7 genes involved in prostate cancer, 6 genes involved in

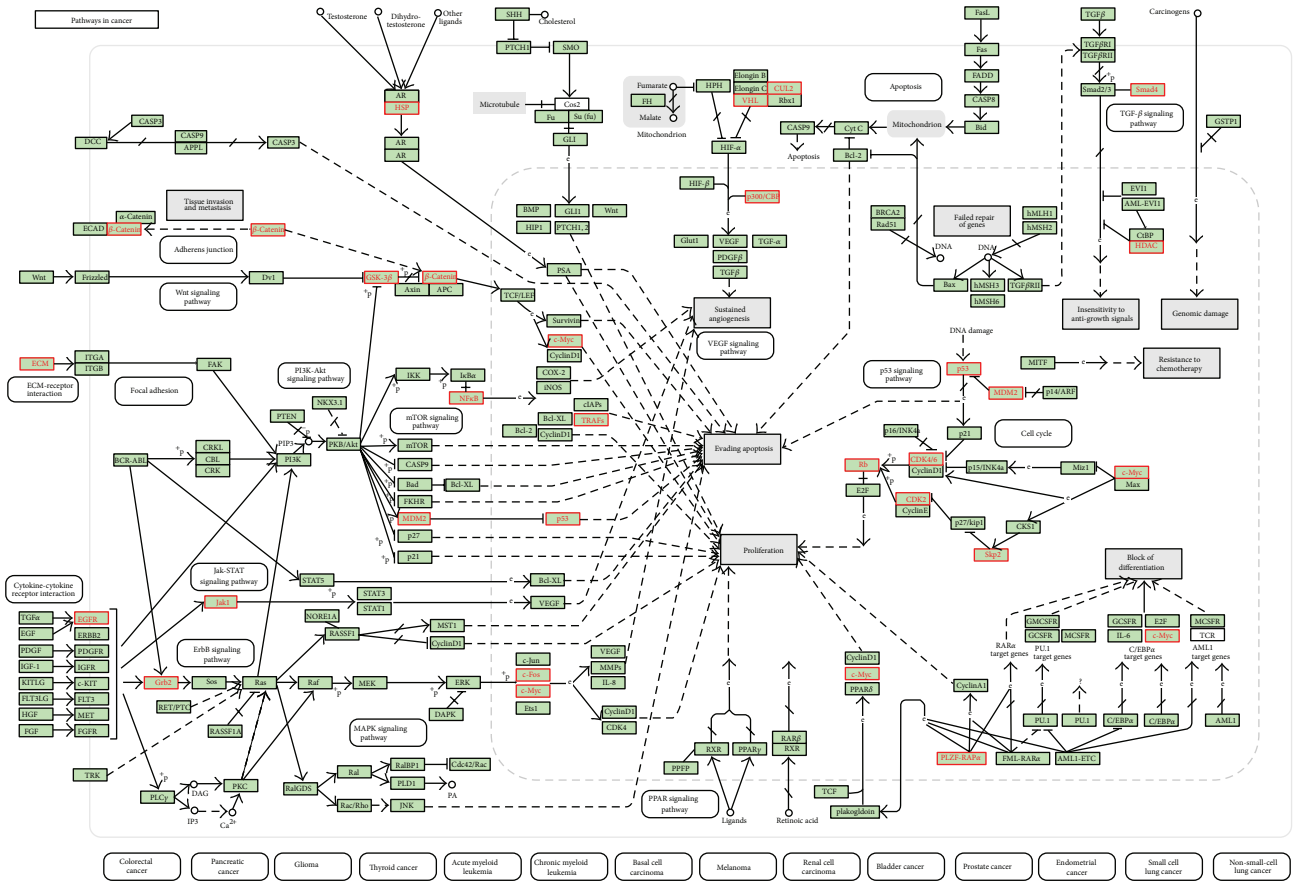
Table 3: The identified top 20 significant proteins in both early and late stage bladder cancer individually.

Early stage bladder cancer ($N = 107$)			Late stage bladder cancer ($N = 107$)		
CRV	Name	P value	CRV	Name	P value
UBC	29.91709	$<1e-5$	UBC	158.5321	$<1e-5$
CUL3	27.96694	$<1e-5$	VCAM1	20.98798103	$<1e-5$
RIOK2	16.02326	$<1e-5$	RPS13	20.09693015	$<1e-5$
CUL5	14.97713	$<1e-5$	TP53	19.5883	$<1e-5$
RPS23	12.13218	$<1e-5$	HDAC1	19.2879	$<1e-5$
RPL12	10.87102	$<1e-5$	HSPA8	17.24137906	$<1e-5$
RPL22	10.47367	$<1e-5$	RPS27A	17.23738059	$<1e-5$
RANBP2	9.8086	$<1e-5$	TUBB	17.03734405	$<1e-5$
PAN2	9.521207	$<1e-5$	CDK2	16.7366	$<1e-5$
DHX9	9.47832	$<1e-5$	VIM	15.89214155	$<1e-5$
RPS8	8.722495	$<1e-5$	KIAA0101	15.8188	$<1e-5$
RPL27	8.641642	$<1e-5$	ITGA4	15.69058519	$<1e-5$
SUMO2	8.391421	$<1e-5$	GSK3B	15.44597966	$<1e-5$
HNRNPH3	8.011681	$<1e-5$	EEF1A1	14.21690842	$<1e-5$
CDC5L	7.950851	$<1e-5$	RUVBL2	13.63207486	$<1e-5$
RUVBL1	7.887244	$<1e-5$	PCNA	13.3217	$<1e-5$
SF3A1	7.468209	$<1e-5$	CUL1	13.2669	$<1e-5$
APP	6.933807	$<1e-5$	MYC	13.0423	$<1e-5$
CCT3	6.860228	$<1e-5$	CUL3	13.0117	$<1e-5$
SH3KBP1	6.765387	$<1e-5$	HNRNPA0	12.15264603	$<1e-5$

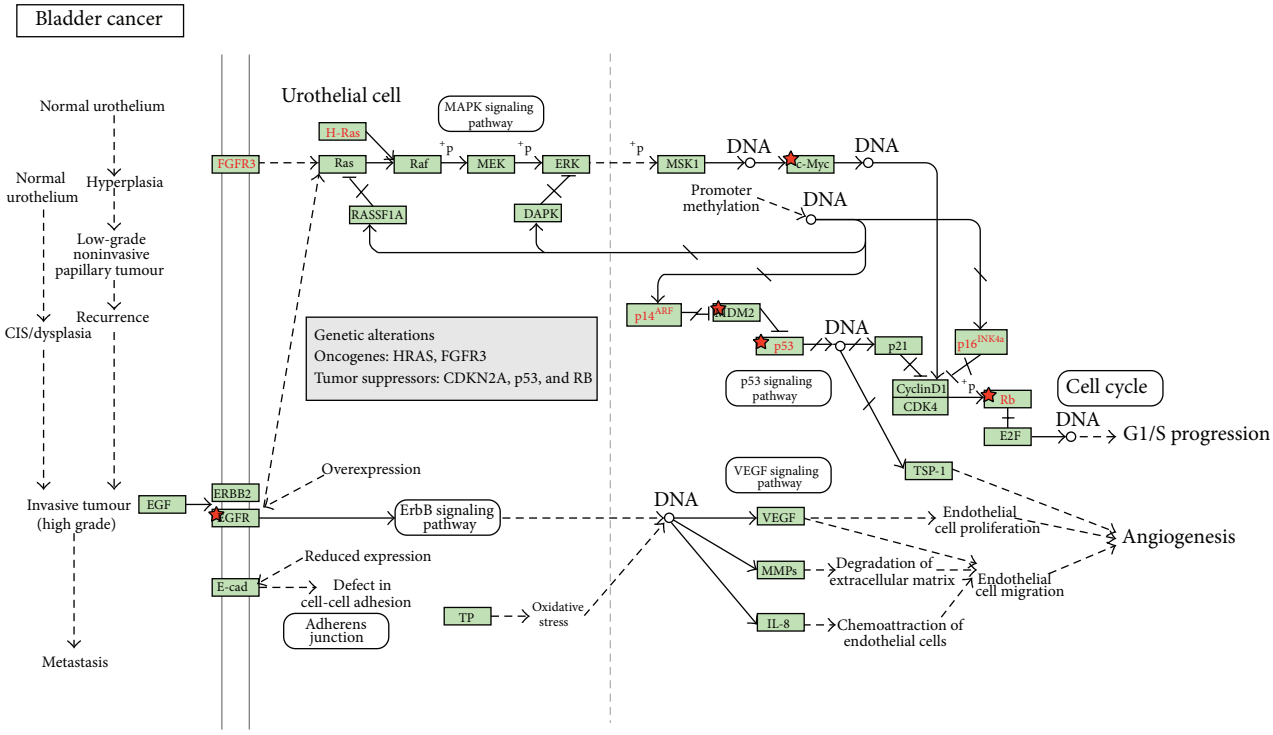
chronic myeloid leukemia, 5 genes involved in small cell lung cancer, 4 genes involved in bladder cancer, and 3 genes involved in thyroid cancer, respectively (Table 3). The four genes of hsa05219 involved in bladder cancer (TP53, MDM2, RN1, and MYC) are principal genes altered in urothelial carcinoma, which is highly related to metastatic bladder cancer and are significant targets of metastatic bladder cancer therapies [30] (Figure 3(b)). Thus, we now note that the 152 candidate proteins are not only related to bladder cancer, but also to other cancers and chronic myeloid leukemia. This would mean that common mechanisms exist between the development of the different cancers in the early stage of carcinogenesis.

Next, we proceeded to analyze the important pathways related to early stage bladder cancer (Table 4). Firstly, the cell cycle is composed of two consecutive periods (Figure 3(c)) characterized by DNA replication, sequential differentiation, and segregation of replicated chromosomes into two separate daughter cells. Both positive-acting and negative-acting proteins control the cells' entry and advancement through the cell cycle, which is composed of four distinct phases: G1 (Gap 1), S (synthesis), G2 (Gap 2), and M (mitosis) [31]. The G1 phase, where the cell grows in size, acts as a quality control check to determine whether the cell is ready to divide. The S phase is where the cell copies its DNA. The G2 phase involves cell checking as to whether all of its DNA has been correctly copied. The M phase is the cell division phase where the cell divides in two. Find out more about how cells prepare to divide and then share out their DNA and split in two. There are many reported discussions in regards to the cell cycle regulators and checkpoint functions involved in bladder cancer [32, 33]. Dysregulation of the cell cycle governs deviant cell proliferation in cancer. Losing the ability to control cell cycle checkpoints induces abnormal genetic instability. This may be due to the activation of tumorigenic mutations, which have been recognized in various tumors at different levels in the mitogenic signal transduction pathways: (1) ligands and receptors (receptor mutations of HER2/neu [ErB2] or the amplification of the HER2 gene), (2) downstream signal transduction networks (Raf/Ras/MAPK or PI3K-AKT-mTOR), and (3) regulatory genes of the cell cycle (cyclin D1/CDK4, CDK6, and cyclin E/CDK2) [34]. Increasing evidence convincingly implicates aberrant expression of cell cycle regulators in multiple cancers. Especially the restriction point (R) is the so-called G1 checkpoint. It separates the cell cycle into a mitogen-dependent phase and a growth factor-independent phase from the commitment to enter S phase. The G1 checkpoint commitment process integrates various and complex extracellular and intracellular signal transduction into the cell nucleus. Any malfunction of the G1 checkpoint may result in uncontrolled cell proliferation or genetic instability, possibly the origin of cancer or other diseases development [35].

The Wnt/β-catenin signaling pathways (Figure 3(d)) are composed of many functional networks, including a bundle of signaling pathways consisting of various proteins that transduce signals from the outside of a cell through the receptors on the cell surface and into the cell interior. They contribute significantly to the developmental process, particularly to direct cell attachment and proliferation. They are one of the most powerful signaling pathways and play critical roles in human development by controlling the genetic

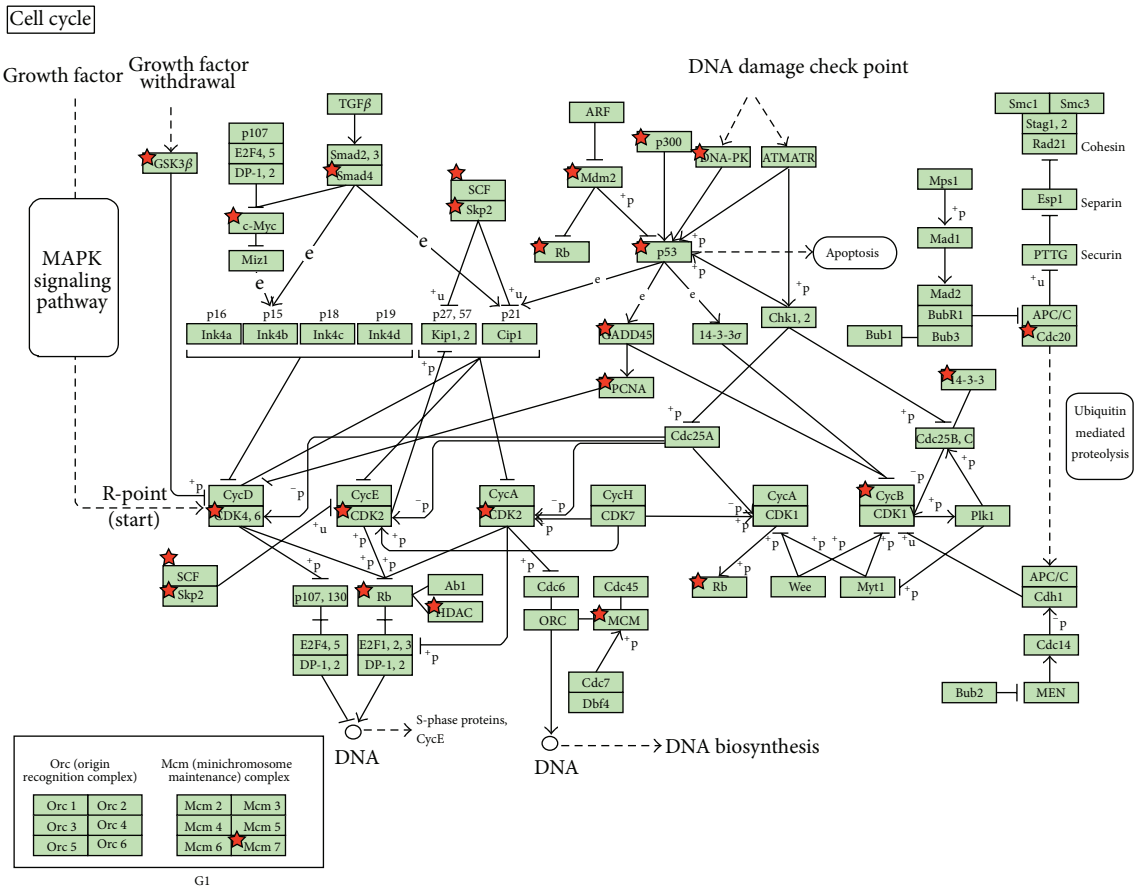


(a) The proteins in the early stage bladder cancer network marker are enriched in "hsa05200:Pathways in cancer" (Rank 2 in Table 4)

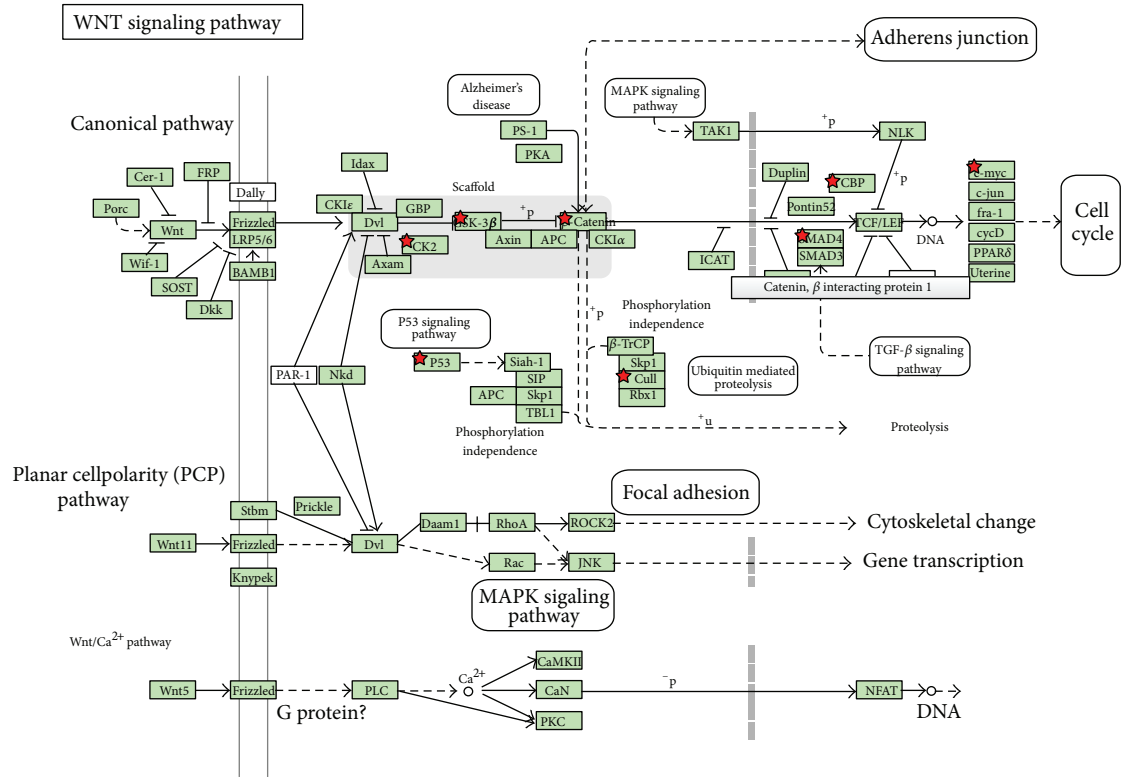


(b) The proteins in the early stage bladder cancer network marker are enriched in "hsa05219:Bladder cancer" (Rank 7 in Table 4)

Figure 3: Continued.

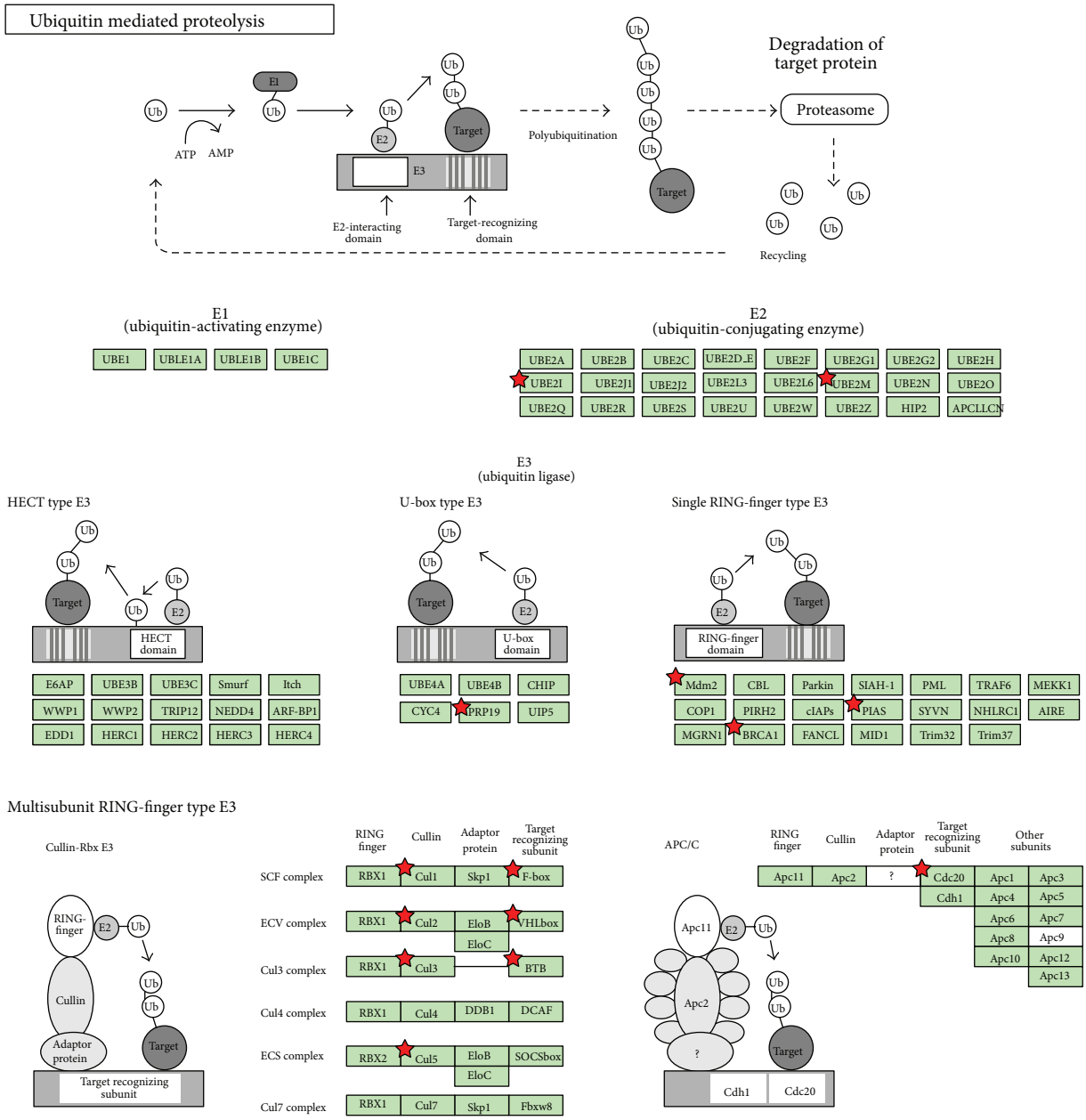


(c) The proteins in the early stage bladder cancer network marker are enriched in "hsa04110:Cell cycle" (Rank 1 in Table 4)



(d) The proteins in the early stage bladder cancer network marker are enriched in "hsa04110:Wnt signaling pathway" (Rank 5 in Table 4)

Figure 3: Continued.

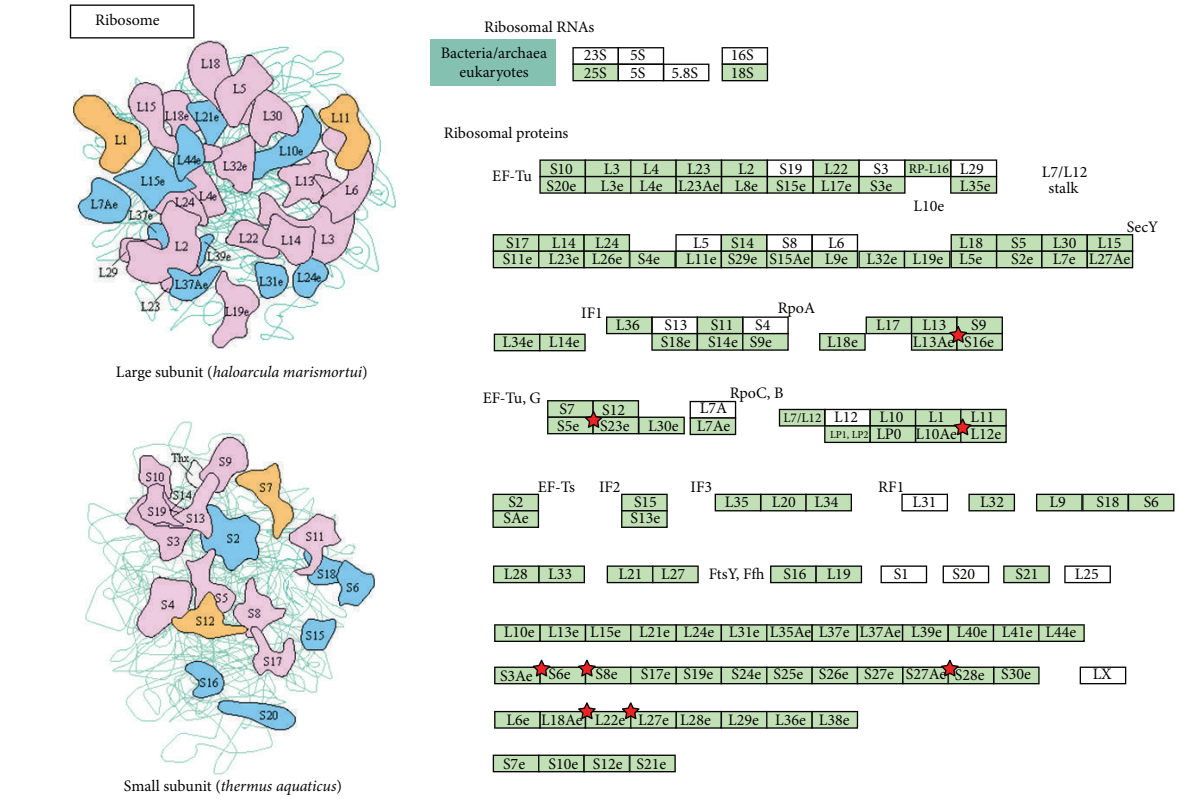


(e) The proteins in the early stage bladder cancer network marker are enriched in "hsa04120:Ubiquitin mediated proteolysis pathway" (Rank 13 in Table 4)

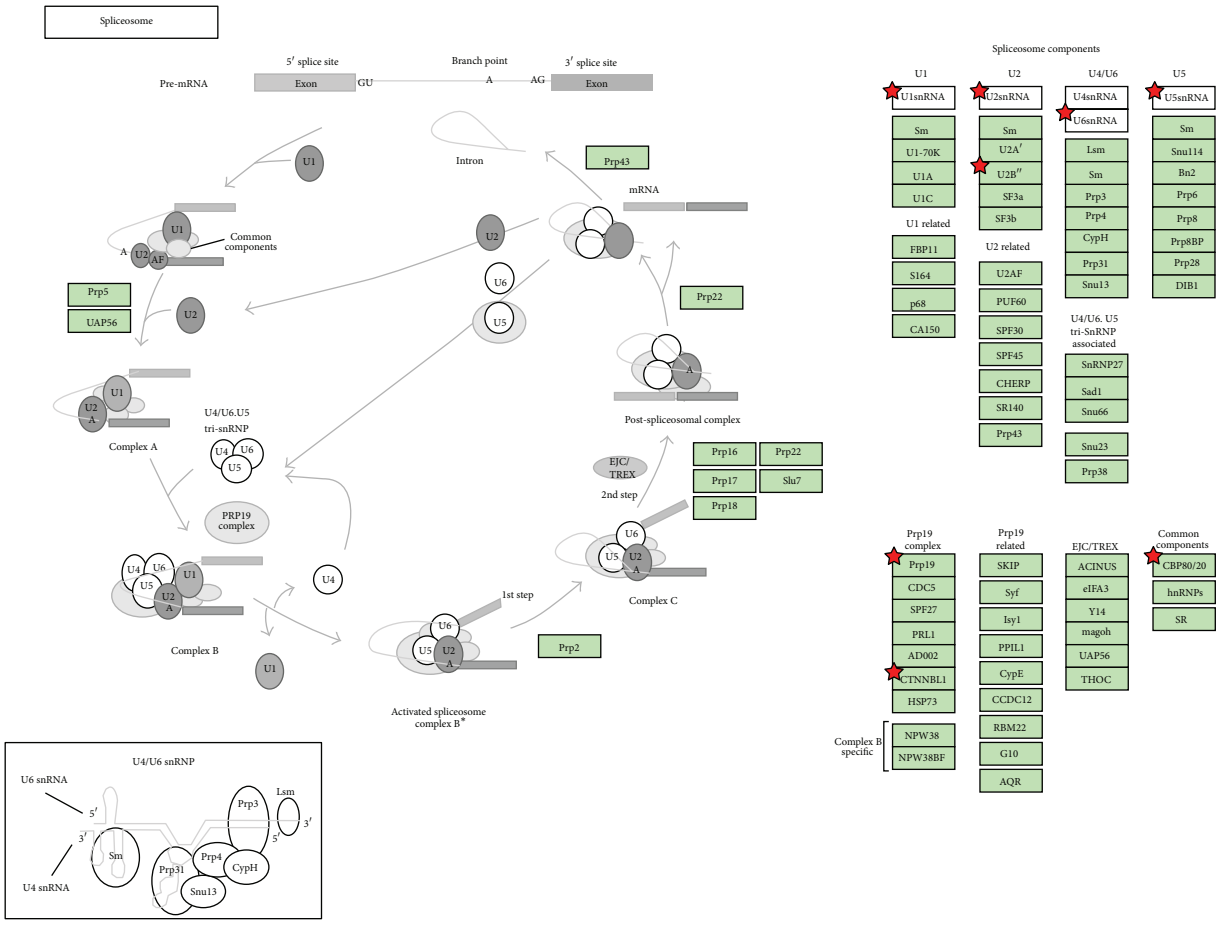
Figure 3: Overview of significant pathways in network marker of early stage bladder cancer. Among these KEGG pathways via DAVID tool (Table 4) showing a significant association with specific proteins of early stage bladder cancer, these molecular pathways are entitled with P value ≤ 0.05. It shows that these pathways are identified to play an important role in the carcinogenesis mechanism of early stage bladder cancer. The proteins in network markers of early stage bladder cancer highlighted by stars show potential targets in the pathways. Due to the different naming system, the same proteins in both these tables and in our text show the different names.

programs of embryonic development and adult homeostasis [36]. Under normal conditions, the Wnt signaling pathway is critical for healthy and normal development, while in adult cells, a dysregulated Wnt signaling pathway can lead to tumorigenesis. For this purpose, cancer cells must have the ability to switch from quiescent mode to proliferation mode, as well as switching between cell proliferation and cell invasion modes. Therefore, the Wnt signaling pathway participates in each of the stages of malignant cancer development and clearly contributes to human tumor progression. Much research has been reported on the relationship between Wnt signaling pathways and urological cancers (including bladder cancer) [37, 38].

Other pathways identified in early stage bladder cancer, such as the Notch signaling pathway, adherens junctions, the TGF-β signaling pathway, ubiquitin-mediated proteolysis

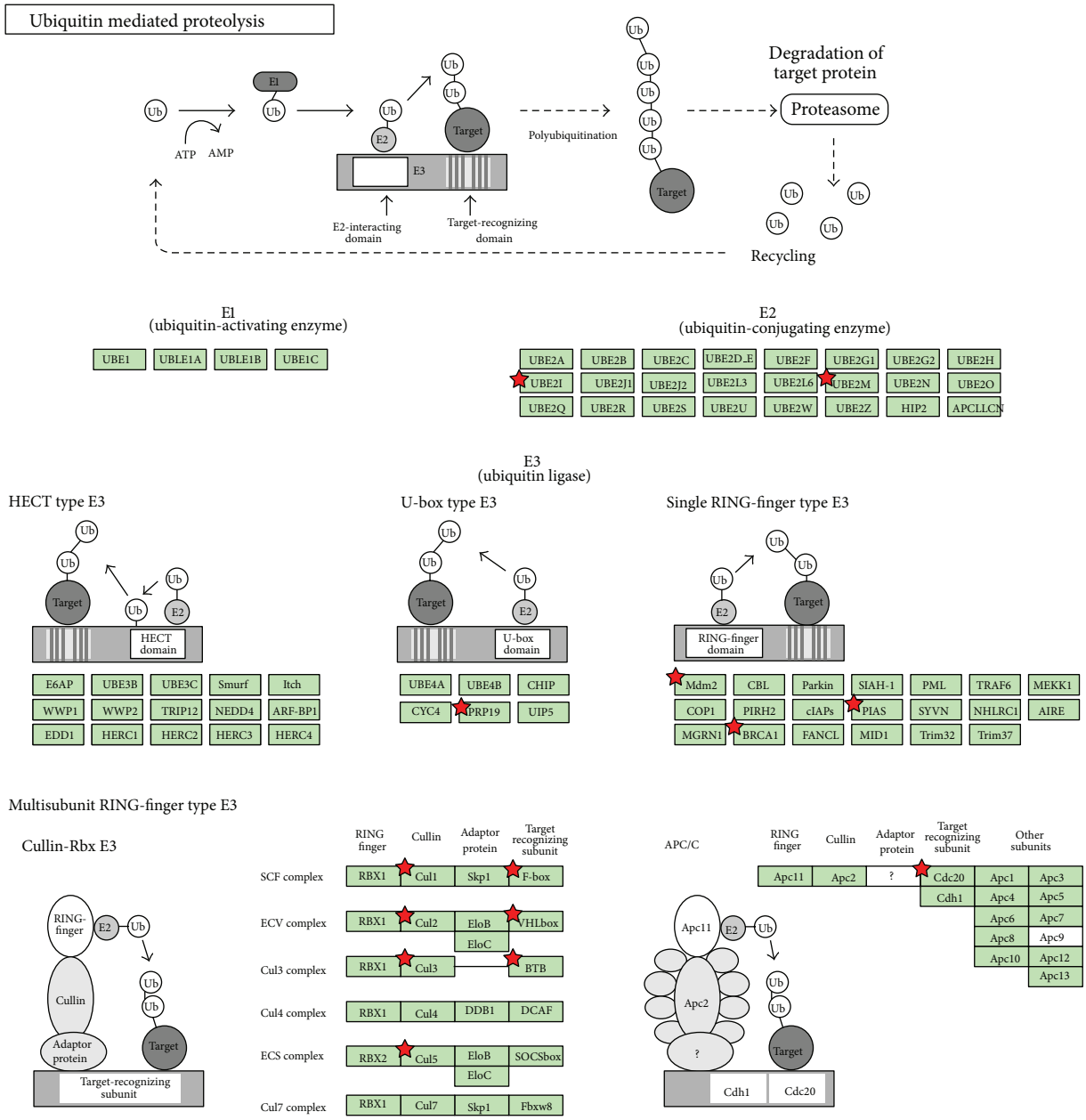


(a) The proteins in the late stage bladder cancer network marker are enriched in "hsa03010:Ribosome" (Rank 1 in Table 5)



(b) The proteins in the late stage bladder cancer network marker are enriched in "hsa03040:Ribosome" (Rank 2 in Table 5)

Figure 4: Continued.



(c) The proteins in the early stage bladder cancer network marker are enriched in "hsa04120:Ubiquitin mediated proteolysis pathway" (Rank 3 in Table 5)

FIGURE 4: Overview of significant pathways in network marker of late stage bladder cancer. Among these KEGG pathways via DAVID tool (Table 5) showing a significant association with specific proteins of late stage bladder cancer, these molecular pathways are entitled with P value ≤ 0.05. It shows that these pathways are identified to play an important role in the carcinogenesis mechanism of late stage bladder cancer. The proteins in network markers of late stage bladder cancer highlighted by stars show potential targets in the pathways. Due to the different naming system, the same proteins in both these tables and in our text show the different names.

(Figures 3(e) and 4(c)), and the p53 signaling pathway are also associated with cancer [39–43].

The NOA analysis results of the pathway and gene enrichment analysis of the early stage bladder cancer is shown in Table 4(b): (1) Biological processes (2) Cellular components (3) Molecular functions. We saw that most of the biological processes are related to the metabolic processes. Second, about the cellular components, there are three of them related to the ribosome. Finally, about the molecular functions, there are RNA binding, heparin binding and cyclin binding, which are very different from the late stage bladder cancer.

Table 4: (a) The pathways analysis for 152 early stage significant proteins in carcinogenesis. (b) The pathway analysis and gene set enrichment analysis of the top 20 proteins of early stage bladder cancer on (1) biological processes, (2) cellular components, and (3) molecular functions by NOA.

(a)

Rank	Term	Count	Symbol	*P* value
1	hsa04110:Cell cycle	13	YWHAZ, CREBBP, TP53, PRKDC, RB1, CDK2, HDAC2, EP300, HDAC1, PCNA, MDM2, MYC, andCUL1	1.50*E* − 14
2	hsa05200:Pathways in cancer	11	TRAF2, EP300, HDAC2, HDAC1, CREBBP, TP53, MDM2, RB1, MYC, CDK2, and CTNNB1	3.51*E* − 07
3	hsa05215:Prostate cancer	7	EP300, CREBBP, TP53, MDM2, RB1, CDK2, and CTNNB1	1.45*E* − 06
4	hsa05220:Chronic myeloid leukemia	6	HDAC2, HDAC1, TP53, MDM2, RB1, and MYC	1.32*E* − 05
5	hsa04310:Wnt signaling pathway	6	EP300, CREBBP, TP53, MYC, CUL1, and CTNNB1	3.78*E* − 04
6	hsa05222:Small cell lung cancer	5	TRAF2, TP53, RB1, MYC, and CDK2	4.04*E* − 04
7	hsa05219:Bladder cancer	4	TP53, MDM2, RB1, and MYC	7.24*E* − 04
8	hsa04330:Notch signaling pathway	4	EP300, HDAC2, HDAC1, and CREBBP	0.001008
9	hsa04520:Adherens junction	4	EP300, CREBBP, SRC, and CTNNB1	0.00418
10	hsa04350:TGF-beta signaling pathway	4	EP300, CREBBP, MYC, and CUL1	0.005889
11	hsa05016:Huntington's disease	5	EP300, HDAC2, HDAC1, CREBBP, and TP53	0.006736
12	hsa05216:Thyroid cancer	3	TP53, MYC, and CTNNB1	0.00676
13	hsa04120:Ubiquitin mediated proteolysis	4	CUL3, MDM2, BRCA1, and CUL1	0.020254
14	hsa05213:Endometrial cancer	3	TP53, MYC, and CTNNB1	0.020793
15	hsa05214:Glioma	3	TP53, MDM2, and RB1	0.029762
16	hsa04115:p53 signaling pathway	3	TP53, MDM2, and CDK2	0.034267
17	hsa05218:Melanoma	3	TP53, MDM2, and RB1	0.037091
18	**hsa05210:Colorectal cancer**	**3**	**TP53, MYC, and CTNNB1**	**0.050311**
19	**hsa04210:Apoptosis**	**3**	**TRAF2, IRAK1, and TP53**	**0.053574**
20	**hsa03450:Non-homologous end-joining**	**2**	**XRCC6and PRKDC**	**0.05487**
21	**hsa04916:Melanogenesis**	**3**	**EP300, CREBBP, and CTNNB1**	**0.067353**
22	**hsa04114:Oocyte meiosis**	**3**	**YWHAZ, CDK2, and CUL1**	**0.080913**
23	**hsa04722:Neurotrophin signaling pathway**	**3**	**IRAK1, YWHAZ, and TP53**	**0.099295**

The significant pathways via DAVID Bioinformatics database are selected for the 152 significant proteins in carcinogenesis. Black background indicates *P* value > 0.05.

(b)

GO:term	*P* value	Corrected *P* value	*R*	*T*	*G*	*O*	Term name
(1) Biological processes							
0044260	5.3*E* − 11	1.7*E* − 8	14791	19	3428	18	Cellular macromolecule metabolic process

(b) Continued.

GO:term	P value	Corrected P value	R	T	G	O	Term name
0043170	$7.3E-10$	$2.4E-7$	14791	19	3975	18	Macromolecule metabolic process
0044237	$3.6E-8$	$1.2E-5$	14791	19	4963	18	Cellular metabolic process
0006414	$1.4E-7$	$4.7E-5$	14791	19	101	5	Translational elongation
0019538	$1.0E-6$	$3.2E-4$	14791	19	2528	13	Protein metabolic process
0008152	$1.1E-6$	$3.6E-4$	14791	19	6033	18	Metabolic process
0044238	$1.7E-6$	$5.6E-4$	14791	19	5258	17	Primary metabolic process
0016071	$4.4E-6$	0.0014	14791	19	364	6	mRNA metabolic process
0044267	$3.3E-5$	0.0110	14791	19	1883	10	Cellular protein metabolic process
0009987	$1.2E-4$	0.0406	14791	19	9216	19	Cellular process
			(2) Cellular components				
0030529	$7.9E-10$	$7.8E-8$	16768	18	510	9	Ribonucleoprotein complex
0032991	$1.8E-7$	$1.8E-5$	16768	18	3312	14	Macromolecular complex
0043228	$1.2E-6$	$1.2E-4$	16768	18	2051	11	Non-membrane-bounded organelle
0043232	$1.2E-6$	$1.2E-4$	16768	18	2051	11	Intracellular non-membrane-bounded organelle
0005840	$1.5E-6$	$1.5E-4$	16768	18	196	5	Ribosome
0005829	$2.0E-6$	$2.0E-4$	16768	18	1269	9	Cytosol
0043229	$8.3E-6$	$8.2E-4$	16768	18	8759	18	Intracellular organelle
0043226	$8.5E-6$	$8.4E-4$	16768	18	8773	18	Organelle
0044445	$1.7E-5$	0.0016	16768	18	150	4	Cytosolic part
0033279	$2.8E-4$	0.0287	16768	18	123	3	Ribosomal subunit
			(3) Molecular functions				
0003735	$1.0E-6$	$1.1E-4$	15767	19	161	5	Structural constituent of ribosome
0003723	$1.8E-4$	0.0193	15767	19	755	6	RNA binding
0005198	$8.0E-4$	0.0825	15767	19	643	5	Structural molecule activity
0031625	0.0013	0.1360	15767	19	45	2	Ubiquitin protein ligase binding
0003678	0.0013	0.1360	15767	19	45	2	DNA helicase activity
0004535	0.0024	0.2480	15767	19	2	1	Poly(A)-specific ribonuclease activity
0033130	0.0048	0.4956	15767	19	4	1	Acetylcholine receptor binding
0008201	0.0079	0.8140	15767	19	112	2	Heparin binding
0030332	0.0096	0.9890	15767	19	8	1	Cyclin binding
0004386	0.0129	1	15767	19	145	2	Helicase activity

R: number of genes in reference set.
T: number of genes in test set.
G: number of genes annotated by given term in reference set.
O: number of genes annotated by given term in test set.

3.4. Pathway Analysis of Late Stage Bladder Cancer. The most important results in this study as compared to our previous work are that we reveal related pathways of late stage bladder cancer in comparison to early stage cancer to reveal the evolution of network biomarkers in the carcinogenesis process. From Table 5, we observed that only three pathways, ribosome, spliceosome, and ubiquitin-mediated proteolysis pathways, were hit by the 50 candidate proteins identified in late stage bladder cancer. This is indicative of the evolution of cancer mechanisms from early stage bladder cancer.

The nucleolus is the site of ribosome biogenesis (Figure 4(a)). Due to the higher concentration of both RNA and proteins in the nucleolus than in the nucleoplasm, the nucleolus is easily detected by microscopy in living cells. From electron microscopy images, three major components were constantly exhibited by mammalian cells. They include fibrillar centers (FCs), which appear as surrounding structures of various sizes, with a very low electron opacity; the dense fibrillar component (DFC), which always constitutes a rim intimately accompanied with the fibrillar centers, composed of densely packed fibrils; and the granular component (GC), which is composed of granules that surround the fibrillar components. There is evidence that changes in nucleolar morphology and function may depend on both the rate and status of ribosome biogenesis and on the proliferative activity of cycling cells [44]. In

Table 5: (a) The pathways analysis for 50 significant proteins in late stage bladder cancer carcinogenesis. (b) The pathway analysis and gene set enrichment analysis of the top 20 proteins of late stage bladder cancer on (1) biological processes, (2) cellular components and (3) molecular functions by NOA.

(a)

Rank	Term	Count	Symbol	*P* value
1	hsa03010:Ribosome	8	RPS28, RPS16, RPL22, RPL27, RPL12, RPS6, RPS8, and RPS23	$2.26E-07$
2	hsa03040:Spliceosome	5	HSPA1L, CDC5L, SF3A1, SNRPE, and HNRNPU	0.004054716
3	hsa04120:Ubiquitin mediated proteolysis	4	CUL3, CUL5, BRCA1, and CUL1	0.034906958

The significant pathways via DAVID Bioinformatics database are selected for the 50 significant proteins in carcinogenesis.

(b)

GO:term	*P* value	Corrected *P* value	*R*	*T*	*G*	*O*	Term name
(1) Biological processes							
GO:0045786	$1.8E-6$	0.0010	14791	18	178	5	Negative regulation of cell cycle
GO:0022402	$2.4E-6$	0.0014	14791	18	562	7	Cell cycle process
GO:0007050	$8.4E-6$	0.0049	14791	18	111	4	Cell cycle arrest
GO:0051726	$8.5E-6$	0.0049	14791	18	435	6	Regulation of cell cycle
GO:0060710	$1.3E-5$	0.0080	14791	18	5	2	Chorioallantoic fusion
GO:0044260	$1.4E-5$	0.0081	14791	18	3428	13	Cellular macromolecule metabolic process
GO:0051052	$1.5E-5$	0.0091	14791	18	130	4	Regulation of DNA Metabolic process
GO:0008629	$2.6E-5$	0.0155	14791	18	49	3	Induction of apoptosis by intracellular signals
GO:0006917	$2.8E-5$	0.0162	14791	18	313	5	Induction of apoptosis
GO:0012502	$2.8E-5$	0.0165	14791	18	314	5	Induction of programmed cell death
(2) Cellular components							
GO:0032991	$5.2E-10$	$5.5E-8$	16768	18	3312	16	Macromolecular complex
GO:0005829	$1.4E-7$	$1.5E-5$	16768	18	1269	10	Cytosol
GO:0043234	$2.5E-6$	$2.6E-4$	16768	18	2748	12	Protein complex
GO:0005654	$6.1E-6$	$6.4E-4$	16768	18	465	6	Nucleoplasm
GO:0044428	$6.4E-5$	0.0067	16768	18	1932	9	Nuclear part
GO:0000307	$9.8E-5$	0.0102	16768	18	14	2	Cyclin-dependent protein kinase holoenzyme complex
GO:0030529	$1.5E-4$	0.0163	16768	18	510	5	Ribonucleoprotein complex
GO:0031461	$4.9E-4$	0.0516	16768	18	31	2	Cullin-RING ubiquitin ligase complex
GO:0022627	$7.4E-4$	0.0777	16768	18	38	2	Cytosolic small ribosomal subunit
GO:0043626	0.0010	0.1116	16768	18	1	1	PCNA complex
(3) Molecular functions							
GO:0019899	$3.1E-5$	0.0037	15767	18	584	6	Enzyme binding
GO:0005515	$1.1E-4$	0.0129	15767	18	8097	17	Protein binding
GO:0030337	0.0011	0.1335	15767	18	1	1	DNA polymerase processivity factor activity
GO:0000701	0.0011	0.1335	15767	18	1	1	Purine-specific mismatch base pair DNA N-glycosylase activity
GO:0031625	0.0011	0.1384	15767	18	45	2	Ubiquitin protein ligase binding
GO:0000166	0.0021	0.2510	15767	18	2283	8	Nucleotide binding
GO:0035033	0.0022	0.2669	15767	18	2	1	Histone deacetylase regulator activity

(b) Continued.

GO:term	*P* value	Corrected *P* value	*R*	*T*	*G*	*O*	Term name
GO:0004696	0.0022	0.2669	15767	18	2	1	Glycogen synthase kinase 3 activity
GO:0000700	0.0022	0.2669	15767	18	2	1	Mismatch base pair DNA N-glycosylase activity
GO:0005200	0.0031	0.3705	15767	18	74	2	Structural constituent of cytoskeleton

R: number of genes in reference set.
T: number of genes in test set.
G: number of genes annotated by given term in reference set.
O: number of genes annotated by given term in test set.

Table 6: The pathways analysis for 18 significant proteins in early and late stage bladder cancer carcinogenesis.

Rank	Term	Count	Symbol	*P* value
1	hsa03010:Ribosome	4	CUL3, CUL5, BRCA1, and CUL1	$1.4E-3$

cancer cells the upregulated ribosome biogenesis leads to an increased demand of ribosomal proteins for rRNA binding. In this way, after ribosome biogenesis alterations, cycling cells can activate the p53 pathway to ensure cell cycle arrest or alternatively to start the apoptotic program [45]. According to our analysis, there were eight significant proteins in the late stage cancer to hit the ribosome pathway.

Alternative splicing is a modification of the premessenger RNA (pre-mRNA) transcript in which internal noncoding regions of pre-mRNA (introns) are removed and then the remaining segments (exons) are joined (Figure 4(b)). The formation of mature messenger RNA (mRNA) is subsequently capped at its 5′ end and polyadenylated at its 3′ end, and transported out of the nucleus to be translated into protein in the cytoplasm. Most genes use alternative splicing to generate multiple spliced transcripts. These transcripts contain various combinations of exons resulting from different mRNA variants and then are synthesized as protein isoforms. The exons are always around 50–250 base pairs, whereas introns could be as long as several thousands of base pairs. For nuclear encoded genes, splicing takes place within the nucleus after or simultaneously with transcription. Splicing is necessary for the eukaryotic messenger RNA (mRNA) before it can be translated into a correct protein. The spliceosome is a dynamic intracellular macromolecular complex of multiple proteins and ribonucleoproteins (snRNPs). For many eukaryotic introns, the spliceosome carries out the two main functions of alternative splicing. First, it recognizes the intron-exon boundaries and second it catalyzes the cut-and-paste reactions that remove introns and concatenate exons. The various spliceosomal machinery complex is formed from 5 ribonucleo-protein (RNP) subunits, termed uridine-rich (U-rich) small nuclear RNP (snRNP), transiently associated with more than 760 non-snRNPs splicing factors (RNA helicases, SR splicing factors, etc.) [46, 47]. Each spliceosomal snRNP (U1, U2, U4, U5, and U6) consists of a uridine-rich small nuclear RNA (snRNA) complexed with a set of seven proteins known as canonical Sm core or SNRP proteins. The seven Sm proteins (B/B′, D1, D2, D3, E, F, and G) form a core ring structure that surrounds the RNA. All Sm proteins contain a conserved sequence motif in two segments (Sm1 and Sm2) that are responsible for the assembly and ordering of the snRNAs. They form the Sm core of the spliceosomal snRNPs [48] and process the pre-mRNA [49]. Spliceosomes not only catalyze splicing by a series of reactions, but they are also the main cellular machinery that guides splicing. Recently, scientists have found two natural compounds that can interfere with spliceosome function that also display anticancer activity in vitro and in vivo [50, 51]. Therefore, it is believable that inhibiting the spliceosome could act as a new target for anticancer drug development [52], and it should be validated in vivo or in vitro in the future.

The NOA analysis results of the pathway and gene enrichment analysis of the late stage bladder cancer is shown in Table 5(b): (1) Biological processes (2) Cellular components (3) Molecular functions. We saw most of the biological processes are related to cell cycle, which are different from the metabolic processes of early stage. Second, about the cellular components, there are complex evolution behaviors of the network compared with the early stage bladder cancer; there is only one intersection of these two stages that is ribonucleoprotein complex. It gives us many clues to develop evolutionary strategies for cancer target therapy. Finally, about the molecular functions, there are enzyme binding, protein binding and nucleotide binding, which are very different from the early stage bladder cancer. All the evolutionary behaviors from early to late stage bladder cancer let us reveal more hidden carcinogenesis mechanism.

3.5. Pathway Analysis of Both Early and Late Stage Bladder Cancer. The only pathway to intersect between early and late stage bladder cancer is the ubiquitin-mediated proteolysis pathway (Table 6). This means it is the only housekeeping pathway for bladder cancer and that the mechanisms of early and late stage bladder cancer are completely different. We hypothesize that this may be a novel concept for target therapy. Various other researches have never built a model in accordance with the network markers at the different stages of cancer. Our results show that the network markers of early stage hit common mechanisms and fundamental pathways, such as cell cycle, cell proliferation, and Wnt signaling, among others, which are implicated in various cancers. These provide

clues in that early stage bladder cancer is active in many related pathways and we can assume that it is an active process to change the cell. In contrast, in the late stage of bladder cancer, the cells were inactive and close to silence. This may mean that the cells are close to death. Should we attempt to save these cells, we should aim to focus on the ribosome and spliceosome pathways. Of course ubiquitin-mediated proteolysis pathways are both active in early and late stage cancer.

4. Conclusions

Bladder cancer is among the 10 most common forms of carcinoma in the USA and worldwide. It is a lethal disease like other cancers and understanding the carcinogenesis mechanism can help to develop new therapeutic strategy. Identifying the PPI interface to develop small molecule inhibitors has become a new direction for targeted cancer therapy. This study, which follows from our prior work, analyzes the carcinogenesis mechanism from early to late stage bladder cancer using a network-based biomarker evolution approach. Other research studies do not distinguish network markers between these two stages of bladder cancer. Thus, our approach is advantageous in that it can provide added insight into the significant network marker evolution of the carcinogenesis process of bladder cancer. The network markers and their related pathways identified in early stage bladder cancer are mostly related to ordinary cancer mechanisms, which just show a highly active state of the early stage and cannot reveal additional novel results. All of these results should be validated in vivo or in vitro in the future. However, from the two specific and significant pathways identified in late stage bladder cancer, ribosome pathway and spliceosome pathway, we identified a novel result, which has potential to become a target for cancer therapy. The only core pathway in these two stages is the ubiquitin-mediated proteolysis pathway, which is a significant cue of carcinogenesis from early to late stage bladder cancer. Applying our method to study more cancers and more classification groups (such as stage, age, ethics, and sex) will give us further insight into the various pathogenesis mechanisms.

Conflict of Interests

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

Acknowledgment

The authors are grateful for the support provided by the Ministry of Science and Technology (NSC-102-2745-E-007-001-ASP).

References

[1] T. N. Seyfried and L. M. Shelton, "Cancer as a metabolic disease," *Nutrition & Metabolism*, vol. 7, article 7, 2010.

[2] http://www.cancer.gov/cancertopics/types/bladder.

[3] D. S. Kaufman, W. U. Shipley, and A. S. Feldman, "Bladder cancer," *The Lancet*, vol. 374, no. 9685, pp. 239–249, 2009.

[4] A. M. Donaldson, J. G. Gonzalez, M. K. B. Parmar, and N. Donaldson, "The MRC superficial bladder cancer trial of intravesical mytomicin-c after complete surgical resection. Sequential statistical methods applied to survival data from a randomised clinical trial," *International Journal of Surgery*, vol. 7, no. 5, pp. 441–445, 2009.

[5] V. H. Cowling and M. D. Cole, "Mechanism of transcriptional activation by the Myc oncoproteins," *Seminars in Cancer Biology*, vol. 16, no. 4, pp. 242–252, 2006.

[6] M. Pacal and R. Bremner, "Insights from animal models on the origins and progression of retinoblastoma," *Current Molecular Medicine*, vol. 6, no. 7, pp. 759–781, 2006.

[7] J. N. Contessa, J. Hampton, G. Lammering et al., "Ionizing radiation activates Erb-B receptor dependent Akt and p70 S6 kinase signaling in carcinoma cells," *Oncogene*, vol. 21, no. 25, pp. 4032–4041, 2002.

[8] A. M. Codegoni, M. I. Nicoletti, G. Buraggi et al., "Molecular characterisation of a panel of human ovarian carcinoma xenografts," *European Journal of Cancer*, vol. 34, no. 9, pp. 1432–1438, 1998.

[9] T. R. Golub, D. K. Slonim, P. Tamayo et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–527, 1999.

[10] H. Han, D. J. Bearss, L. W. Browne, and et al, "Identification of differentially expressed genes in pancreatic cancer cells using cDNA microarray," *Cancer Research*, vol. 62, pp. 2890–2896, 2002.

[11] K.-Q. Liu, Z.-P. Liu, J.-K. Hao, L. Chen, and X.-M. Zhao, "Identifying dysregulated pathways in cancers from pathway interaction networks," *BMC Bioinformatics*, vol. 13, no. 1, article 126, 2012.

[12] H. Uramoto, K. Sugio, T. Oyama et al., "Expression of the p53 family in lung cancer," *Anticancer Research*, vol. 26, no. 3, pp. 1785–1790, 2006.

[13] E. A. Horvat, J. D. Zhang, U. Stefan, O. Sahin, and K. A. .Zweig, "A network-based method to assess the statistical significance of mild co-regulation effects," *PLoS ONE*, vol. 8, Article ID e73413, 2013.

[14] Y.-C. Wang and B.-S. Chen, "A network-based biomarker approach for molecular investigation and diagnosis of lung cancer," *BMC Medical Genomics*, vol. 4, article 2, 2011.

[15] A. A. Ivanov, F. R. Khuri, and H. Fu, "Targeting protein-protein interactions as an anticancer strategy," *Trends in Pharmacological Sciences*, vol. 34, no. 7, pp. 393–400, 2013.

[16] L. Chen, R. Liu, Z.-P. Liu, M. Li, and K. Aihara, "Detecting early-warning signals for sudden deterioration of complex diseases by dynamical network biomarkers," *Scientific Reports*, vol. 2, article 342, 2012.

[17] R. Liu, X. Wang, K. Aihara, and L. Chen, "Early diagnosis of complex diseases by molecular biomarkers, network biomarkers, and dynamical network biomarkers," *Medicinal Research Reviews*, vol. 34, pp. 455–478, 2014.

[18] R. Liu, X. Yu, X. Liu, D. Xu, K. Aihara, and L. Chen, "Identifying critical transitions of complex diseases based on a single sample," *Bioinformatics*, vol. 30, no. 11, pp. 1579–1586, 2014.

[19] C. Zhang, J. Wang, K. Hanspers, D. Xu, L. Chen, and A. R. Pico, "NOA: a cytoscape plugin for network ontology analysis," *Bioinformatics*, vol. 29, no. 16, pp. 2066–2067, 2013.

[20] W.-J. Kim, E.-J. Kim, S.-K. Kim et al., "Predictive value of progression-related gene classifier in primary non-muscle invasive bladder cancer," *Molecular Cancer*, vol. 9, article 3, 2010.

[21] A. Chatr-Aryamontri, B.-J. Breitkreutz, S. Heinicke et al., "The BioGRID interaction database: 2013 update," *Nucleic Acids Research*, vol. 41, no. 1, pp. D816–D823, 2013.

[22] J. M. Bland and D. G. Altman, "Multiple significance tests: the Bonferroni method," *British Medical Journal*, vol. 310, article 170, no. 6973, 1995.

[23] R. Johansson, *System Modeling and Identification*, Prentice Hall, 1993.

[24] M. Pagano and K. Gauvreau, *Principles of Biostatistics*, 2000.

[25] M. Kanehisa, "Molecular network analysis of diseases and drugs in KEGG," *Methods in Molecular Biology*, vol. 939, pp. 263–275, 2013.

[26] J.-I. Satoh, "Molecular network of microRNA targets in Alzheimer's disease brains," *Experimental Neurology*, vol. 235, no. 2, pp. 436–446, 2012.

[27] 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.

[28] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists," *Nucleic Acids Research*, vol. 37, no. 1, pp. 1–13, 2009.

[29] J. Wang, Q. Huang, Z.-P. Liu et al., "NOA: a novel Network Ontology Analysis method," *Nucleic Acids Research*, vol. 39, no. 13, p. e87, 2011.

[30] M. Fassan, E. J. Trabulsi, L. G. Gomella, and R. Baffa, "Targeted therapies in the management of metastatic bladder cancer," *Biologics*, vol. 1, no. 4, pp. 393–406, 2007.

[31] J. Camacho, *Molecular Ontology: Principles and Recent Advances*, 2012.

[32] P. Korkolopoulou, P. Christodoulou, A.-E. Konstantinidou, E. Thomas-Tsagli, P. Kapralos, and P. Davaris, "Cell cycle regulators in bladder cancer: a multivariate survival study with emphasis on p27Kip1," *Human Pathology*, vol. 31, no. 6, pp. 751–760, 2000.

[33] S. C. Doherty, S. R. McKeown, C. Stephen Downes et al., "Cell cycle checkpoint function in bladder cancer," *Journal of the National Cancer Institute*, vol. 95, no. 24, pp. 1859–1868, 2003.

[34] G. H. Williams and K. Stoeber, "The cell cycle and cancer," *Journal of Pathology*, vol. 226, no. 2, pp. 352–364, 2012.

[35] J. Lukas, J. Bartkova, and J. Bartek, "Convergence of mitogenic signalling cascades from diverse classes of receptors at the cyclin D-cyclin-dependent kinase-pRb-controlled G1 checkpoint," *Molecular & Cellular Biology*, vol. 16, no. 12, pp. 6917–6925, 1996.

[36] T. Grigoryan, P. Wend, A. Klaus, and W. Birchmeier, "Deciphering the function of canonical Wnt signals in development and disease: conditional loss- and gain-of-function mutations of β-catenin in mice," *Genes and Development*, vol. 22, no. 17, pp. 2308–2341, 2008.

[37] S. Majid, S. Saini, and R. Dahiya, "Wnt signaling pathways in urological cancers: past decades and still growing," *Molecular Cancer*, vol. 11, article 7, 2012.

[38] I. Ahmad, *The role of Wnt signalling in urothelial cell carcinoma [Ph.D. thesis]*, University of Glasgow, 2011.

[39] H. Lodish et al., *Molecular Cell Biology*, 2013.

[40] E. J. Allenspach, I. Maillard, J. C. Aster, and W. S. Pear, "Notch signaling in cancer," *Cancer Biology and Therapy*, vol. 1, pp. 466–476, 2002.

[41] V. Bolos, J. Grego-Bessa, and J. L. de la Pompa, "Notch signaling in development and cancer," *Endocrine Reviews*, vol. 28, no. 3, pp. 339–363, 2007.

[42] J. Schneikert and J. Behrens, "The canonical Wnt signalling pathway and its APC partner in colon cancer development," *Gut*, vol. 56, no. 3, pp. 417–425, 2007.

[43] R. Derynck, R. J. Akhurst, and A. Balmain, " TGF-β signaling in tumor suppression and cancer progression (Nature Genetics (2001) 29 (117–129))," *Nature Genetics*, vol. 29, no. 3, p. 351, 2001.

[44] L. Montanaro, D. Treré, and M. Derenzini, "Nucleolus, ribosomes, and cancer," *The American Journal of Pathology*, vol. 173, no. 2, pp. 301–310, 2008.

[45] L. Montanaro, D. Treré, and M. Derenzini, "Changes in ribosome biogenesis may induce cancer by down-regulating the cell tumor suppressor potential," *Biochimica et Biophysica Acta—Reviews on Cancer*, vol. 1825, no. 1, pp. 101–110, 2012.

[46] M. S. Jurica and M. J. Moore, "Pre-mRNA splicing: awash in a sea of proteins," *Molecular Cell*, vol. 12, no. 1, pp. 5–14, 2003.

[47] Z. Zhou, L. J. Licklider, S. P. Gygi, and R. Reed, "Comprehensive proteomic analysis of the human spliceosome," *Nature*, vol. 419, no. 6903, pp. 182–185, 2002.

[48] R. S. Pillai, M. Grimmler, G. Meister et al., "Unique Sm core structure of U7 snRNPs: assembly by a specialized SMN complex and the role of a new component, Lsm11, in histone RNA processing," *Genes and Development*, vol. 17, no. 18, pp. 2321–2333, 2003.

[49] L. Agranat-Tamir, N. Shomron, J. Sperling, and R. Sperling, "Interplay between pre-mRNA splicing and microRNA biogenesis within the supraspliceosome," *Nucleic Acids Research*, vol. 42, no. 7, pp. 4640–4651, 2014.

[50] D. Kaida, H. Motoyoshi, E. Tashiro et al., "Spliceostatin A targets SF3b and inhibits both splicing and nuclear retention of pre-mRNA," *Nature Chemical Biology*, vol. 3, no. 9, pp. 576–583, 2007.

[51] Y. Kotake, K. Sagane, T. Owa et al., "Splicing factor SF3b as a target of the antitumor natural product pladienolide," *Nature Chemical Biology*, vol. 3, no. 9, pp. 570–575, 2007.

[52] R. J. van Alphen, E. A. C. Wiemer, H. Burger, and F. A. L. M. Eskens, "The spliceosome as target for anticancer treatment," *British Journal of Cancer*, vol. 100, no. 2, pp. 228–232, 2009.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 456323, 10 pages
http://dx.doi.org/10.1155/2014/456323

Research Article

MicroRNA Expression Profiling Altered by Variant Dosage of Radiation Exposure

Kuei-Fang Lee,[1,2] Yi-Cheng Chen,[3] Paul Wei-Che Hsu,[4] Ingrid Y. Liu,[5] and Lawrence Shih-Hsin Wu[1]

[1] *Institute of Medical Sciences, Tzu Chi University, No. 701, Zhongyang Road, Section 3, Hualien 97004, Taiwan*
[2] *Laboratory for Cytogenetics, Center for Genetic Counseling, Buddhist Tzu Chi General Hospital, Hualien 97004, Taiwan*
[3] *Department of Computer Science & Information Engineering, Tamkang University, New Taipei City 25137, Taiwan*
[4] *Bioinformatics Core Laboratory, Institute of Molecular Biology, Academia Sinica, Taipei 11529, Taiwan*
[5] *Department of Molecular Biology and Human Genetics, Tzu Chi University, Hualien 97004, Taiwan*

Correspondence should be addressed to Lawrence Shih-Hsin Wu; lshwu@hotmail.com

Received 11 April 2014; Revised 14 August 2014; Accepted 16 August 2014; Published 16 September 2014

Academic Editor: Tzong-Yi Lee

Various biological effects are associated with radiation exposure. Irradiated cells may elevate the risk for genetic instability, mutation, and cancer under low levels of radiation exposure, in addition to being able to extend the postradiation side effects in normal tissues. Radiation-induced bystander effect (RIBE) is the focus of rigorous research as it may promote the development of cancer even at low radiation doses. Alterations in the DNA sequence could not explain these biological effects of radiation and it is thought that epigenetics factors may be involved. Indeed, some microRNAs (or miRNAs) have been found to correlate radiation-induced damages and may be potential biomarkers for the various biological effects caused by different levels of radiation exposure. However, the regulatory role that miRNA plays in this aspect remains elusive. In this study, we profiled the expression changes in miRNA under fractionated radiation exposure in human peripheral blood mononuclear cells. By utilizing publicly available microRNA knowledge bases and performing cross validations with our previous gene expression profiling under the same radiation condition, we identified various miRNA-gene interactions specific to different doses of radiation treatment, providing new insights for the molecular underpinnings of radiation injury.

1. Introduction

Radiation exists everywhere in our daily life. It is used in medical treatments and also utilized to generate electricity. Risk assessment of acute radiation injury caused by high-dose radiation exposure has been the focus of extensive research, but the mechanisms underlying the effect of low-dose radiation, whether short- or long-term, remain elusive [1]. Response to radiation-induced damages varies due to many confounding factors such as the immune status, age, and genetics [2]. However, in some cases, signs of radiation damage may not be immediately apparent, or not present at all.

Radiation studies primarily concentrate on examining the biological effects of radiation on cell death, chromosomal impairments, mutagenesis, carcinogenesis, and structural alterations of the cell, as well as direct or indirect damage to the DNA double helix via the production of free radical [3]. These damaging consequences of radiation exposure may require several months to years or even generations to develop [4, 5].

In addition to cellular and molecular damages caused by radiation exposure, the radiation-induced bystander effect (RIBE) is also an important topic of rigorous research. The RIBE theory describes the condition in which nonirradiated cells become irradiated by receiving radiation from neighboring irradiated cells. In other words, cells that are not directly hit by an alpha particle but are in the vicinity of one that has been hit also contribute to the genotoxic response of the cell population [6]. RIBE is also suggested

to play a role in the biological consequences of exposure to low doses of radiation [7]. Immune cells such as T-lymphocytes and dendritic cells have been particularly implicated in this process [8], though there is currently insufficient evidence to demonstrate that the bystander effect is able to promote carcinogenesis in human at low doses [1]. The mechanisms underlying the bystander effect are complex and a comprehensive understanding of this process has yet to be established [9]. It is known that irradiated cells affect nonirradiated cells through intracellular communications. Molecular signals may be transmitted from the irradiated cells to the nonirradiated ones via gap junctions between cells or through ligand-receptor interaction when the signals are secreted as soluble factors into the culture medium [10]. These transmittable factors are diverse. At present, it is not definitely established as to how many types of molecular signals are involved and to what extent they modulate the transmission of irradiation effect. Nonetheless, RIBE has clear negative implications on health. In the context of RIBE, even at low levels of radiation exposure [11], irradiated cells may still elevate the risk for genetic instability, mutation, and cancer, in addition to being able to extend the postradiation side effects in normal tissues [12]. The bystander effects seem to be tissue-specific as demonstrated by in vivo and in vitro studies [13, 14], and radiation-induced genomic instability (RIGI) as a result of RIBE may play an important role in carcinogenesis [15], though the mechanisms are less clear under fractionated irradiation. RIGI is one of the postirradiation outcomes that appear in nonirradiated progeny cells much later after the initial exposure [15]. Alterations in the DNA sequence could not explain these biological effects of radiation and it is thought that epigenetics, including DNA methylation, histone modifications, chromatin remodeling, and noncoding RNA modulation, may be involved [16].

Indeed, some microRNAs (or miRNAs) were found to be correlated with the biological effects caused by radiation in recent findings. Studies in humans [17, 18] and mice [19, 20] have revealed an association between miRNA regulation and radiation exposure that is dependent on the dosages and time after irradiation. However, the regulatory role miRNA plays in this aspect is still unclear. In the present study, we profiled the expression changes in miRNA under fractionated radiation exposure in human peripheral blood mononuclear cells. By utilizing publicly available microRNA knowledge bases and cross validating with our previous gene expression profiling under the same radiation set-up, our analysis identified specific miRNA-gene interactions characteristic of various doses of radiation treatment, providing new insights for the molecular underpinnings of radiation injury.

2. Materials and Methods

2.1. Sample Preparation. Whole blood samples (30 mL) were drawn from each of the five participants and collected into vacutainers containing sodium heparin. Samples were irradiated using ^{60}Co at a dose rate of 0.546 Gy/min (The Institute of Nuclear Energy Research (INER), Taoyuan, Taiwan). The radiation doses used in these experiments were chosen to cover a range of 0.5 Gy, 1 Gy, 2.5 Gy, and 5 Gy. The control samples were not exposed to any radiation. Samples were harvested after 24 hours of treatment with radiation [21, 22]. Informed consents were obtained from all participants. All procedures were approved by the Institutional Review Board at Tzu Chi General Hospital, Hualien, Taiwan.

2.2. RNA Preparation. Total RNA was extracted from peripheral blood mononuclear cells using Trizol (Invitrogen, Carlsbad, CA, USA) according to the manufacturer's instruction. RNA quantity and purity were assessed using NanoDrop ND-1000 (Thermo Fisher Scientific, Waltham, Massachusetts, USA). A260/A280 $\geq$ 1.6 and A260/A230 $\geq$ 1 indicate acceptable RNA purity, while acceptable RIN value $\geq$ 5 using Agilent RNA 6000 Nanoassay (Agilent Technologies, Inc., Santa Clara, California, USA). gDNA contamination was evaluated by gel electrophoresis.

2.3. miRNA Expression Profiling. Total RNA samples (2.5 μg) were subjected to microarray analysis of microRNA expression using the Human miRNA OneArray v5 (Phalanx Biotech Group, Hsinchu, Taiwan). Labeling efficiency was calculated by the concentration of CyDye and RNA was measured by NanoDrop ND-1000. Normalization and statistical analysis were conducted with R/Bioconductor (Bioconductor, Fred Hutchinson Cancer Research Center). Expression profiles of changes induced by the various radiation doses (0.5 Gy, 1 Gy, 2.5 Gy, and 5 Gy) were each normalized to the control without any radiation exposure. Significantly differentially expressed miRNAs (normalized intensity $\geq$ 300, absolute Log_2 ratio $\geq$ 1, absolute fold change $\geq$ 1, and FDR $<$ 0.05) were categorized into up- and downregulated genes for each radiation dose.

2.4. miRNA Target Prediction. We adopted an integrative approach, utilizing publicly available databases and our own gene expression data, to identify the target genes for the differentially expressed miRNAs (Figure 1). First, a list of validated targets for a specific differentially expressed miRNA was generated by performing a search through the validated data in miRWalk database [23]. For the predicted data, the accepted target predictions were those identified by at least four out of the five well-established databases, including miRWalk [23], miRANDA [24], miRDB [25], RNA22 [26], and TargetScan [27]. Next, we utilized a list of differentially expressed genes identified from our previous gene expression study of changes in human peripheral blood mononuclear cells induced by varying doses of ^{60}Co radiation (absolute Log_2 ratio $\geq$ 1, absolute fold change $\geq$ 1, and FDR $<$ 0.05). For each ^{60}Co radiation dosage, we grouped the downregulated genes with upregulated microRNAs and vice versa. On another web-based system, miRTar [28], we input these genes to look for possible miRNA-gene interactions based on the target genes' 3′UTR (untranslated regions), 5′UTR, and coding regions. This systematic approach filtered out previously identified candidate genes that did not match the predicted or validated miRNA-gene interaction list.

2.5. miRNA-Gene Interaction Analysis. By employing the gene enrichment function in miRTar [28], the putative or

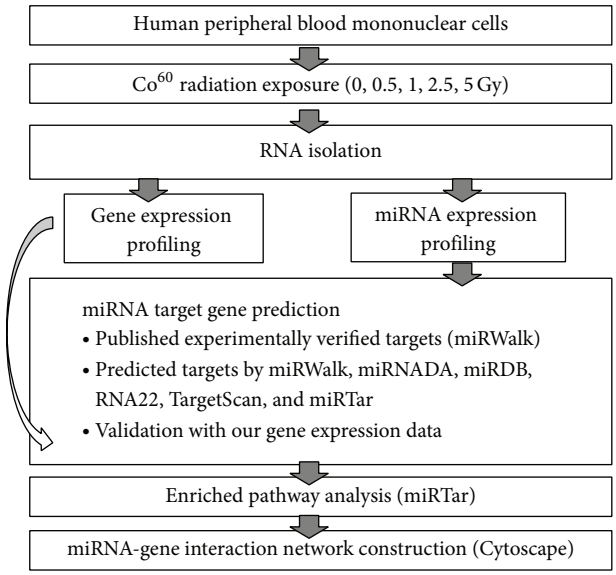


Figure 1: System flow of our analysis.

Table 1: Number of significantly differentially expressed miRNAs in human peripheral blood mononuclear cells exposed to varying doses of ^{60}Co radiation (absolute fold change ≥ 1; FDR < 0.05).

Comparison (Gy)	Upregulated miRNAs	Downregulated miRNAs
0 versus 0.5	hsa-miR-185-5p	0
0 versus 1	hsa-miR-107 hsa-miR-126-3p hsa-miR-144-3p hsa-miR-17-5p hsa-miR-185-5p hsa-miR-20b-5p hsa-miR-5194	hsa-miR-3180 hsa-miR-4730
0 versus 2.5	0	0
0 versus 5	0	hsa-miR-142-3p hsa-miR-142-5p hsa-miR-223-3p hsa-miR-451a

experimentally verified miRNA-gene interactions identified through the integrative database searches were mapped to KEGG (Kyoto Encyclopedia of Genes and Genomes) [29] to reveal the biological importance of these molecular relationships. Only those enriched pathways with a P value < 0.05 were considered significant. After applying such a threshold, only miRNAs differentially expressed under 1 Gy of radiation exposure were retained. To visualize the relationships between the candidate miRNAs and their differentially expressed target genes, an interaction network integrating the expression values of each miRNA and gene was constructed using Cytoscape v.3.1.0 (National Institute of General Medical Sciences and National Resource for Network Biology, USA) [30, 31].

3. Results

Few miRNAs showed significant changes in expression under 0.5 and 2.5 Gy of ^{60}Co radiation exposure (Table 1). Whereas only four miRNAs exhibited enhanced expression under 5 Gy of radiation, a dosage of 1 Gy appeared to induce expression changes in the greatest number of microRNAs, with seven being upregulated and two downregulated. The pattern is consistent with our findings [32] on the gene expression changes in human peripheral blood mononuclear cells exposed to the same varying doses of ^{60}Co radiation.

To elucidate the potential regulatory relationship between the differentially expressed miRNAs and our previously identified gene candidates [32], we performed a systematic search utilizing a variety of database tools. Table 2 presents the potential relationships between the candidate miRNAs and their corresponding target genes as validated by our list of differentially expressed genes. It appears that most of the molecular changes occurred at 1 Gy of radiation dosage.

In order to identify the biological importance of the molecular changes occurring under 1 Gy of radiation exposure, the differentially expressed miRNAs and their corresponding target genes were mapped to their respective KEGG pathways. The overrepresented pathways seem to be related to homeostasis, metabolism, neuronal survival, and cellular control (Table 3). Three microRNAs, hsa-miR-20b-5p, hsa-miR-17-5p, and hsa-miR-185-5p, appear to regulate the highest number of radiation sensitive genes compared to the other differentially expressed microRNAs (Table 2).

Moreover, these miRNAs' target genes are enriched in cancer and cell cycle-related pathways (Table 3). Consequently, hsa-miR-20b-5p, hsa-miR-17-5p, and hsa-miR-185-5p may be involved in modulating genes underlying cell cycle control and the development of thyroid cancer and prostate cancer. The pathway-specific association between these miRNAs and their corresponding target genes is shown in Figures 2–4.

Our data suggest that the miRNA-gene interactions associated with 1 Gy of radiation dosage treatment may be the key molecular signatures underlying the damages caused by radiation exposure. In order to visualize the relationships between these miRNA and gene candidates, we constructed an interaction network that illustrates the complex regulatory relationships among these genes and miRNAs. Note that hsa-miR-20b-5p and hsa-miR-17-5p share many target genes, suggesting that they modulate gene expression through a cooperative manner (Figure 5).

4. Discussion

In the current study, we profiled miRNA expression changes under varying doses of radiation exposure through an array-based approach. Our results support the emerging evidence that tissue and cellular injuries may alter miRNA expression [33]. In addition, by utilizing publicly available bioinformatics resources and comparing with our previous gene expression profiling data, we have mapped out a potential miRNA-gene

Table 2: Putative and validated interactions between the differentially expressed microRNAs and gene candidates specific to each dose of ^{60}Co radiation.

Dose (Gy)	miRNA	Fold change	Target gene	Fold change
0.5	hsa-miR-185-5p	1.18	*TNFSF10*[a]	−1.21
			GLUL[c]	−1.49
1	hsa-miR-107	1.09	14 genes[c]	Decreased
	hsa-miR-144	1.60	*SELL*	−1.04
			COTL	−1.06
			CEP63	−1.14
	hsa-miR-17-5p	1.09	*MCL1*[a]	−1.26
			FGL2[b]	−1.33
			30 other genes[c]	Decreased
	hsa-miR-185-5p	1.01	32 genes[c]	Decreased
	hsa-miR-20b-5p	1.01	27 genes[c]	Decreased
5	hsa-miR-142-3p	−1.02	*MAP4K3*[b]	1.32
			DIRC2[b]	1.01
			TIPARP[b]	1.05
			PDE4B[c]	1.54
	hsa-miR-142-5p	−1.09	*AHR*[a]	1.31
			SLC36A4[b]	1.11
	hsa-miR-223-3p	−1.27	*DUSP10*[b,c]	1.07
			EFNA1[b]	1.01
	hsa-miR-451a	−1.28	*SLC7A11*[c]	1.14

[a]Experimentally validated miRNA-gene interaction as identified by miRWalk.
[b]miRNA-gene interaction predicted by at least four out of the five selected miRNA target prediction databases.
[c]miRNA-gene interaction predicted by miRTar; the complete gene list is provided in Supplementary Material 1 (see Supplementary Material available online at http://dx.doi.org/10.1155/2014/456323).

Table 3: Enriched KEGG pathways associated with specific miRNA-gene interactions under 1 Gy of ^{60}Co radiation exposure.

microRNA	Pathway	Genes	*P* value
hsa-miR-185-5p	Cell cycle	*YWHAG, YWHAB,* and *PCNA*	0.0019
	Prostate cancer	*TCF7, HSP90AA1*	0.0380
hsa-miR-107	Neurotrophin signaling pathway	*YWHAB, CRKL*	0.0200
	Renin-angiotensin system	*MME*	0.0351
hsa-miR-20b-5p	Thyroid cancer	*TCF7*	0.0490
hsa-miR-17-5p	Pentose phosphate pathway	*TALDO1*	0.0490
	Thyroid cancer	*TCF7*	0.0483

interactome map that may underlie the molecular changes induced by radiation treatment.

Most of the upregulated miRNAs were found under low doses of ^{60}Co radiation exposure (≤1 Gy), while only significantly downregulated miRNAs were found in the 5 Gy radiation dosage range. This indicates that response to radiation exposure at the miRNA level is dose dependent. From publicly available miRNA knowledge bases, we retrieved a list of genes that have been validated to interact with these miRNAs, namely, hsa-miR-185-5p, hsa-miR-107, hsa-miR-20b-5p, and hsa-miR-17-5p for the low radiation doses and hsa-miR-142, hsa-miR-223-3p, and hsa-miR-451a for the 5 Gy radiation exposure. These genes were compared to those identified from our previous gene expression profiling experiment. Matched genes were considered the most promising targets.

Our analysis showed that two miRNAs, hsa-miR-142-3p and hsa-miR-223-3p, with decreased expression after exposure to 5 Gy of ^{60}Co radiation may be potential modulators of *MAP4K3* (mitogen-activated protein kinase kinase kinase kinase 3) and *DUSP10* (dual specificity phosphatase 10), respectively. In particular, MAP4K3 is an apoptosis inducer that is activated upon UV radiation [34]. Mutation in the *MAP4K3* gene sequence was predicted to modulate cancer progression [35], and this hypothesis was supported by abnormal levels of MAP4K3 expression in pancreatic cancer tissues and enhanced cellular proliferation by RNAi-induced suppression of *MAP4K3* [36]. On the other hand, increased transcript abundance of DUSP10 was shown to influence gut homeostasis by suppressing proliferation and apoptosis, while promoting differentiation [37]. Both DUSP10 and

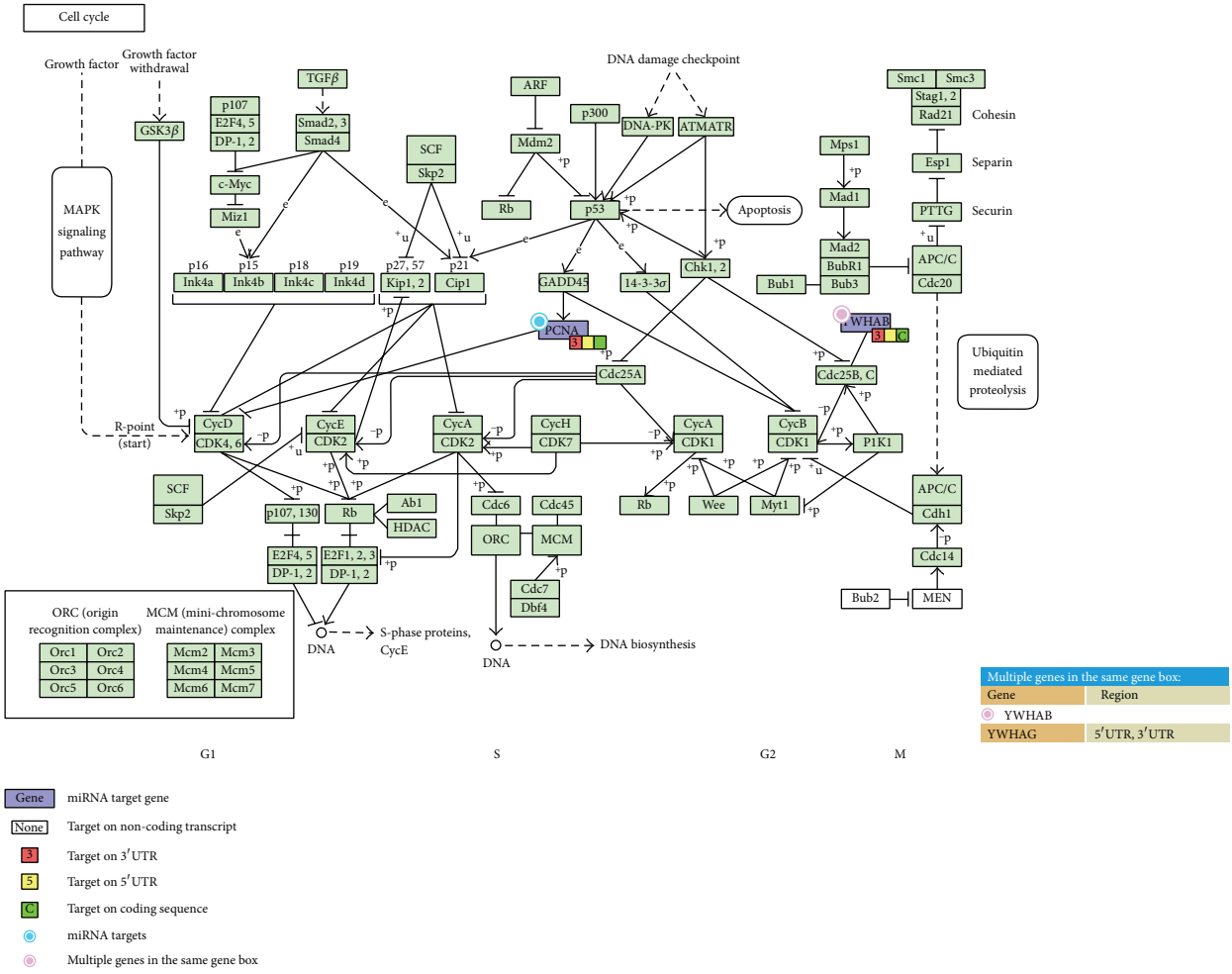


Figure 2: Regulation of *YWHAG*, *YWHAB*, and *PCNA* by hsa-miR-185-5p in a cell cycle pathway.

MAP4K3 play important roles in the MAPK signaling pathway [38]. Our results suggest that the increase in *MAP4K3* and *DUSP10* transcript abundance as a result of 5 Gy radiation exposure may be modulated by their interactions with hsa-miR-142-3p and hsa-miR-223-3p, respectively. These putative relationships might even be one of the underlying regulatory mechanisms of radiation-induced changes in cell cycle signaling.

As the upregulated miRNAs were primarily found in the low-dose radiation range (0.5 Gy and 1 Gy), we performed a series of gene set enrichment analysis on their candidate target genes and mapped these miRNA-gene interactions to pathways related to the cell cycle, neurotrophin signaling, renin-angiotensin system, and pentose phosphate pathways, as well as prostate and thyroid cancers. This is in line with the observation that radiation induced apoptosis through the neurotrophin signaling pathway [39]. Delayed wound healing following low-dose radiation exposure was also partially attributed to reduced activity of the renin-angiotensin system [40]. Moreover, evidence has shown that increased activity of the pentose phosphate cycle can protect cells from programmed cell death induced by low doses of ionizing radiation [41]. According to our results, it is possible that specific miRNAs are upregulated to modulate genes involved in these pathways in response to low-dose radiation.

The miRNA hsa-miR-185-5p exhibited increased expression when exposed to 0.5 Gy and 1 Gy dosage of radiation and was predicted to interact with *YWHAG* (tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, gamma polypeptide), *YWHAB* (tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, beta polypeptide), and *PCNA* (proliferating cell nuclear antigen), which appeared to be downregulated under the same condition. Pathway analysis (see Figure 2) suggested that these three genes are involved in cell cycle pathways. In particular, YWHAB is known to regulate the cyclin B and cyclin-dependent kinase 1 complex, while PCNA is an inhibitor of the cyclin D and cyclin-dependent kinase 4 or 6 complex [42, 43]. According to a porcine model study, YWHAG and YWHAB mediate insulin-like growth factor signaling and the G2/M DNA damage checkpoint in cell cycle control [44]. Maternal high protein diet has been shown to associate with increasing expression levels of *YWHAG* and *YWHAB* [44]. Our results support that these three

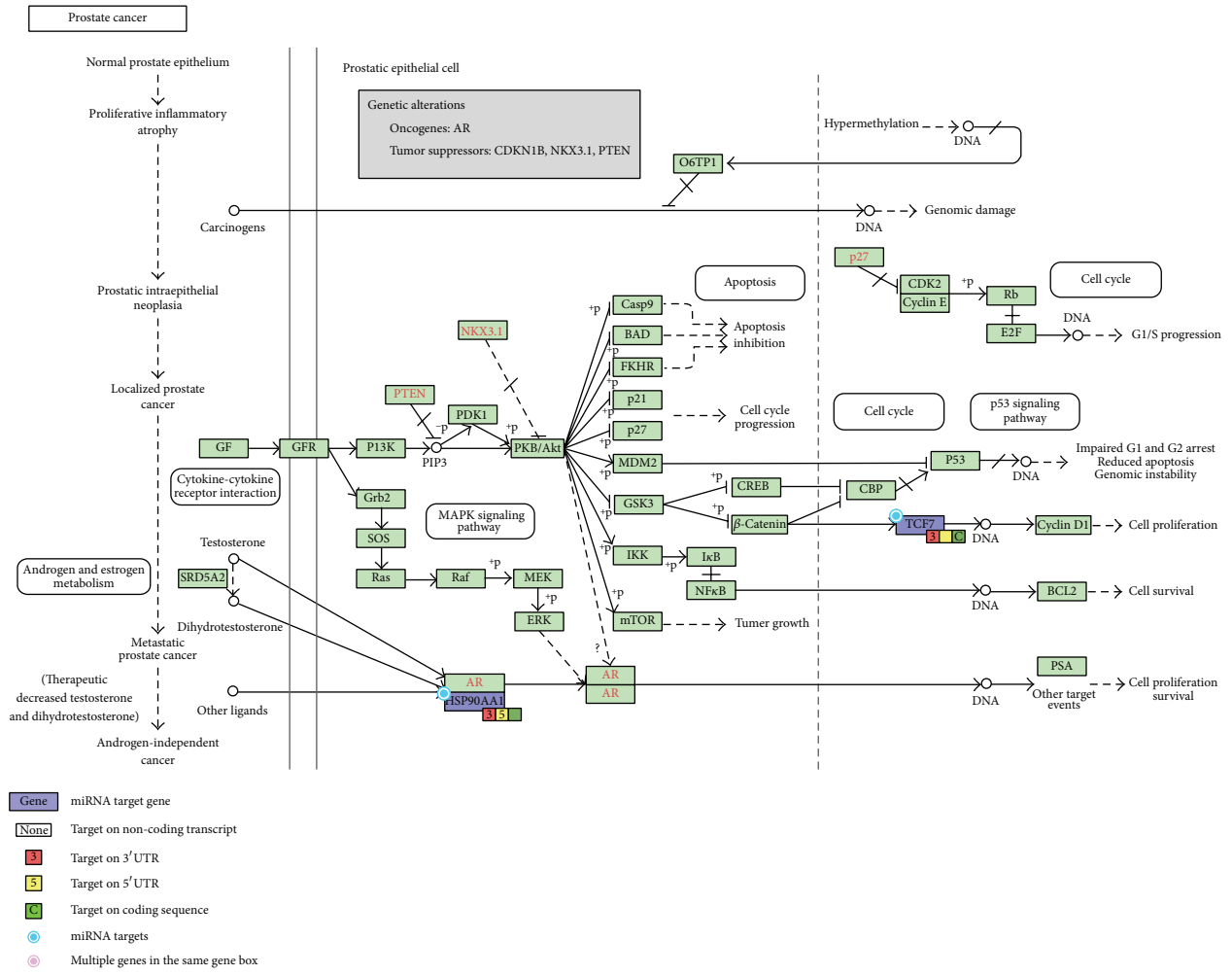


Figure 3: Regulation of *TCF7* and *HSP90AA1* by hsa-miR-185-5p in a prostate cancer pathway.

genes regulate the cell cycle pathway and are particularly sensitive to changes in the external environment, especially radiation exposure, and that hsa-miR-185-5p may be involved in mediating this response. In addition, we showed that hsa-miR-185 may control the expression of *TCF7* (T-cell-specific transcription factor 7) and *HSP90AA1* (heat shock protein 90 kDa alpha, class A member 1) in Figure 3. TCF7 is a known regulator of the Wnt signaling pathway [45] and an important modulator of the self-renewal and differentiation processes in hematopoietic cells [46]. In contrast, *HSP90AA1* is responsible for the degradation of androgen receptor and cell killing following radiation exposure in a prostate cancer cell line [47]. In addition, hsa-miR-185 is also known to be involved in the development of prostate cancer [48]. Our pathway analysis further supports the roles these three molecules play in carcinogenesis by mapping the interaction among hsa-miR185-5p, *TCF7*, and *HSP90AA1* to prostate cancer through the PI3K signaling pathway.

Previous study reported that hsa-miR-107 regulates the DNA damage response (DDR) and sensitizes tumor cells by repressing expression of *RAD51* and corporation with miR-222 in olaparib, an experimental chemotherapeutic agent, thus impairing DSB repair by HR [49]. Elevated expression of miR-107 has been correlated with *PARP* inhibitor sensitivity and reduced *RAD51* expression in a subset of ovarian clear cell carcinomas [49]. The miRNAs hsa-mir-103 and hsa-mir-107 are upregulated in relation to insulin sensitivity in an obese mouse model. This suggested that these miRNAs represent potential biomarkers for type 2 diabetes (the miR-103 microRNA precursor is homologous to miR-107) [50]. The miR-107 negatively regulates the miRNA let-7 via direct interaction in tumors and in cancer cell line. Previous study showed that miR-107 increased the tumourigenic and metastatic potential via inhibition of let-7 and upregulation of let-7 targets in human breast cancer cell line and in mice model [51]. Our result indicates that miR-107 is upregulated under 1 Gy dosage of radiation exposure and this increase in expression is associated with metabolic pathways and potentially involved in cancer development. Three genes, *MME*, *YWHAB*, and *CRKL*, were predicted to interact with miR-107. *MME* was involved in rennin-angiotensin pathway and *CRKL* and *YWHAB* were involved

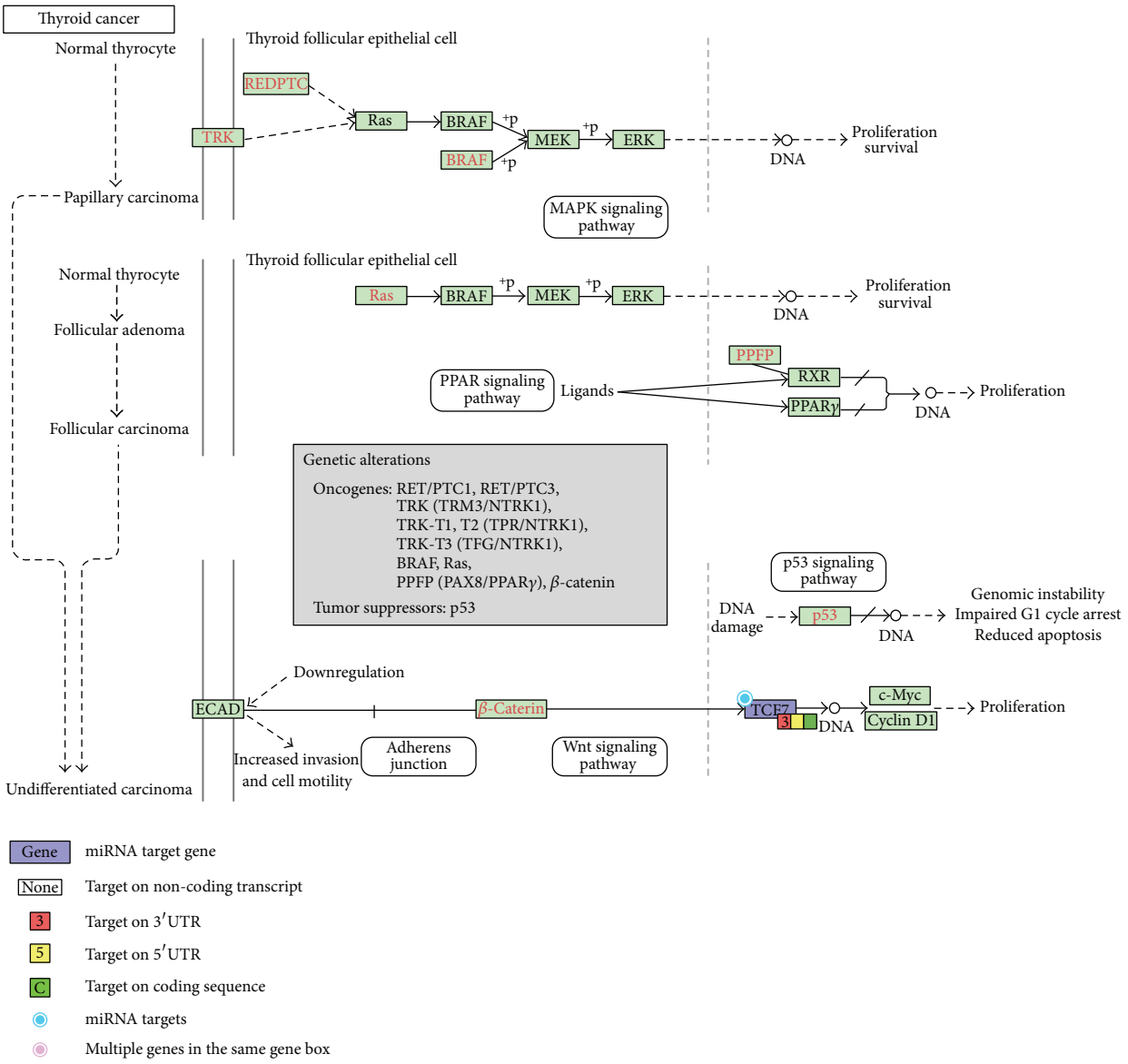


Figure 4: Regulation of *TCF7* by hsa-miR-20b-5p and hsa-miR-17-5p in a thyroid cancer pathway.

in neurotrophin signaling pathway in our data. Our analysis demonstrated that miR-107 may cooperate with miR-185-5p to regulate the cell cycle via *YWHAB*. This result corresponds to a previous study, which showed that the miR-107 and miR-185, localized in frequently altered chromosomal regions in human lung cancers, may contribute to regulate cell cycle in human malignant tumors [52].

In addition, our analysis in Figure 4 suggests that miR-17-5p and miR-20b-5p may cooperate to exert regulatory effects on the *TCF7* gene. In fact, miR-17-5p and miR-20b-5p are mature forms of the same precursor family. The microRNA miR-17-5p has been shown to mediate the transition from G1 to S phase of the cell cycle and initiate the signal for proliferation [53, 54]. Moreover, miR-17-5p can act as both an oncogene and a tumor suppressor gene in different cellular contexts, underscoring its importance in cell cycle control [53]. Our finding indicates that miR-17-5p and miR-20b-5p may modulate the Wnt signaling pathway by regulating *TCF7* expression, which in turn affects the activity of the c-Myc and cyclin D1 complex. This particular process is associated with thyroid cancer. Thus, under low-dose radiation, changes in the abundance of miR-17-5p and miR-20b-5p may influence the cell cycle via interaction with their target gene *TCF7* and modulate the development of thyroid cancer [55, 56].

Our study demonstrates that many miRNAs were upregulated in response to low-dose radiation, and these radiation-induced changes may alter cell cycle regulation, affecting cell rescue, interrupt the generation of NADPH and pentoses through glycolysis, create imbalance in body fluid homeostasis via the renin-angiotensin system (RAS), and finally modulate cell survival through the neurotrophin signaling. In conclusion, we have provided comprehensive miRNA-gene

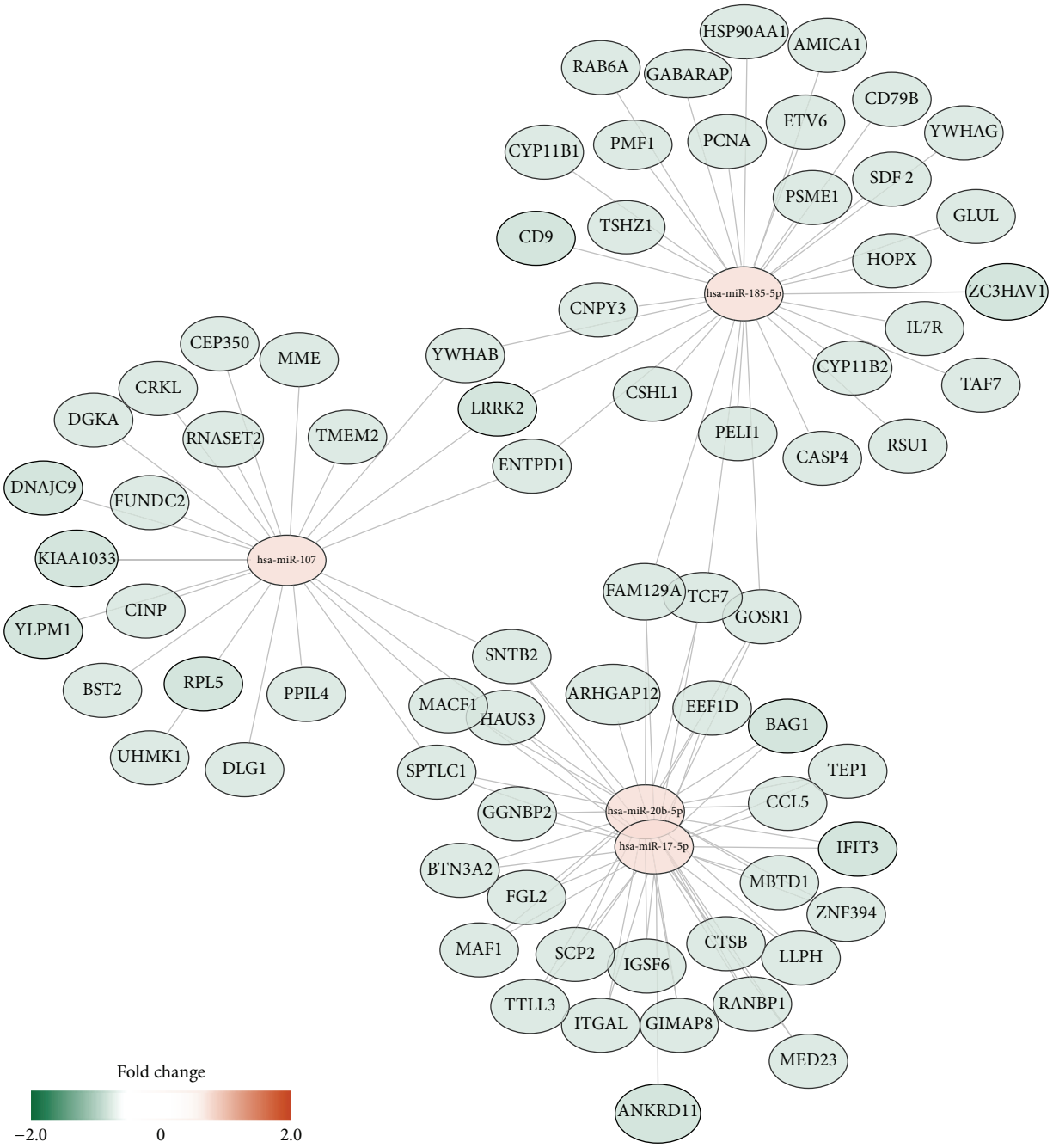


FIGURE 5: Potential miRNA-gene interaction network associated with the changes induced by 1 Gy of ^{60}Co radiation exposure in human peripheral blood mononuclear cells.

interaction networks that underlie the mechanisms of damages induced by varying doses of radiation. Our findings have built a framework for further validation studies to investigate the specific molecular signatures of radiation exposure.

Conflict of Interests

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

References

[1] B. J. Blyth and P. J. Sykes, "Radiation-induced bystander effects: what are they, and how relevant are they to human radiation exposures?," *Radiation Research*, vol. 176, no. 2, pp. 139–157, 2011.

[2] N. K. Jacob, J. V. Cooley, T. N. Yee et al., "Identification of sensitive serum microrna biomarkers for radiation biodosimetry," *PLoS ONE*, vol. 8, no. 2, Article ID e57603, 2013.

[3] H. Lodish, A. Berk, and P. Matsudaira, "Cancer: the role of carcinogens and DNA repair in cancer," *Molecular Biology of the Cell*, p. 963, 2004.

[4] E. H. Donnelly, J. B. Nemhauser, J. M. Smith et al., "Acute radiation syndrome: assessment and management," *Southern Medical Journal*, vol. 103, no. 6, pp. 541–546, 2010.

[5] M. Xiao and M. H. Whitnall, "Pharmacological countermeasures for the acute radiation syndrome," *Current Molecular Pharmacology*, vol. 2, no. 1, pp. 122–133, 2009.

[6] E. I. Azzam and J. B. Little, "The radiation-induced bystander effect: evidence and significance," *Human & Experimental Toxicology*, vol. 23, no. 2, pp. 61–65, 2004.

[7] M. Mancuso, E. Pasquali, S. Leonardi et al., "Oncogenic bystander radiation effects in Patched heterozygous mouse cerebellum," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 34, pp. 12445–12450, 2008.

[8] S. Demaria, B. Ng, M. L. Devitt et al., "Ionizing radiation inhibition of distant untreated tumors (abscopal effect) is immune mediated," *International Journal of Radiation Oncology Biology Physics*, vol. 58, no. 3, pp. 862–870, 2004.

[9] M. H. Barcellos-Hoff, C. Adams, A. Balmain et al., "Systems biology perspectives on the carcinogenic potential of radiation," *Journal of Radiation Research*, vol. 55, supplement 1, pp. i145–i154, 2014.

[10] W. F. Morgan and M. B. Sowa, "Effects of ionizing radiation in nonirradiated cells," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 40, pp. 14127–14128, 2005.

[11] R. Baskar, "Emerging role of radiation induced bystander effects: cell communications and carcinogenesis," *Genome Integrity*, vol. 1, article 13, 2010.

[12] U. Aypar, W. F. Morgan, and J. E. Baulch, "Radiation-induced genomic instability: are epigenetic mechanisms the missing link?" *International Journal of Radiation Biology*, vol. 87, no. 2, pp. 179–191, 2011.

[13] W. F. Morgan, "Non-targeted and delayed effects of exposure to ionizing radiation: I. Radiation-induced genomic instability and bystander effects in vitro," *Radiation Research*, vol. 159, no. 5, pp. 567–580, 2003.

[14] W. F. Morgan, "Non-targeted and delayed effects of exposure to ionizing radiation: II. Radiation-induced genomic instability and bystander effects in vivo, clastogenic factors and transgenerational effects," *Radiation Research*, vol. 159, no. 5, pp. 581–596, 2003.

[15] S. Nagar, L. E. Smith, and W. F. Morgan, "Variation in apoptosis profiles in radiation-induced genomically unstable cell lines," *Radiation Research*, vol. 163, no. 3, pp. 324–331, 2005.

[16] E. Li, "Chromatin modification and epigenetic reprogramming in mammalian development," *Nature Reviews Genetics*, vol. 3, no. 9, pp. 662–673, 2002.

[17] S. Ma, X. Liu, B. Jiao, and Y. Yang, "Low-dose radiation-induced responses: focusing on epigenetic regulation," *International Journal of Radiation Biology*, vol. 86, no. 7, pp. 517–528, 2010.

[18] N. K. Jacob, J. V. Cooley, T. N. Yee et al., "Identification of sensitive serum microRNA biomarkers for radiation biodosimetry," *PLoS ONE*, vol. 8, no. 2, Article ID e57603, 2013.

[19] W. Cui, J. Ma, Y. Wang, and S. Biswal, "Plasma miRNA as biomarkers for assessment of total-body radiation exposure dosimetry," *PLoS ONE*, vol. 6, no. 8, Article ID e22988, 2011.

[20] T. Templin, S. A. Amundson, D. J. Brenner, and L. B. Smilenov, "Whole mouse blood microRNA as biomarkers for exposure to γ-rays and 56Fe ions," *International Journal of Radiation Biology*, vol. 87, no. 7, pp. 653–662, 2011.

[21] International Organization for Standardization (ISO), "Radiation protection—performance criteria for service laboratories performing biological dosimetry by cytogenetics," ISO 19238, ISO, Geneva, Switzerland, 2004.

[22] IAEA, "Cytogenetic analysis for radiation dose assessment: a manual," Tech. Rep. 405, International Atomic Energy Agency, Vienna, Austria, 2001.

[23] H. Dweep, C. Sticht, P. Pandey, and N. Gretz, "MiRWalk-database: prediction of possible miRNA binding sites by "walking" the genes of three genomes," *Journal of Biomedical Informatics*, vol. 44, no. 5, pp. 839–847, 2011.

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

[25] X. Wang, "miRDB: a microRNA target prediction and functional annotation database with a wiki interface," *RNA*, vol. 14, no. 6, pp. 1012–1017, 2008.

[26] P. Loher and I. Rigoutsos, "Interactive exploration of RNA22 microRNA target predictions," *Bioinformatics*, vol. 28, no. 24, pp. 3322–3323, 2012.

[27] B. P. Lewis, C. B. Burge, and D. P. Bartel, "Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets," *Cell*, vol. 120, no. 1, pp. 15–20, 2005.

[28] J. B. Hsu, C.-M. Chiu, S.-D. Hsu et al., "MiRTar: an integrated system for identifying miRNA-target interactions in human," *BMC Bioinformatics*, vol. 12, article 300, 2011.

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

[30] M. S. Cline, M. Smoot, E. Cerami et al., "Integration of biological networks and gene expression data using Cytoscape," *Nature Protocols*, vol. 2, no. 10, pp. 2366–2382, 2007.

[31] 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.

[32] K. F. Lee, Y. C. Chen, P. W. C. Hsu et al., "Gene expression profiling of biological pathway alterations by radiation exposure," *BioMed Research International*, vol. 2014, Article ID 834087, 9 pages, 2014.

[33] X. Ji, R. Takahashi, Y. Hiura, G. Hirokawa, Y. Fukushima, and N. Iwai, "Plasma miR-208 as a biomarker of myocardial injury," *Clinical Chemistry*, vol. 55, no. 11, pp. 1944–1949, 2009.

[34] K. Diener, X. S. Wang, C. Chen et al., "Activation of the c-Jun N-terminal kinase pathway by a novel protein kinase related to human germinal center kinase," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 94, no. 18, pp. 9687–9692, 1997.

[35] S. Jones, X. Zhang, D. W. Parsons et al., "Core signaling pathways in human pancreatic cancers revealed by global genomic analyses," *Science*, vol. 321, no. 5897, pp. 1801–1806, 2008.

[36] D. Lam, D. Dickens, E. B. Reid, S. H. Y. Loh, N. Moisoi, and L. M. Martins, "MAP4K3 modulates cell death via the post-transcriptional regulation of BH3-only proteins," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 29, pp. 11978–11983, 2009.

[37] R. Wang, I.-K. Kwon, N. Singh et al., "Type 2 cGMP-dependent protein kinase regulates homeostasis by blocking c-Jun N-terminal kinase in the colon epithelium," *Cell Death and Differentiation*, vol. 21, no. 3, pp. 427–437, 2014.

[38] C.-Y. Yang, J.-P. Li, L.-L. Chiu et al., "Dual-specificity phosphatase 14 (DUSP14/MKP6) negatively regulates TCR signaling by inhibiting TAB1 activation," *The Journal of Immunology*, vol. 192, no. 4, pp. 1547–1557, 2014.

[39] D. H. Kim, X. Zhao, C. H. Tu, P. Casaccia-Bonnefil, and M. V. Chao, "Prevention of apoptotic but not necrotic cell death following neuronal injury by neurotrophins signaling through the tyrosine kinase receptor," *Journal of Neurosurgery*, vol. 100, no. 1, pp. 79–87, 2004.

[40] S. S. Jadhav, N. Sharma, C. J. Meeks et al., "Effects of combined radiation and burn injury on the rennin-angiotensin system," *Wound Repair and Regeneration*, vol. 21, no. 1, pp. 131–140, 2013.

[41] S. Tuttle, T. Stamato, M. L. Perez, and J. Biaglow, "Glucose-6-phosphate dehydrogenase and the oxidative pentose phosphate cycle protect cells against apoptosis induced by low doses of ionizing radiation," *Radiation Research*, vol. 153, no. 6, pp. 781–787, 2000.

[42] A. L. Gartel and A. L. Tyner, "The role of the cyclin-dependent kinase inhibitor p21 in apoptosis," *Molecular Cancer Therapeutics*, vol. 1, no. 8, pp. 639–649, 2002.

[43] W. Strzalka and A. Ziemienowicz, "Proliferating cell nuclear antigen (PCNA): a key factor in DNA replication and cell cycle regulation," *Annals of Botany*, vol. 107, no. 7, pp. 1127–1140, 2011.

[44] M. Oster, E. Murani, C. C. Metges, S. Ponsuksili, and K. Wimmers, "A high protein diet during pregnancy affects hepatic gene expression of energy sensing pathways along ontogenesis in a porcine model," *PLoS ONE*, vol. 6, no. 7, Article ID e21691, 2011.

[45] H. Clevers and M. van de Wetering, "TCF/LEF factors earn their wings," *Trends in Genetics*, vol. 13, no. 12, pp. 485–489, 1997.

[46] J. Q. Wu, M. Seay, V. P. Schulz et al., "Tcf7 is an important regulator of the switch of self-renewal and differentiation in a multipotential hematopoietic cell line," *PLoS Genetics*, vol. 8, no. 3, Article ID e1002565, 2012.

[47] K. Harashima, T. Akimoto, T. Nonaka, K. Tsuzuki, N. Mitsuhashi, and T. Nakano, "Heat shock protein 90 (Hsp90) chaperone complex inhibitor, Radicicol, potentiated radiation-induced cell killing in a hormone-sensitive prostate cancer cell line through degradation of the androgen receptor," *International Journal of Radiation Biology*, vol. 81, no. 1, pp. 63–76, 2005.

[48] X. Li, Y.-T. Chen, S. Josson et al., "MicroRNA-185 and 342 inhibit tumorigenicity and induce apoptosis through blockade of the SREBP metabolic pathway in prostate cancer cells," *PLoS ONE*, vol. 8, no. 8, Article ID e70987, 2013.

[49] S. Neijenhuis, I. Bajrami, R. Miller, C. J. Lord, and A. Ashworth, "Identification of miRNA modulators to PARP inhibitor response," *DNA Repair*, vol. 12, no. 6, pp. 394–402, 2013.

[50] M. Trajkovski, J. Hausser, J. Soutschek et al., "MicroRNAs 103 and 107 regulate insulin sensitivity," *Nature*, vol. 474, no. 7353, pp. 649–653, 2011.

[51] P.-S. Chen, J.-L. Su, S.-T. Cha et al., "miR-107 promotes tumor progression by targeting the let-7 microRNA in mice and humans," *The Journal of Clinical Investigation*, vol. 121, no. 9, pp. 3442–3455, 2011.

[52] N. Cloonan, M. K. Brown, A. L. Steptoe et al., "The miR-17-5p microRNA is a key regulator of the G1/S phase cell cycle transition," *Genome Biology*, vol. 9, no. 8, article R127, 2008.

[53] A. Hossain, M. T. Kuo, and G. F. Saunders, "Mir-17-5p regulates breast cancer cell proliferation by inhibiting translation of AIB1 mRNA," *Molecular and Cellular Biology*, vol. 26, no. 21, pp. 8191–8201, 2006.

[54] R. Hrdličková, J. Nehyba, W. Bargmann, and H. R. Bose Jr., "Multiple tumor suppressor microRNAs regulate telomerase and TCF7, an important transcriptional regulator of the Wnt pathway," *PLoS ONE*, vol. 9, no. 2, Article ID e86990, 2014.

[55] A. Sastre-Perona and P. Santisteban, "Role of the Wnt pathway in thyroid cancer," *Frontiers in Endocrinology*, vol. 3, article 31, 10 pages, 2012.

[56] S. D. Solomon, N. Anavekar, FRACP, and NHMRC Faculty and Disclosures, "A Brief Overview of Inhibition of the Renin-Angiotensin System: Emphasis on Blockade of the Angiotensin II Type-1 Receptor," 2005.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 807839, 12 pages
http://dx.doi.org/10.1155/2014/807839

Research Article

EXIA2: Web Server of Accurate and Rapid Protein Catalytic Residue Prediction

Chih-Hao Lu,[1] Chin-Sheng Yu,[2] Yu-Tung Chien,[3] and Shao-Wei Huang[3]

[1] *Graduate Institute of Molecular Systems Biomedicine, China Medical University, Taichung 40402, Taiwan*
[2] *Department of Information Engineering and Computer Science, Feng Chia University, Taichung 40724, Taiwan*
[3] *Department of Medical Informatics, Tzu Chi University, Hualien 97004, Taiwan*

Correspondence should be addressed to Shao-Wei Huang; swhwang.orz@gmail.com

Received 21 February 2014; Revised 27 May 2014; Accepted 11 June 2014; Published 11 September 2014

Academic Editor: Tzong-Yi Lee

We propose a method (EXIA2) of catalytic residue prediction based on protein structure without needing homology information. The method is based on the special side chain orientation of catalytic residues. We found that the side chain of catalytic residues usually points to the center of the catalytic site. The special orientation is usually observed in catalytic residues but not in noncatalytic residues, which usually have random side chain orientation. The method is shown to be the most accurate catalytic residue prediction method currently when combined with PSI-Blast sequence conservation. It performs better than other competing methods on several benchmark datasets that include over 1,200 enzyme structures. The areas under the ROC curve (AUC) on these benchmark datasets are in the range from 0.934 to 0.968.

1. Introduction

Enzymes play important roles in various biological processes. As the number of sequenced genomes rapidly grows, the function of the majority of proteins can only be annotated computationally. While a number of methods have been reported to predict protein function from protein sequence [1–3], protein structure [4, 5], protein-protein interaction network [6, 7], and evolutionary relationships [8–10], the complexity of protein function makes function prediction challenging. In addition, prediction of protein function is distinct from actual identification of functional regions or residues. To identify the location of functional regions in protein, a number of methods have been reported to predict protein functional site, including ligand binding sites [11–13], phosphorylation sites [14, 15], protein-protein interaction sites [16–18], and ubiquitination site [19] from protein sequence, structure, or high-throughput experimental data. Here we focus on the prediction of protein catalytic residues. Although only a small number of residues compose enzyme catalytic site, they are the most crucial part for enzymes to perform their function. Identifying these critical residues and characterizing their features are crucial to understanding the molecular basis of protein function.

Sequence or structure homology information is the primary feature used in catalytic residue prediction [20–26] because catalytic residues are usually evolutionarily conserved. One of the most successful sequence-based methods is CRpred [27], which used several types of sequence-based features including position-specific scoring matrix and entropy of weighted observed percentages extracted from multiple sequence alignment using PSI-BLAST [28]. Another method, ConSurf, identifies functionally important regions in proteins by estimating the degree of conservation of the amino acid sites among their close sequence homologues [29]. However, homology-based methods require reliable evolutionary information, for example, multiple sequence alignment constructed from enough number of protein sequences. Recent studies show that the evolutionary origin of one-third of genes is not clear in yeast genome [30]. For a protein that lacks reliable homology information, it is important to develop prediction method based on information contained in the protein itself. Several methods have been proposed to extract as much information as possible from

protein structure. A method is based on the observation that catalytic residues are usually moderately exposed residues that are located closest to the protein centroid [31]. It was shown that if a protein was converted to a network in which the residues are vertices and their interactions are edges, the central hubs are usually functionally important residues or their neighboring residues [32]. It was also reported that catalytic residues usually have higher force constant, that is, the ease of moving a given residue with respect to the other residues in the protein [33]. The theoretical microscopic titration curves (THEMATICS) [34] method detects catalytic residues by calculating theoretical residue electrostatic properties from protein structure. THEMATICS was then enhanced by structure geometric features [35] to detect catalytic residues from protein structure using a monotonicity-constrained maximum likelihood approach, called partial order optimum likelihood (POOL). A more recent study [36] models the properties, such as physicochemical properties, atomic density, flexibility, and presence of water molecules or heteroatom, of spherical regions around target residues. Such methods are helpful for proteins that do not have reliable homology information. However, current methods that do not rely on homology information perform worse than other homology-based methods.

Here we propose a method (EXIA2) of catalytic residue prediction based on protein structure without needing homology information. The method is an improved version of our previous work [37], which is based on the special side chain orientation of catalytic residues. We found that the side chain of catalytic residues usually points to the center of the catalytic site. The special orientation is usually observed in catalytic residues but not in noncatalytic residues, which usually have random side chain orientation. The feature is effective in the identification of catalytic residues from enzyme structure. In this work, we further add a new property, the amino acid combination feature, which is a general composition of amino acids in enzyme catalytic site. We implement the web server and optimize its computation efficiency for practical use. The prediction performance of EXIA2 web server is better than those of other state-of-the-art prediction methods. In addition to better prediction performance, it is more efficient than other structure-based prediction methods.

2. Description of Web Server

2.1. Input. The input for the web server is a 3D protein structure in Protein Data Bank (PDB) [38] format. Users can upload their own protein structure file or input a PDB id. Each submission allows a structure of up to 5000 residues. The results are displayed instantly for small and medium size proteins. For proteins of larger size (more than 3000 residues), the processing time is normally from several seconds to a minute.

2.2. Output. The web server first predicts possible catalytic residues based on information extracted from the input structure. Users can optionally choose to combine the structure-based results with evolutionary information from PSI-Blast position-specific substitution matrix. The web server provides a one-click link for users to submit a PSI-Blast search and the evolutionary information is automatically combined with the structure-based prediction results when PSI-Blast search is finished. The computation results include possible catalytic residues ranked by their scores, which are calculated based on various sequence and structure features. The detailed scoring of each feature is also provided for users to judge and interpret the prediction results. In addition to raw computation data, the web server visualizes structures around the predicted catalytic residues for users to easily inspect the regions in which they are interested. Figure 1 shows a brief overview of the web server.

3. Methods

The method uses the following features to predict catalytic residues: residue side chain orientation, theoretical structural flexibility, and amino acid combination. The success of the method is to scan one small region of the structure at a time, instead of just considering the properties of each single residue. For each region that is probable to be catalytic site, we calculate the side chain orientation of residues in the region. A region is more probable to be catalytic site if the side chain of residues in the region tends to point to the center of the region. In addition to side chain orientation, we also calculate the structure flexibility and amino acid combination for each region. The following sections explain the detailed calculations.

3.1. Side Chain Orientation. A vector s_k is defined as the side chain direction of residue k:

$$s_k = \mathrm{X}_k^F - \mathrm{X}_k^{\mathrm{CA}}, \tag{1}$$

where X_k^F and $\mathrm{X}_k^{\mathrm{CA}}$ are the crystallographic position of the side chain vector atom and Cα atom of residue k. The side chain vector atom is carefully chosen for each amino acid type. It is either the most frequent functional atom defined in catalytic site atlas (CSA) [39] or the atom located near the centroid of multiple possible functional atoms. Here, side chain vector atoms are only defined for residues whose functional atom is usually on its side chain: Arg: CZ, Asn: CG, Asp: CG, Cys: SG, Gln: CD, Glu: CD, His: NE, Lys: NZ, Ser: OG, Thr: OG, Trp: CZ, and Tyr: OH. Only amino acid types defined here are included in the calculations of side chain orientations.

3.2. Prediction Procedure. All nonprotein ligands are removed from the input structure. The structure is then embedded in a three-dimensional $30 \times 30 \times 30$ grid of points. The reason of using fixed grid spacing is that we want to make sure the server finishes the calculations in reasonable time even for larger proteins. The grid size is the optimal balance between computation time and prediction performance. The grid spacing is from 1.33 angstrom to 2.13 angstrom depending on the protein size and is small enough to scan possible catalytic site even for large proteins. Each grid point is a probable position of catalytic site. For

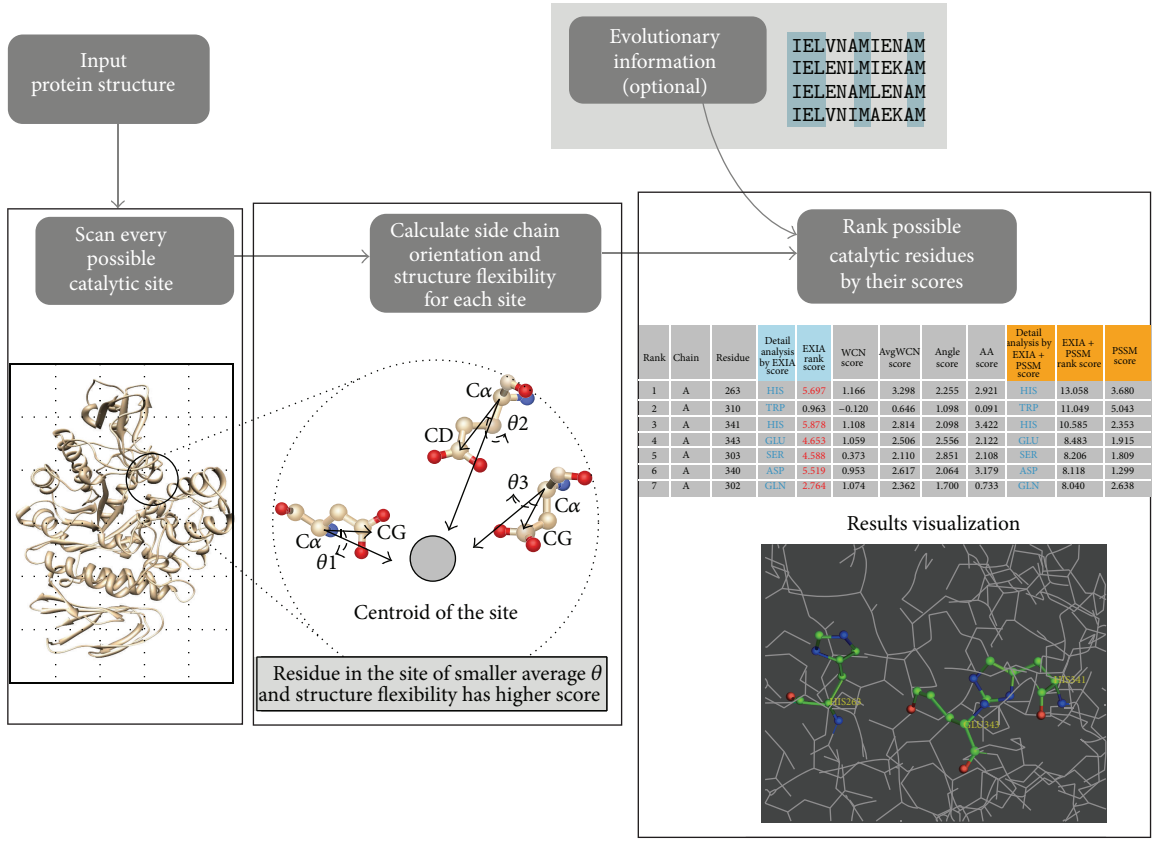


FIGURE 1: Overview of the EXIA2 web server. The input of the EXIA2 web server is a protein structure. EXIA2 first scans all possible catalytic sites in the structure and then computes the side chain orientation of residues in each candidate site. A residue receives higher score when it and its neighboring residues have their side chains pointing to their center position. The structure flexibility of residues is also included in scoring. After the structure-based calculation is finished, users may optionally add the sequence conservation from PSI-Blast, which usually takes longer calculation time. The final results are the possible catalytic residues ranked by their scores and all the detailed scores (side chain orientation score, structure flexibility score, and sequence conservation score) calculated in the prediction process.

each grid point i, the *surrounding residues* of point i are the residues whose distance between its Cα atom and point i is less than 10Å. Grid points with less than three surrounding residues are removed. For each point i and any one of its surrounding residues j, the vector between point i and Cα atom of residue j is defined as

$$v_{ij} = X_i - X_j, \tag{2}$$

where X_i and X_j are the position of point i and Cα atom of residue j. We compute the angle θ_{ij} between v_{ij} and s_j, which is the side chain vector of residue j:

$$\theta_{ij} = \operatorname{acos} \frac{v_{ij} \cdot s_j}{\|v_{ij}\| \|s_j\|}. \tag{3}$$

For a grid point within the area of the catalytic site, its surrounding residues usually have smaller θ angles. Based on this observation, we calculate the averaged angle θ_i among all of the surrounding residues for point i, as in

$$\theta_i = \sum \frac{\theta_{ij}}{N}, \tag{4}$$

where N is the number of surrounding residues of point i. We assume that points near catalytic site have smaller averaged θ and remove the points that have averaged $\theta > 80^\circ$. The cut-off value is chosen based on a statistics of θ angles for all residues in the PW79 dataset (as shown in Figure S1 in Supplementary Material available online at http://dx.doi.org/10.1155/2014/807839). About 80% of catalytic residues have the angle $\theta \leq 80$ degrees. We tried different cut-off values ranging from 30 to 100 degrees and found that the prediction performance is the best when the cut-off value is 80 degrees on the PW79 dataset. For every remaining point (points with $\theta \leq 80^\circ$), we select three residues at a time from its surrounding residues and give each selected residue a score (referred to as *feature score*) according to their features. For each point, the selection process is repeated for all possible combinations of any three surrounding residues (referred to as triplet). Each time a residue is involved in a selected triplet, the residue receives a feature score based on the features of selected triplets. A residue is possible to be the surrounding residue of multiple grid points and therefore involved in triplets that belong to different grid points. An example of score calculation is

available in supplementary Figure S2. Residues are finally ranked by their sum of all feature scores (denoted by S) received from all grid point that includes the residue. The final result is a list of residues ranked by their S score, that is, the likelihood of being a catalytic residue according to our prediction.

3.3. Feature Scores. The feature score is calculated based on theoretical structural flexibility and amino acid combination of residues. The weighted-contact number model (WCN) [41, 42] is used to compute structural flexibility. Catalytic residues usually have more rigid structures, that is, having higher WCN [43–45]. B-factor is not directly used because, in many cases, crystal structures of the same enzyme have almost identical 3D structure but have quite different B-factor profiles. In addition, B-factor is only available for structures solved by X-ray crystallography but not available for structures solved by NMR. WCN is more reliable than B-factor to measure the structural flexibility. For a residue k in a structure, its WCN w_k is defined as

$$w_k = \sum_{m \neq k} \frac{1}{r_{km}^2}, \tag{5}$$

where m is any other residues in the structure and r_{km} is the distance between the Cα atoms of residues k and m. The WCN scores are normalized as in

$$z_k^w = \frac{w_k - \overline{w}}{\sigma}, \tag{6}$$

where z_k^w is the normalized WCN of residue k and $\overline{w}$ and σ are the mean and standard deviation of the WCN of all residues in the protein. As described in the previous section, for every remaining point with N surrounding residues, we select three residues (triplet) from surrounding residues and give each selected residue a feature score. The purpose is to give higher score to residues involved in "better" combination, that is, triplet that are more structurally rigid. For a selected triplet (denoted by n, a subset of the N surrounding residues), we define an averaged WCN w_n, which is the average structure flexibility of these three residues:

$$w_n = \sum_{j \in n} \frac{z_j^w}{3}, \tag{7}$$

where z_j^w is the normalized WCN, w_j, of residue j. Among the three residues, each residue receives a feature score S:

$$S = w_n + z_j^w + z_j^a, \tag{8}$$

where w_n is the averaged WCN, z_j^w is the normalized WCN of residue j, and z_j^a is the normalized amino acid combination score of residue j. The amino acid combination score is based on statistics of the amino acids of catalytic sites in the PW79 dataset [37]. For each type of amino acid, a profile p_x containing 20 elements is constructed:

$$p_x = \left(p_x^{\mathrm{ALA}}, p_x^{\mathrm{CYS}}, p_x^{\mathrm{GLY}}, \ldots, p_x^{\mathrm{VAL}}\right), \tag{9}$$

where p_x denotes the profile of amino acid type x; each element in the profile is the frequency of an amino acid type appearing in the same catalytic site as amino acid type x. Here residues are defined as in the same catalytic site if they are all annotated as catalytic residue and located in the same catalytic site defined by the CSA database. The z_j^a score in (8) is calculated as in

$$z_j^a = \frac{a_j - \overline{a}}{\sigma^a} = \frac{\left(p_x^y + p_x^z\right) - \overline{a}}{\sigma^a}, \tag{10}$$

where a_j denotes the amino acid combination score of residue j and $\overline{a}$ and σ^a are the mean and standard deviation of the amino acid combination score of all residues in the protein. a_j is the sum of p_x^y and p_x^z, where x is the amino acid type of residue j and y and z are the amino acid types of the other two residues in the subset n, which has three selected residues as described previously.

Most catalytic residues have their functional atom on the side chain; there are about 5% of catalytic residues that have functional atom on the backbone. These catalytic residues are usually hydrophobic and nonpolar amino acids and are not involved in the calculations of side chain orientations. In the results of the web server, we provide users the structural flexibility and amino acid combination scores for these residues.

3.4. Evolutionary Information. The method becomes more powerful by including evolutionary information. Users may include evolutionary sequence conservation from PSI-Blast [28] position-specific substitution matrix (PSSM) to the prediction. EXIA2 web server provides users a one-click option to submit PSI-Blast search on the web server. The sequence conservation information is automatically combined with structure-based features. PSI-Blast is set to search against the nonredundant (nr) database for three iterations with an E-value threshold of 5×10^{-3}. The nr database is a default built-in protein sequence database in PSI-Blast. The sequence conservation score c_j of residue j is directly taken from the "information per position" column in PSSM. The sequence conservation score c_j is included in the final score S_j of residue j as in

$$S_j' = S_j + 1.6 \times z_j^c, \tag{11}$$

where S_j is the final score of residue j based on structure information and z_j^c is the normalized c_j of residue j as in

$$z_j^c = \frac{c_j - \overline{c}}{\sigma^c}, \tag{12}$$

where c_j is the original sequence conservation score of residue j and $\overline{c}$ and σ^c are the mean and standard deviation of all the sequence conservation scores in the protein. The parameter in (11) was tuned based on the PW79 dataset. The prediction performance (AUCROC) is optimal for the dataset when the value is set to 1.6.

Table 1: Comparison of prediction performance of EXIA2 web server and competing method using both sequence and structure information.

	Benchmark datasets				
	PW79	POOL160	EF fold	EF superfamily	EF family
Competing method[1]					
Recall (R)	46.0	78.1	64.2	67.3	61.7
Precision (P)	28.0	19.0	17.1	16.9	18.5
AUCROC	0.963	0.948	—	—	—
EXIA2 + PSSM[2]					
Recall at equal P	63.3	77.8	71.0	70.4	64.3
Precision at equal R	36.5	18.7	19.4	17.7	19.6
AUCROC	0.973	0.965	0.968	0.968	0.968
EXIA2[3]					
Recall at equal P	48.8	68.6	43.8	46.9	42.2
Precision at equal R	30.5	14.5	12.0	11.6	12.9
AUCROC	0.962	0.960	0.943	0.944	0.946

[1]Prediction results of Cilia and Passerini [36].
[2]Prediction results of EXIA2 combined with PSI-Blast PSSM.
[3]Prediction results of EXIA2 without evolutionary information.

3.5. Datasets. The PW79, POOL160, EF, and P100 datasets are from [35, 40, 46, 47], respectively. The proteins in the L55 dataset (Table S1) are selected from the PW79, POOL160, and EF datasets. Among all proteins in these datasets, we first picked the proteins without bounding ligand (78 proteins selected). Then for each protein, a structure that has the most similar sequence (most of them have completely the same sequence) and has bounding ligand in the catalytic pocket was selected from the PDB database. 23 proteins among the 78 structures that have no available structure with bounding ligand in PDB are removed. There are totally 55 pairs of enzyme structures selected as the L55 dataset. The EX79 dataset (Table S2) is built by combining all proteins in the POOL160 and EF datasets and excluding all proteins that are in the PW79 dataset. The definition of catalytic residues is based on Catalytic Site Atlas version 2.2.12.

4. Performance

We compared the prediction performance of EXIA2 web server with that of three state-of-the-art prediction methods on six benchmark datasets [37], PW79, POOL160, EF fold, EF superfamily, EF family, and P100 which include over 1,200 proteins and 861,404 residues (3,664 catalytic residues and 857,740 noncatalytic residues). Tables 1 and 2 summarize the comparison of prediction performances, including recall (R), precision (P), and area under ROC curve (AUCROC).True positives are correctly predicted catalytic residues; false positives are noncatalytic residues incorrectly predicted to be catalytic residues; true negatives are correctly predicted noncatalytic residues; false negatives are catalytic residues incorrectly predicted to be noncatalytic residues. For each protein, we calculate the prediction result under different cut-off values (true positive rate from 0 to 1) and draw the ROC curve and recall-precision curve. The overall prediction result of a dataset is by averaging the per-protein ROC curves (or recall-precision curves) in the dataset. When comparing the prediction performance with other methods, the recall and precision values are directly retrieved from the overall recall-precision curve for a dataset.

4.1. Comparison with Other Methods. Table 1 compares the prediction results of EXIA2 and the results of a prediction method [36], which uses many sequence, structure, and evolutionary features to model residue structural neighborhood. The prediction performance of EXIA2 combined with PSI-Blast evolutionary information (PSSM) is better than that of the competing method. Among the six benchmark datasets, the recall (or precision) is higher than that of the competing method when the precision (or recall) is equal to theirs. EXIA2 also has higher AUCROC than theirs in the PW79 and POOL160 datasets (the AUCROC for the other three datasets are not provided in the report of the competing method).

We also compare the prediction results of EXIA2 web server *without* using PSSM information and those of two other prediction methods: POOL that uses only structure information and CRpred [27] that uses only sequence information. We compared the precision when our recall is equal to theirs and the recall when our precision is equal to theirs (Table 2). The results show that EXIA2 performs better than these two methods. It has higher recall (70.8) and higher precision (20.2) when the precision and recall are equal to those of POOL (18.1 and 61.7, resp.) for the POOL160 dataset. Most current prediction methods do not perform well when only structure-based features are used. Evolutionary information is usually required for prediction methods to have better prediction results. POOL [35], which calculates theoretical residue electrostatic property and structure shape, is one of the best structure-based prediction methods. EXIA2 performs better than POOL for the POOL160 dataset. The results indicate that EXIA2, which uses side chain orientation and structure flexibility, is more effective than the structure features used by POOL. In addition to prediction performance, EXIA2 web server is more computationally efficient

Table 2: Comparison of prediction performance of EXIA2 web server and CRpred, POOL, ConSurf, and ResBoost.

Competing method	CRpred	POOL[1]	POOL[2]	ConSurf[3]	ResBoost[3]
Benchmark datasets	EF fold	POOL160	POOL160	P100	P100
Recall (R)	48.2	61.7	64.7	55.0	55.0
Precision (P)	17.0	18.1	19.1	5.0	17.0
AUCROC[4]	—	0.907	0.925	—	—
EXIA2 + PSSM[5]					
Recall at equal P	72.7	80.0	77.8	96.0	74.3
Precision at equal R	27.3	24.3	23.3	25.3	25.3
AUCROC	0.968	0.965	0.965	0.966	0.966
EXIA2[6]					
Recall at equal P	45.1	70.8	68.6	90.6	58.0
Precision at equal R	16.2	22.2	20.8	18.3	18.3
AUCROC	0.943	0.960	0.960	0.952	0.952

[1]Prediction results of POOL.
[2]Prediction results of POOL combined with evolutionary information.
[3]Prediction results published in [40]. Complete comparison of ROC and recall-precision curves is available in supplementary Figure S4.
[4]Some AUC values are not available in the publications.
[5]Prediction results of EXIA2 combined with PSI-Blast PSSM.
[6]Prediction results of EXIA2 without evolutionary information.

than POOL web server [48]. The prediction results of EXIA2 web server are usually displayed instantly. POOL web server usually needs several minutes to finish the calculations. We compared the computation time used by EXIA2 and POOL by submitting the 79 proteins in the PW79 dataset to the two web servers. For the POOL web server, the submission of three proteins (PDB ID: 1B57, 1DCO, and 1DQS) did not finish correctly (server crash due to parameter errors during calculation). These proteins are excluded in the comparison. The average computation time of EXIA2 is 6.25 seconds and that of POOL server is 868.14 seconds (the time for generating PSI-Blast profile not included). For POOL web server, the computation time of 46% of proteins in the dataset is more than 600 seconds. For EXIA2 web server, the maximum computation time is 25 seconds on a protein of about 3000 residues. The results show that EXIA2 web server is very efficient and stable. Figure 2 shows the distribution of computation time for the EXIA2 and POOL web server for the PW79 dataset.

CRpred is currently the best sequence-based catalytic residue prediction method. In the prediction of protein catalytic residue, prediction results using only sequence information are usually much better than the results only using structure information. The reason may be that sequence information includes evolutionary conservation and catalytic residues are usually highly evolutionarily conserved. Here, the prediction results of EXIA2 without adding evolutionary conservation are better than those of CRpred. EXIA2 has higher recall (67.8) and higher precision (24.7) when the precision and recall is equal to those of CRpred (17.5 and 53.7, resp.) for the PW79 dataset. It also performs better than CRpred on the EF fold dataset. Although EXIA2 has slightly smaller recall and precision values than those reported by CRpred (Table 2), it still performs better than CRpred by looking at their ROC curve (Figure S3), which is a much more complete performance measure. The results show that the structure features used by EXIA2 are very effective.

We also compare the performance of EXIA2 with that of ConSurf [29], ResBoost [40], and a recent structure-based prediction method [49] on two test datasets. ConSurf and ResBoost are both based on evolutionary conservation, various sequence, and structure features. ConSurf identifies functionally important regions in proteins by estimating the degree of conservation of the amino acid sites among their close sequence homologues. ResBoost predicts catalytic residues based on several features, including evolutionary conservation, 3D clustering, residue solvent accessibility, and hydrophobicity. EXIA2 server performs better than both ConSurf and ResBoost even without using sequence conservation information (Table 2) on the P100 dataset. The comparisons of their ROC curves and recall-precision curves are available in supplementary Figure S4. Another recent structure-based prediction method [49] is based on various centrality measures of nodes in graphs of interacting residues: closeness, betweenness and page-rank centrality, general center of mass of the structure, relative solvent accessibility, and sequence conservation. EXIA2 also performs better than the method on a test set of 29 proteins. EXIA2 has higher precision (21.7 versus 17.1) when the sensitivity is equal to theirs and has higher sensitivity (71.9 versus 63.1) when the precision is equal to theirs. However, the prediction results of the method without sequence conservation are not available on the test dataset.

4.2. Performance for Enzyme Structure without Bounded Ligand. We construct a dataset (L55, PDB IDs listed in Table S1) that contains 55 enzymes and their structures crystallized with substrates (denoted by L55-Bound) and without substrates (L55-Unbound). Figure 3 shows the ROC curves for the structures of L55-Bound and L55-Unbound. The performances of EXIA2 on these two sets of structures are very

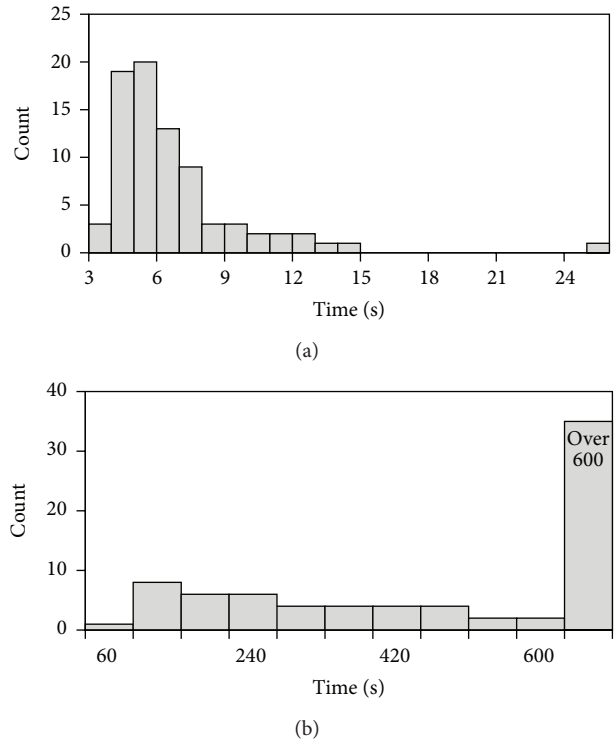


Figure 2: Distribution of computation time for the PW79 dataset. The figure shows the computation time for (a) EXIA2 web server and (b) POOL web server.

similar. The AUCROC for L55-Bound and L55-Unbound are 0.968 and 0.967, respectively, when both structure features (side chain orientation and flexibility) and sequence conservation are used. The AUCROC are 0.950 and 0.947 for L55-Bound and L55-Unbound, respectively, when only structure features are used to perform prediction. The results suggest that the special side chain orientation of catalytic residue exists not only in substrate-bounded structures but also in structures without bounded ligands. In Figure 4, we further analyze the angle between the side chain vector of catalytic residue and the vector of the residue Cα atom to the center of the catalytic site (as the θ angle described in Figure 1 or (3)). Residues whose side chain tends to point to the center of the catalytic site have smaller angles. For catalytic residues of both substrate-bounded (orange bar) and substrate-unbounded (green bar) structure, their angles are smaller than those of noncatalytic residues. For the angle calculation of noncatalytic residues, we randomly pick noncatalytic residues and include its structurally neighboring residues within 10 angstroms. For each random "noncatalytic site" selected, we calculate the angle between the side chain vector of these residues and the vector from their Cα atom to the center of the site. The results indicate that the special side chain orientation only exists in catalytic residues but not in noncatalytic residues. More importantly, the results also suggest that the side chain structures of catalytic residues are ready to interact with substrates even before substrate binding. The observation also explains the success of EXIA2 to identify the catalytic residues for enzymes without bounded ligands.

4.3. Effect of Amino Acid Combination Feature. In the EXIA2 server, we add the amino acid combination feature, which is a general composition of amino acid types in enzyme catalytic sites. The scoring of the feature is calculated based on the enzymes in the PW79 dataset. To evaluate the performance of the feature, we construct a dataset (EX79 dataset, PDB IDs listed in Table S2) that combines the POOL160, EF fold, EF superfamily, and EF family datasets and excludes all of the enzymes from the PW79 dataset. Figure 5 shows the receiver operating characteristic (ROC) curve for the PW79, POOL160, EF fold, and the EX79 dataset. The ROC curve of EF family and EF superfamily is similar to that of EF fold and not shown in the figure. The results show that the performance (AUCROC = 0.964) of the EX79 dataset is similar to that of the EF fold dataset (AUCROC = 0.968). It suggests that the feature is still effective for enzymes that were not used to calculate the amino acid combination feature.

To see the effect of amino acid combination feature on the prediction performance, we compare the AUCROC of prediction using structure feature only and using both structure feature and amino acid combination. The AUCROC is improved from 0.938 to 0.944 on the EX79 dataset. Figure S5 shows the ROC curve of prediction with and without amino acid combination feature on the EX79 dataset. The TPR values are improved especially when FPR is smaller than 0.15. EXIA2 is primarily based on the intrinsic structure features, side chain orientation, and structure flexibility, of the input protein. In this work, the amino acid combination

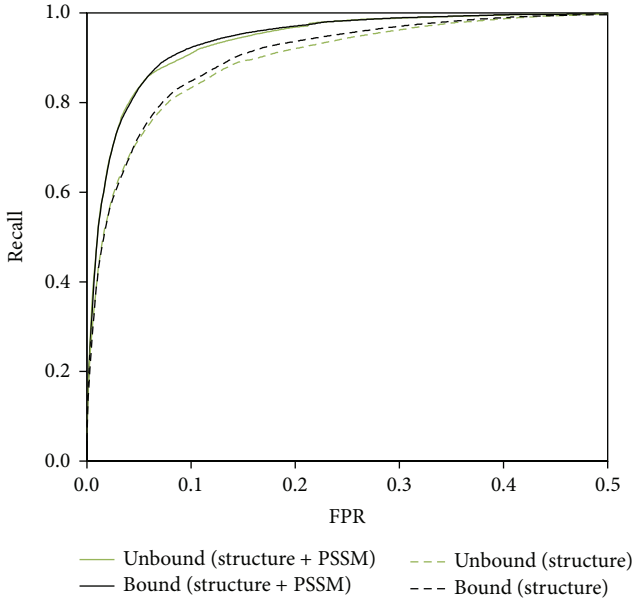


Figure 3: ROC curves for L55-Bound and L55-Unbound enzyme structures. Comparison of performance on 55 enzymes in their substrate-bounded (L55-Bound) form and substrate-unbounded form (L55-Unbound). The performances of L55-Unbound are similar to those of L55-Bound using only structure features (dashed lines) or using both structure features and sequence conservation (solid lines). The AUCROC for L55-Bound and L55-Unbound are 0.968 and 0.967, respectively, when both structure features and sequence conservation are used. The AUCROC are 0.950 and 0.947 for L55-Bound and L55-Unbound, respectively, when only structure features are used.

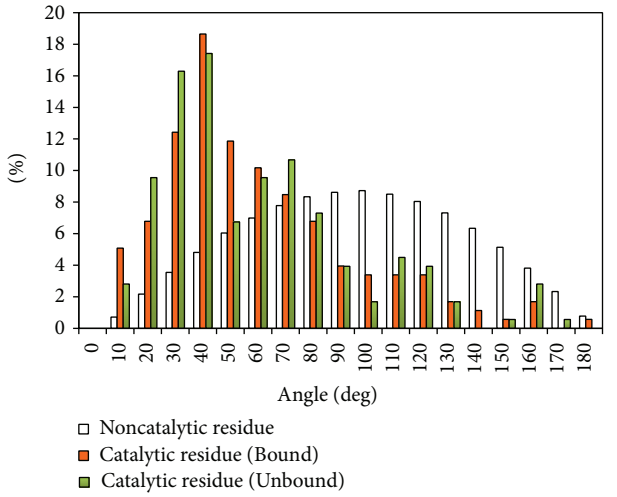


Figure 4: Comparison of side chain orientation of catalytic and noncatalytic residues. Comparison of side chain orientation for catalytic residues in structures of L55-Bound (orange bar) and L55-Unbound (green bar) set and random selected noncatalytic sites (white bar). Residues with small angles have their side chain pointing to the center of the site (see Figure 1 for the definition of angle calculation).

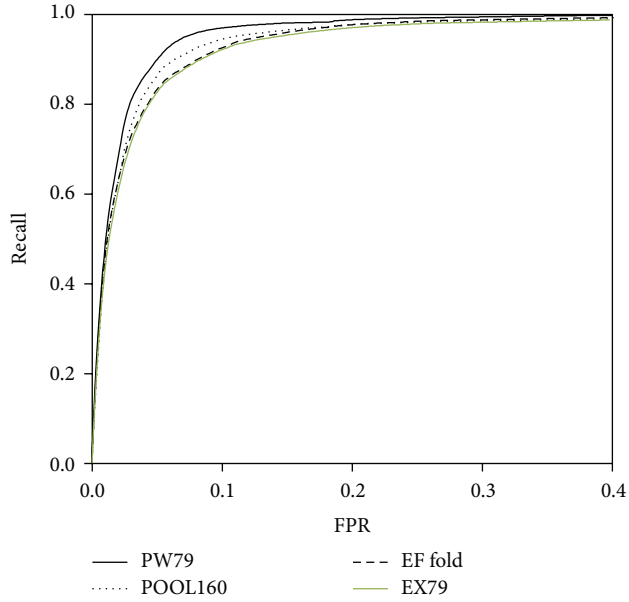


Figure 5: ROC curves for PW79, POOL160, EF fold, and EX79 dataset. The prediction performance using structure features combined with sequence conservation. The AUCROC for the PW79, POOL160, EF fold, and EX79 dataset are 0.973, 0.965, 0.968, and 0.964, respectively.

feature is added to the web server because of its practical usage to identify possible catalytic site.

5. Prediction Examples

5.1. Human Ferrochelatase. The catalytic site of Human ferrochelatase (PDB ID: 1HRK) includes three catalytic residues, H263, H341, and E343. Figures 6(a) and 6(b) show the structures of the catalytic site and demonstrate a good example of the side chain orientations of catalytic residues. The side chain of the three catalytic residues point to the center of the catalytic site to interact with the ligand (ligand information is not used in the prediction). Catalytic residues H341, H263, and E343 are ranked 1st, 2nd, and 5th, respectively, in the prediction using only structure information. The output results are shown in Figure 7. Although the two noncatalytic residues, D340 and E369, are ranked 3rd and 4th, they have low WCN score (more flexible structure) and are less likely to be catalytic residues. The prediction results are further improved by adding evolutionary information (PSI-Blast PSSM). Catalytic residues H263, H341, and E343 are ranked 1st, 3rd, and 4th, respectively (Figure 7). The noncatalytic residue W310 is ranked 2nd because it is extremely evolutionarily conserved (Figure 6(b)). However, it has very low WCN score and is less probable to be catalytic residue (Figure 6(a)). The catalytic residues of the enzyme are correctly predicted because they have stable structure and proper side chain orientations. One of the successful designs of EXIA2 is that we consider not only the properties of single residue but also the properties of its neighboring residues. A residue receives high score when the residue and its neighbors have their side chain

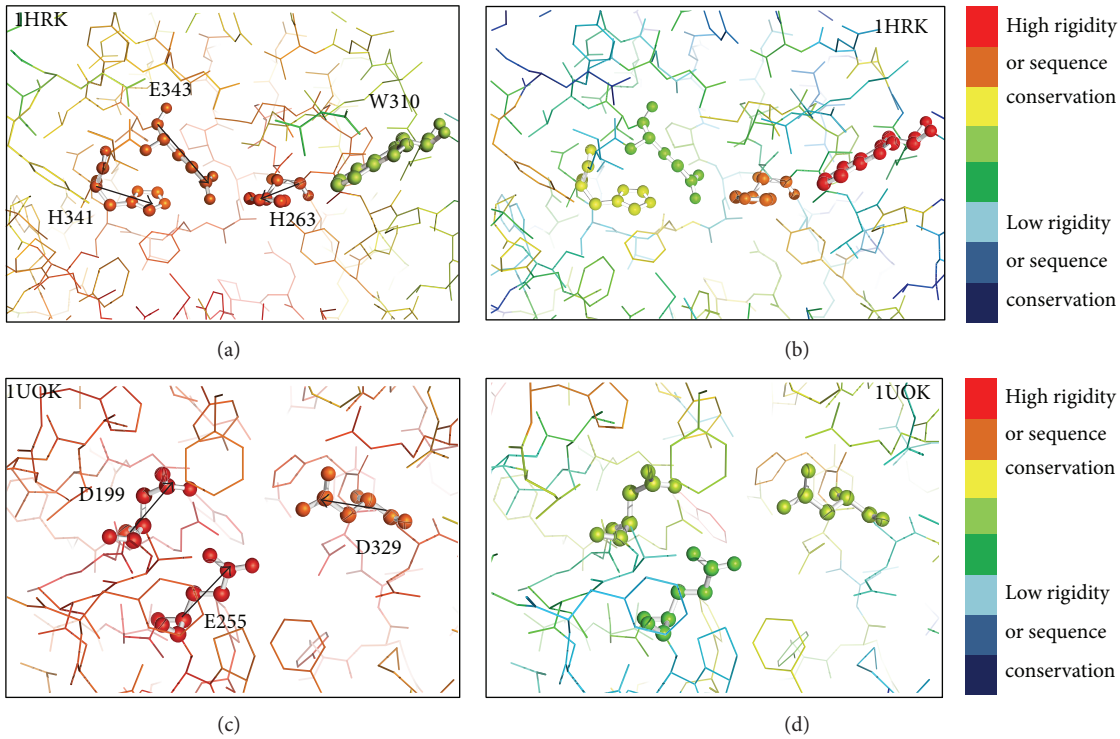


Figure 6: Structures around the catalytic residues of two example enzymes colored by residue structure flexibility and sequence conservation. The structures around the catalytic residues of Human ferrochelatase (PDB: 1HRK) and Oligo-1,6-glucosidase (PDB: 1UOK). The rainbow coloring in (a) and (c) is from blue (low structure rigidity or high structure flexibility) to red (high structure rigidity or low structure flexibility). (b) and (d) are colored from blue (low sequence conservation) to red (high sequence conservation). Structure flexibility and sequence conservation are based on the WCN model and PSI-Blast PSSM as described in Section 3. The black arrows in the figure indicate the direction of side chain vector for the catalytic residues.

pointing to their centroid position and their average structure flexibility is low.

5.2. Oligo-1,6-glucosidase. The catalytic residues of oligo-1,6-glucosidase (PDB ID: 1UOK) are D199, E255, and D329. They are the top three ranked residues in the prediction results using only structure information. Each of the three catalytic residues has low structural flexibility (Figure 6(c)). In addition, they also have high average WCN score, which means that these residues and their neighboring residues form very stable structures. The enzyme shows a good example on the effect of calculating *average WCN score*. There are several noncatalytic residues that have better WCN score (the structure flexibility of the residue itself) than the three catalytic residues, but these catalytic residues have higher average WCN score (the average structure flexibility of the residue and its neighboring residues) than all the other residues in the enzyme. It suggests that considering the structural flexibility of single residue is not enough in the prediction of catalytic residues. The three catalytic residues also have extremely high side chain orientation score; that is, these residues and their neighboring residues have side chains pointing to their centroid (Figure 6(c)). The side chain orientation score of these catalytic residues is higher than those of all the other residues in the enzyme. Side chain orientation score helps to easily identify the most probable catalytic residues in this example. It also suggests that the side chain orientation feature is unique enough to be used in the prediction of catalytic residues because noncatalytic residues do not seem to have such property.

6. Conclusion

EXIA2 is an accurate and efficient catalytic residue prediction method. In addition to accurate identification of catalytic residues, the web server provides detailed scoring data, including the side chain orientation, structural flexibility, amino acid combination, and sequence conservation scores, for users to inspect and analyze the enzyme structure.The advantage of EXIA2 is that it does not rely on sequence or structure homology information. The fundamental feature used in EXIA2 is to detect the regions in which the residues' side chain points to the center of the region. We found that the special side chain orientation is usually observed for catalytic residues but not for noncatalytic residues. The prediction performance based on the phenomenon is better than those of existing prediction methods and is tested on various datasets, including a dataset of enzymes that do not have any bounded ligand in their crystallographic structures. The results suggest that the special side chain orientation exists not only in ligand-bounded structure but also in the apo form of enzymes.

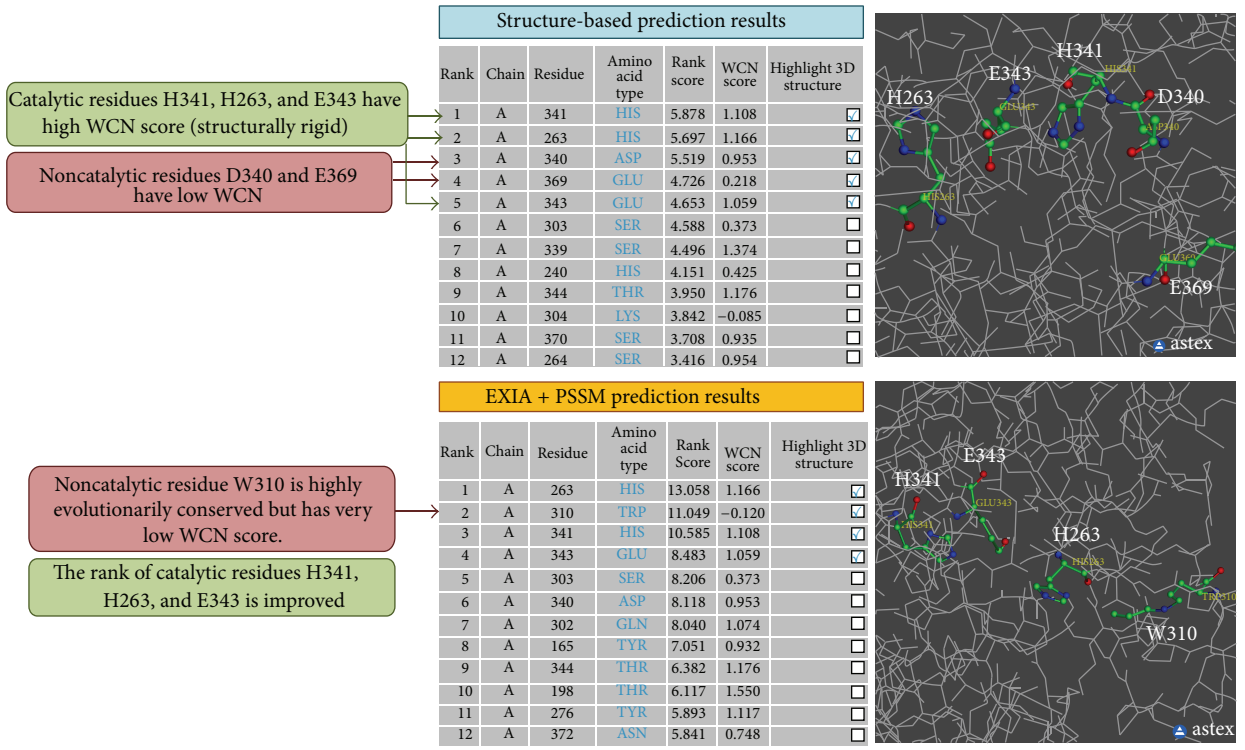


FIGURE 7: Server output results of protein Human ferrochelatase. The figure shows part of the output results of the EXIA2 web server for Human ferrochelatase (PDB: 1HRK). The main results are possible catalytic residues predicted by the server ranked by their rank score. The prediction results are improved when evolutionary information from PSI-Blast PSSM is included in the prediction. The WCN score of each residue is also provided for users to further analyze the results. The detailed scores, including side chain orientation score, average structure flexibility, and PSSM scores, used in the prediction are also provided (not shown here).

EXIA2 is different from other competing machine learning methods (except POOL, which is also a heuristic-based approach). The performance of EXIA2 is mostly contributed from the intrinsic properties of input structure, the side chain orientation, and structure flexibility feature. There is no training process required to calculate these structure features. The prediction performance only based on these structure features is more accurate than those of other existing structure-based methods. Although there are few parameters that need to be optimized, most of them are based on statistics and observation of general enzyme properties. We also used the EX79 dataset, which exclude the 79 proteins used for parameter optimization, to test the performance of EXIA2. The results show that performance on EX79 dataset is similar to those of the EF fold, EF family, and EF superfamily datasets.

Conflict of Interests

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

References

[1] M. N. Wass and M. J. E. Sternberg, "ConFunc—functional annotation in the twilight zone," *Bioinformatics*, vol. 24, no. 6, pp. 798–806, 2008.

[2] W. T. Clark and P. Radivojac, "Analysis of protein function and its prediction from amino acid sequence," *Proteins: Structure, Function and Bioinformatics*, vol. 79, no. 7, pp. 2086–2096, 2011.

[3] T. Hawkins, S. Luban, and D. Kihara, "Enhanced automated function prediction using distantly related sequences and contextual association by PFP," *Protein Science*, vol. 15, no. 6, pp. 1550–1556, 2006.

[4] F. Pazos and M. J. E. Sternberg, "Automated prediction of protein function and detection of functional sites from structure," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 41, pp. 14754–14759, 2004.

[5] R. A. Laskowski, J. D. Watson, and J. M. Thornton, "Protein function prediction using local 3D templates," *Journal of Molecular Biology*, vol. 351, no. 3, pp. 614–626, 2005.

[6] A. Vazquez, A. Flammini, A. Maritan, and A. Vespignani, "Global protein function prediction from protein-protein interaction networks," *Nature Biotechnology*, vol. 21, no. 6, pp. 697–700, 2003.

[7] S. Letovsky and S. Kasif, "Predicting protein function from protein/protein interaction data: a probabilistic approach," *Bioinformatics*, vol. 19, supplement 1, pp. i197–i204, 2003.

[8] B. E. Ersgelhardt, M. I. Jordan, K. E. Muratore, and S. E. Brersfser, "Protein molecular function prediction by bayesian phylogenomics," *PLoS Computational Biology*, vol. 1, article e45, 2005.

[9] M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates, "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles," *Proceedings of*

the National Academy of Sciences of the United States of America, vol. 96, no. 8, pp. 4285–4288, 1999.

[10] P. Gaudet, M. S. Livstone, S. E. Lewis, and P. D. Thomas, "Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium," *Briefings in Bioinformatics*, vol. 12, no. 5, pp. 449–462, 2011.

[11] B. Huang, "Metapocket: a meta approach to improve protein ligand binding site prediction," *A Journal of Integrative Biology*, vol. 13, no. 4, pp. 325–330, 2009.

[12] S. Pérot, O. Sperandio, M. A. Miteva, A.-C. Camproux, and B. O. Villoutreix, "Druggable pockets and binding site centric chemical space: a paradigm shift in drug discovery," *Drug Discovery Today*, vol. 15, no. 15-16, pp. 656–667, 2010.

[13] M. N. Wass, L. A. Kelley, and M. J. E. Sternberg, "3DLigandSite: predicting ligand-binding sites using similar structures," *Nucleic Acids Research*, vol. 38, no. 2, pp. W469–W473, 2010.

[14] M. L. Miller and N. Blom, "Kinase-specific prediction of protein phosphorylation sites," *Methods in Molecular Biology*, vol. 527, pp. 299–310, 2009.

[15] F. Gnad, L. M. F. de Godoy, J. Cox et al., "High-accuracy identification and bioinformatic analysis of *in vivo* protein phosphorylation sites in yeast," *Proteomics*, vol. 9, no. 20, pp. 4642–4652, 2009.

[16] I. Ezkurdia, L. Bartoli, P. Fariselli, R. Casadio, A. Valencia, and M. L. Tress, "Progress and challenges in predicting protein-protein interaction sites," *Briefings in Bioinformatics*, vol. 10, no. 3, pp. 233–246, 2009.

[17] M. Sikic, S. Tomic, and K. Vlahovicek, "Prediction of protein-protein interaction sites in sequences and 3D structures by random forests," *PLoS Computational Biology*, vol. 5, Article ID e1000278, 2009.

[18] Q. C. Zhang, L. Deng, M. Fisher, J. Guan, B. Honig, and D. Petrey, "PredUs: a web server for predicting protein interfaces using structural neighbors," *Nucleic Acids Research*, vol. 39, no. 2, pp. W283–W287, 2011.

[19] P. Radivojac, V. Vacic, C. Haynes et al., "Identification, analysis, and prediction of protein ubiquitination sites," *Proteins*, vol. 78, no. 2, pp. 365–380, 2010.

[20] J. A. Capra and M. Singh, "Predicting functionally important residues from sequence conservation," *Bioinformatics*, vol. 23, no. 15, pp. 1875–1882, 2007.

[21] J. D. Fischer, C. E. Mayer, and J. Söding, "Prediction of protein functional residues from sequence by probability density estimation," *Bioinformatics*, vol. 24, no. 5, pp. 613–620, 2008.

[22] D. La, B. Sutch, and D. R. Livesay, "Predicting protein functional sites with phylogenetic motifs," *Proteins*, vol. 58, no. 2, pp. 309–320, 2005.

[23] M. Ota, K. Kinoshita, and K. Nishikawa, "Prediction of catalytic residues in enzymes based on known tertiary structure, stability profile, and sequence conservation," *Journal of Molecular Biology*, vol. 327, no. 5, pp. 1053–1064, 2003.

[24] S. Pande, A. Raheja, and D. R. Livesay, "Prediction of enzyme catalytic sites from sequence using neural networks," in *Proceedings of the IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology (CIBCB '07)*, pp. 247–253, 2007.

[25] B. Sterner, R. Singh, and B. Berger, "Predicting and annotating catalytic residues: an information theoretic approach," *Journal of Computational Biology*, vol. 14, no. 8, pp. 1058–1073, 2007.

[26] J. W. Torrance, G. J. Bartlett, C. T. Porter, and J. M. Thornton, "Using a library of structural templates to recognise catalytic sites and explore their evolution in homologous families," *Journal of Molecular Biology*, vol. 347, no. 3, pp. 565–581, 2005.

[27] T. Zhang, H. Zhang, K. Chen, S. Shen, J. Ruan, and L. Kurgan, "Accurate sequence-based prediction of catalytic residues," *Bioinformatics*, vol. 24, no. 20, pp. 2329–2338, 2008.

[28] S. F. Altschul, T. L. Madden, A. A. Schäffer et al., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[29] H. Ashkenazy, E. Erez, E. Martz, T. Pupko, and N. Ben-Tal, "ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids," *Nucleic Acids Research*, vol. 38, no. 2, Article ID gkq399, pp. W529–W533, 2010.

[30] D. Tautz and T. Domazet-Lošo, "The evolutionary origin of orphan genes," *Nature Reviews Genetics*, vol. 12, no. 10, pp. 692–702, 2011.

[31] A. Ben-Shimon and M. Eisenstein, "Looking at enzymes from the inside out: the proximity of catalytic residues to the molecular centroid can be used for detection of active sites and enzyme-ligand interfaces," *Journal of Molecular Biology*, vol. 351, no. 2, pp. 309–326, 2005.

[32] G. Amitai, A. Shemesh, E. Sitbon et al., "Network analysis of protein structures identifies functional residues," *Journal of Molecular Biology*, vol. 344, no. 4, pp. 1135–1146, 2004.

[33] S. Sacquin-Mora, É. Laforet, and R. Lavery, "Locating the active sites of enzymes using mechanical properties," *Proteins: Structure, Function and Genetics*, vol. 67, no. 2, pp. 350–359, 2007.

[34] Y. Wei, J. Ko, L. F. Murga, and M. J. Ondrechen, "Selective prediction of interaction sites in protein structures with THEMATICS," *BMC Bioinformatics*, vol. 8, article 119, 2007.

[35] W. Tong, Y. Wei, L. F. Murga, M. J. Ondrechen, and R. J. Williams, "Partial Order Optimum Likelihood (POOL): maximum likelihood prediction of protein active site residues using 3D structure and sequence properties," *PLoS Computational Biology*, vol. 5, no. 1, Article ID e1000266, 2009.

[36] E. Cilia and A. Passerini, "Automatic prediction of catalytic residues by modeling residue structural neighborhood," *BMC Bioinformatics*, vol. 11, article 115, 2010.

[37] Y.-T. Chien and S.-W. Huang, "Accurate prediction of protein catalytic residues by side chain orientation and residue contact density," *PLoS ONE*, vol. 7, no. 10, Article ID e47951, 2012.

[38] H. M. Berman, J. Westbrook, Z. Feng et al., "The protein data bank," *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, 2000.

[39] C. T. Porter, G. J. Bartlett, and J. M. Thornton, "The Catalytic Site Atlas: s resource of catalytic sites and residues identified in enzymes using structural data," *Nucleic Acids Research*, vol. 32, pp. D129–D133, 2004.

[40] R. Alterovitz, A. Arvey, S. Sankararaman, C. Dallett, Y. Freund, and K. Sjölander, "ResBoost: characterizing and predicting catalytic residues in enzymes," *BMC Bioinformatics*, vol. 10, article 197, 2009.

[41] C.-P. Lin, S.-W. Huang, Y.-L. Lai et al., "Deriving protein dynamical properties from weighted protein contact number," *Proteins: Structure, Function and Genetics*, vol. 72, no. 3, pp. 929–935, 2008.

[42] S.-W. Huang, C.-H. Shih, C.-P. Lin, and J.-K. Hwang, "Prediction of NMR order parameters in proteins using weighted protein contact-number model," *Theoretical Chemistry Accounts*, vol. 121, no. 3-4, pp. 197–200, 2008.

[43] Y.-T. Chien and S.-W. Huang, "On the structural context and identification of enzyme catalytic residues," *BioMed Research International*, vol. 2013, Article ID 802945, 9 pages, 2013.

[44] S.-W. Huang, S.-H. Yu, C.-H. Shih, H.-W. Guan, T.-T. Huang, and J.-K. Hwang, "On the relationship between catalytic residues and their protein contact number," *Current Protein and Peptide Science*, vol. 12, no. 6, pp. 574–579, 2011.

[45] Y.-T. Chien and S.-W. Huang, "Prediction of protein catalytic residues by local structural rigidity," in *Proceeding of the 6th International Conference on Complex, Intelligent, and Software Intensive Systems (CISIS '12)*, pp. 592–596, Palermo, Italy, July 2012.

[46] N. Petrova and C. Wu, "Prediction of catalytic residues using Support Vector Machine with selected protein sequence and structural properties," *BMC Bioinformatics*, vol. 7, article 312, 2006.

[47] E. Youn, B. Peters, P. Radivojac, and S. D. Mooney, "Evaluation of features for catalytic residue prediction in novel folds," *Protein Science*, vol. 16, no. 2, pp. 216–226, 2007.

[48] S. Somarowthu and M. J. Ondrechen, "POOL server: machine learning application for functional site prediction in proteins," *Bioinformatics*, vol. 28, no. 15, Article ID bts321, pp. 2078–2079, 2012.

[49] J. E. Fajardo and A. Fiser, "Protein structure based prediction of catalytic residues," *BMC Bioinformatics*, vol. 14, no. 1, article 63, 2013.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 764102, 11 pages
http://dx.doi.org/10.1155/2014/764102

Research Article

High-Throughput Functional Screening of Steroid Substrates with Wild-Type and Chimeric P450 Enzymes

Philippe Urban,[1,2,3] Gilles Truan,[1,2,3] and Denis Pompon[1,2,3]

[1] *Université de Toulouse, INSA, UPS, INP, LISBP, 135 Avenue de Rangueil, 31077 Toulouse, France*
[2] *INRA, UMR792 Ingénierie des Systèmes Biologiques et des Procédés, 31400 Toulouse, France*
[3] *CNRS, UMR 5504, 135 Avenue de Rangueil, 31400 Toulouse, France*

Correspondence should be addressed to Philippe Urban; urban@insa-toulouse.fr

Received 11 April 2014; Accepted 21 July 2014; Published 26 August 2014

Academic Editor: Wen-Chi Chang

The promiscuity of a collection of enzymes consisting of 31 wild-type and synthetic variants of CYP1A enzymes was evaluated using a series of 14 steroids and 2 steroid-like chemicals, namely, nootkatone, a terpenoid, and mifepristone, a drug. For each enzyme-substrate couple, the initial steady-state velocity of metabolite formation was determined at a substrate saturating concentration. For that, a high-throughput approach was designed involving automatized incubations in 96-well microplate with sixteen 6-point kinetics per microplate and data acquisition using LC/MS system accepting 96-well microplate for injections. The resulting dataset was used for multivariate statistics aimed at sorting out the correlations existing between tested enzyme variants and ability to metabolize steroid substrates. Functional classifications of both CYP1A enzyme variants and steroid substrate structures were obtained allowing the delineation of global structural features for both substrate recognition and regioselectivity of oxidation.

1. Introduction

In mammals, P450 enzymes of hepatocyte endoplasmic reticulum play a major role in the oxidative metabolism not only of xenobiotics (drugs, environmental pollutants, and phytochemicals) [1–4] but also of endogenous compounds such as steroids [5]. It has been shown that, similarly to drug chemicals, steroids are frequently transformed into several metabolites by P450 enzymes involved in drug metabolism [6, 7]. Progesterone is hydroxylated at the 6 and 21 positions by human CYP2C9 in addition to carbon-16 for human CYP3A4 and 16 and 17 for human CYP2C19 [8]. However, in contrast to the wide variety of drug shapes, steroids present a common chemical androstane skeleton which limits the structural diversity of the nature of substituting groups at the different available position of the androstane ring as shown in Figure 7.

The particular diversity of steroids therefore appears particularly adapted to explore the structural factors that control enzyme regiospecificity of action and, reciprocally, to identify the general features of steroid structures which control recognition by P450 enzymes.

We focused this work on natural and synthetic CYP1A enzymes and a collection of 16 steroidal substrates, 14 steroids, and 2 steroid analogues, a terpenoid and a drug. The effect of the diversity of steroid chemical structures was assessed by using different substrates of increasing complexity in their substituent groups. The effect of the diversity of P450 enzyme structures was assessed by using a library of artificial chimeric CYP1A enzymes of increasing shuffling complexity as described previously [9]. The purpose of this study was to assess to what extent a statistical approach relying on kinetics could be used to identify and predict global structural determinants of substrate-enzyme recognition. Interestingly, the X-ray structural information now available for CYP1A P450 proteins [10, 11] does not evidence any clear features explaining how major differences of activities among CYP1A enzymes could be related to local structural differences. Statistical approaches thus constitute an alternative but they require sufficiently large datasets of experiments thus relying on high-throughput methods. The steady-state rates were determined by monitoring metabolites produced from a given steroid substrate by LC/MS following different incubation times in 96-well microplate using yeast-expressed

recombinant mammalian CYP1A enzyme. Due to the large number of enzyme-substrate couples, each giving frequently several metabolites, only measurements at saturating substrate concentrations were performed.

The choice to study CYP1A enzymes instead of CYP2 or CYP3 enzymes, which are well known to oxidize steroids, is mainly due to the fact that CYP1A enzymes are less numerous than CYP2C ones and, thus, more easily amenable to interpretable combinatorial studies. Moreover, CYP1A enzymes do not exhibit cooperativity in their kinetics contrary to what is observed with CYP3A enzymes. In these respects, CYP1A enzymes are simpler and, therefore, better adapted to be used as models of steroid oxidation, as is the case in this work.

2. Materials and Methods

2.1. Substrates. The steroid series comprises testosterone, 17-methyltestosterone, progesterone, pregnenolone, estrone, cortisol, 19-norandrostenedione, dehydroepiandrosterone (DHEA), cortexolone, corticosterone, 17-hydroxy- and 21-hydroxyprogesterone, *cis*-androsterone (3α-hydroxy-5α-androstan-17-one), *trans*-androsterone (3β-hydroxy-5α-androstan-17-one), and two steroid analogs: nootkatone and mifepristone (RU486). Mifepristone was a gift from Roussel-Uclaf (Romainville, France). All other steroids were from Sigma. The different substrates were solubilized either in methanol (since ethanol is a known CYP1A inhibitor) or in dimethylformamide (DMF).

2.2. Plasmids and Yeast Strain. Vectors p1A1/V60 and p1A2/V60 for human wt CYP1A expression, pP1V8 for mouse wt 1A1 expression, and pLM4V8 for rabbit wt 1A2 expression were described before [12, 13]. The pYeDP60 vector contains both *URA3* and *ADE2* as selection markers, whereas pYeDP8 only bears *URA3*. The inserted coding sequence is placed under the transcriptional control of a *GAL10-CYC1* hybrid artificial promoter and *PGK* terminator.

Saccharomyces cerevisiae W(R) strain is a derivative of the W303-1B which, when cultivated on galactose, overexpresses yeast NADPH-P450 reductase. For expression of P450 enzymes, the W(R) yeast strain was chosen since yeast NADPH-P450 reductase overexpression optimizes the activities of any recombinant P450. After transformation and growth on selective medium, positive clones were selected for mitochondrial respiration on plates containing glycerol [14]. Several well-growing clones were used to inoculate 50 mL of SGI selective liquid media. This culture was grown overnight to reach stationary phase and was used to inoculate 250 mL of YPGE liquid culture. After 48 hours growing on constant shaking and at 28°C, 5 g of galactose was added to the culture for the overnight induction of the expression of both the cloned gene and the yeast P450 reductase.

2.3. Microsomal Fractions Preparation. Briefly, yeast cells were harvested by centrifugation, suspended, and washed in 50 mM Tris-HCl 1 mM EDTA 0.6 M Sorbitol (TES) buffer pH7.3. Cells were disrupted by manual shaking with 0.4 mm diameter glass beads. Cellular debris were removed by centrifugation (10 min at 10,000 rpm). The supernatant was transferred to another centrifuge tube, and NaCl and PEG4000 were added at final concentrations of 0.1 M and 10%, respectively, and kept on ice for 30 minutes. The precipitated microsomes were then pelleted by centrifugation 10 min at 10,000 rpm, washed, and resuspended in 50 mM Tris-HCl 1 mM EDTA 20% glycerol, pH7.4 buffer, and stored in −80°C [14].

2.4. Chimeric CYP1A Enzyme Variants. Four parental P450s were used in DNA shuffling experiments: human CYP1A1 and CYP1A2 mouse CYP1A1 and rabbit CYP1A2 coding sequences. Two shuffling methods have been applied to build a library of increasing complexity ranging from bi- or tripartite chimera up to highly mosaic structures (average 5-6 crossovers per sequence). The bi- or tripartite chimeras were obtained by *in vivo* gap-repair technology [15] between mouse CYP1A1 and rabbit CYP1A. The mosaic CYP1A variants were obtained using the mixed *in vivo, in vitro* CLERY recombination procedure between human CYP1A1 and CYP1A2 [16]. These two libraries of chimeric sequences coding for functional variant mammalian CYP1A enzymes have been previously described [9].

2.5. Enzyme Activities. Incubations with steroids (initiated by NADPH addition), reaction quenching with trifluoroacetic acid (2 : 250 by vol.), and acetonitrile extractions were automatically performed with a QIAgen 8000 robot using a homemade QIAsoft program. The high-throughput procedure consists in incubation of the different enzyme, substrate pairs in each well of a 96-well microplate. Microsomal fractions of recombinant yeast clones each expressing a particular CYP1A enzyme were assayed with the 16 steroid substrates in a round-bottom 96-well microplate (well volume = 0.30 mL) by incubation at 30°C for 0, 5, 10, 15, and 20 min. Incubation mixtures contained 0.3–0.6mg of microsomal yeast proteins and their concentration in recombinant P450 was ranging from about 5 up to 30 nM (as assessed by CO-reduced differential spectrum on some preparations). These variations were eliminated by the statistical treatment described further. A systematic control was included for each chemical tested and consisted of 20 min incubation with microsomal fractions prepared from W(R) cells transformed by void vector. After incubation completion, microplate content was transferred by the robot to a second microplate for acetonitrile treatment (1 : 1 by vol.) and, after centrifugation (5 min at 3,500 rpm), the 96 supernatants of this microplate were transferred to a third 96-well Porvair microplate which fits in the automatic injector compartment of an Alliance HT2795 HPLC Waters module. For all steroid substrates, the initial concentration in incubation mixtures was 100 μM, a value that generally corresponds to enzyme saturation while remaining below the solubility limits. Microsomal steroid hydroxylation and mifepristone N-demethylation activities were shown to be strictly NADPH-dependent.

2.6. Analytical LC/MS Methods. The acetonitrile supernatants were LC-separated at 40°C with an XTerraMS C_{18} 5 μM

4.6 × 100 mm column (Waters) and analyzed on a Micromass ZQ single quadrupole mass spectrometer (Waters). The solvent system consisted of H_2O + 0.01% formic acid (by vol.) in acetonitrile + 0.01% formic acid (by vol.). The gradient used (flow rate of 1.0 mL/min) for all steroid metabolites starts at 90 : 10 (water : acetonitrile), followed by linear increase from 85 : 15 to 0 : 100 over 8 min, return to initial conditions and hold for 2 min; 10 min of total run length. Parameters of the electrospray positive ionization were as follows: capillary voltage 3.4 kV, cone voltage 20.0 V, desolvation gas flow 550 L · h^{-1}, desolvation temperature 350°C, and source temperature 120°C. Continuous metabolite mass detection was using both full scan spectra by scanning mass range 200–500 amu and several SIR channels (single ion response) set at the precise m/z corresponding to the expected protonated metabolites (hydroxylated derivatives of all substrates plus the N-demethylated metabolite of mifepristone). With the masses of each substrate being known, it was possible to specifically detect the mass of predicted hydroxylated (+16 amu) and N-demethylated products (−14 amu). The detected m/z corresponds to [M+H]+ since positive electrospray mode is used. Metabolites were quantified by measuring peak areas on mass spectra chromatograms. The area of each MS detected product peak was plotted versus the time of incubation (0, 5, 10, 15, and 20 min) yielding the rate of reaction. MS response coefficients are linked to ionization efficiencies and can differ between metabolites. In the absence of suitable standards for all metabolites, absolute calibration cannot be performed but due to the similar ionization mechanisms for all ketosteroids, the observed signal was considered to correlate relative metabolite amounts. However, a suitable normalization procedure was applied for statistical analysis.

2.7. Statistics. A normalization procedure based on the variance in the dataset was used as previously described [9]. CYP1A activities for the different steroidal substrates were analyzed using principal component analysis (PCA) and multidimensional scaling (MDS). PCA visualizes systematic patterns or trends of variation in large data set. The different trends of variation hidden in the initial multidimensional space are evidenced since the new orthogonal axes of the projected space (the two first principal components) are derived from the directions of larger variability [17]. MDS is a nonlinear projection of the distances separating each object from the others in the original multidimensional space into a 2- or 3-dimensional diagram designated as the MDS configuration plot [18]. MDS enables one to easily visualize and evaluate distances between two objects considering at the same time the influence of all other objects. An indicator to evaluate the quality of a MDS analysis is calculation of a diagnostic index known as Kruskal's stress which varies from 0 (a perfect fitting) up to 1 (no fitting). It measures the closeness of the distance mapping in the 2D MDS plot compared to the "real" distances in the original space. The multivariate statistics and dendrogram construction were performed by using Addinsoft XLSTAT software. Datasets and correlation matrices used throughout this work are available upon email request from urban@insa-toulouse.fr.

2.8. Human CYP1A Structures. The atomic coordinates were taken from the Protein Data Bank, RCSB, Rutgers University (entry: 4i8v for human CYP1A1 at 2.6 Å resolution and 2hi4 for human CYP1A2 at 1.95 Å resolution) [10, 11]. Protein visualizations were performed by using PyMol (Delano Scientific LLC, San Carlos, CA, USA).

3. Results and Discussion

3.1. Experimental Design for Activity Matrix Acquisition. This work is a combinatorial approach of protein quantitative structure-activity relationships (QSAR) [19]. A large collection of chimeric enzymes (all related by sharing a high degree of sequence similarity due to the fact that the parental enzymes are homologous) was prepared by shuffling sequence elements between two parental wild-type enzymes [9].

The library of chimeric structures combining members of the CYP1A subfamily was built starting from four parental wild-type enzymes, two CYP1A1s and two CYP1A2s, originating from different mammals (human, mouse, and rabbit). Human CYP1A1 and CYP1A2 amino acid sequences present a 73% identity, placing them over the limits for effective recombination using annealing-based methods. In our case, these two sequences were shuffled using the CLERY method [16]. This shuffling produced mosaic sequences with an average of 5 crossovers per sequence. The expressed functional CYP1A mosaic variants are collectively designated as ChiMo enzymes and numerated individually from ChiMo1 up to ChiMo11. The second library of chimera was constructed by shuffling sequences between mouse CYP1A1 and rabbit CYP1A2. Their 65% amino acid identity places them below the limits for effective PCR-based recombination techniques. The chimeras were obtained by using an *in vivo* technique based on gap repair in yeast [15]. This shuffling produced bi- or tripartite chimeric sequence. The expressed functional CYP1A chimeric variants are collectively designated as Chim enzymes and numerated individually from Chim3 up to Chim14.

We assayed the wild-type and chimeric P450 enzymes ($n = 31$) with a collection of 16 steroids and 2 analogs (Figure 1). Mifepristone is an N-dimethylated steroid analog and both hydroxylation reactions and N-demethylation reactions can be observed with CYP1A enzymes. All other substrates were monitored for CYP1A-catalyzed hydroxylase activity. With this set of substrates, 82 different metabolites were detected by mass spectrometry after separation of the incubation mixture by reverse-phase LC. Each of these metabolites specifies an activity; the profile of all detected metabolites produced from a single steroid substrate allows accessing a regiospecificity index (to be defined latter) for the action of a particular CYP1A enzyme. For each pair (enzyme, activity), the initial rates of metabolite formations were measured using the same substrate concentration (100 μM) chosen to be generally saturating and within the range of solubility of the whole set of chemicals. Experimentally the 31 different enzymes and 18 substratesrepresent 558

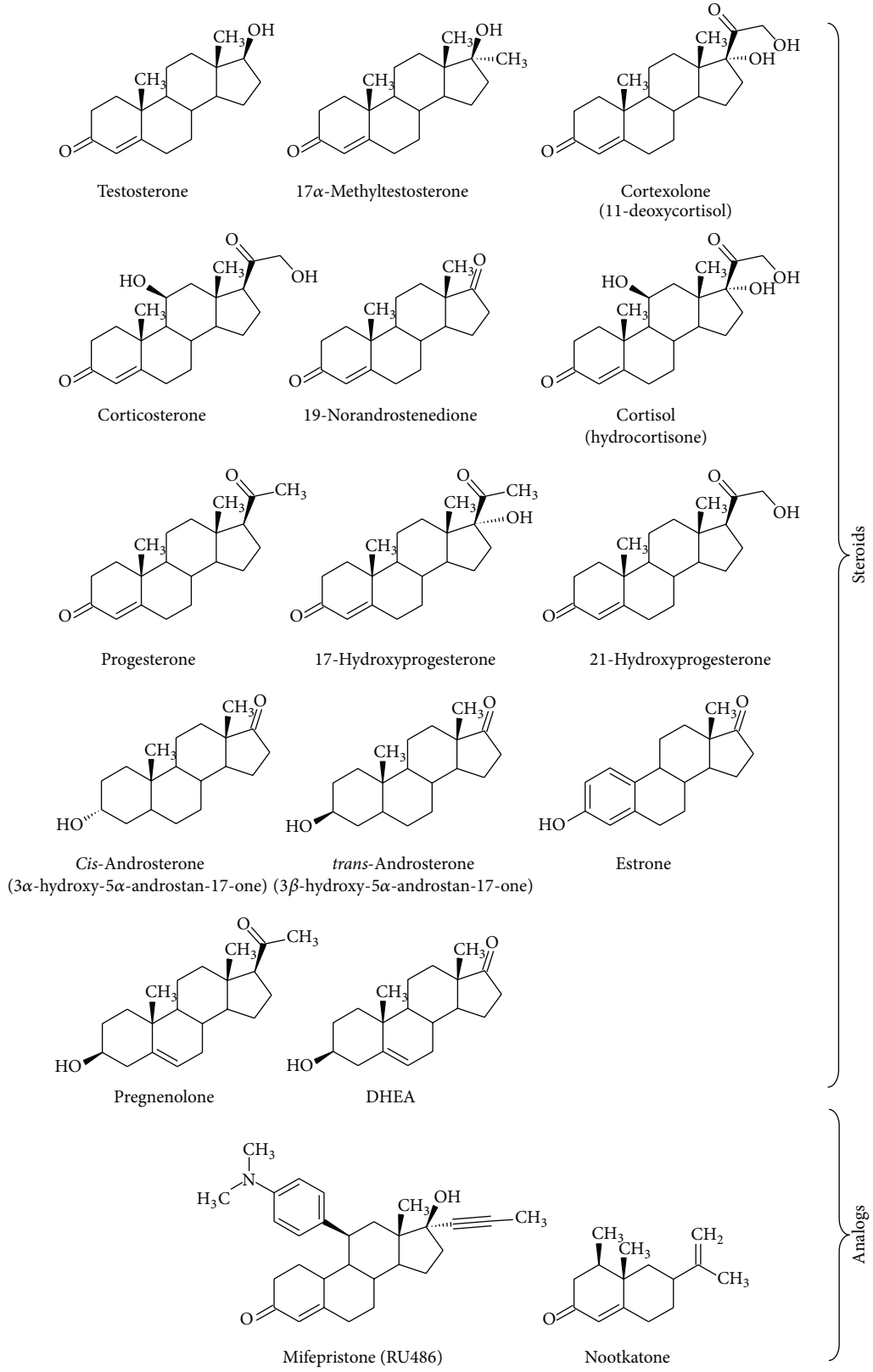


Figure 1: Chemical structures of the steroids and steroid analogues used in this work as substrates to assay CYP1A enzymes.

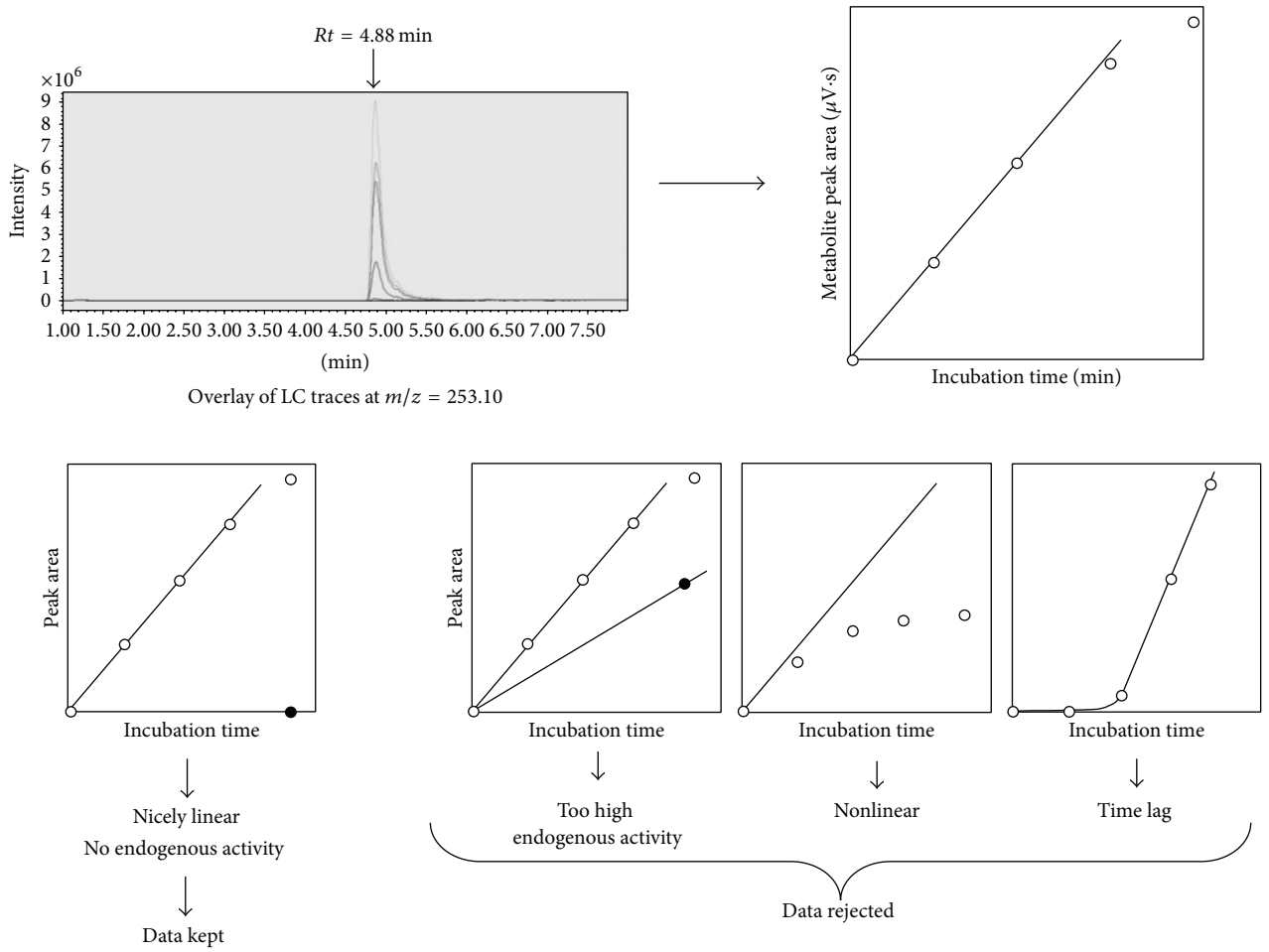


Figure 2: Flowchart for HT activity analysis. Upper left panel, LC/MS screenshot of the overlay of six chromatograms observed for CYP1A1-catalyzed hydroxylation of a substrate of $m/z = 237$ at different incubation times (hydroxylated substrate [M + H] + = 253). Upper right panel, area of the metabolite peak at $Rt = 4.88$ min versus incubation time. The four lower panels present different types of observed kinetics and rejection cases. Left panel, the case is kept for further analysis (metabolite production is linear over time).

(enzyme, steroid) pairs, leading to 3348 individual incubations considering that a 5-point time series (0, 5, 10, 15, and 20 min) and a negative control (void expression vector and 20 min incubation time) were performed for each condition (see Table S1 in Supplementary Material available online at http://dx.doi.org/10.1155/2014/764102). Assays were duplicated using at least two independent microsomal preparations and hidden replicates were included. The initial rates determination was achieved in a reasonable amount of time with the help of high-throughput technology. Only activity-enzyme pairs for which the time course of metabolite accumulation appeared to be linear with a negligible background endogenous activity were further considered for analysis (Figure 2).

3.2. Data Normalization for Analysis. Since kinetic analyses were not carried out with purified enzymes and because P450 contents in microsomal fractions were not systematically determined due to limited sensitivity of spectral quantification with certain microsomal fractions, actual catalyst contents from a yeast microsomal preparation to the other were not systematically assessed. Two parallel normalization treatments were thus carried out on the raw activity data.

The first treatment of dataset consisted in unit scaling of activities; it is aimed to easily visualize the full dataset by normalizing the activities of all metabolites in a way that each activity value ranges from 0 to 1. This makes a direct comparison of the dataset possible even if the raw unprocessed values widely differ, as is the case here. To do so, for each metabolite, the highest signal observed is arbitrarily set at 1. All other activities determined for this metabolite with all other enzymes are expressed as a ratio of their values to the highest activity observed. This makes all activity values for each metabolite to range from 0 to 1 after this data processing. Such a processing was already used for activity data visualization purposes by others [20].

The second treatment of dataset consisted in a normalization based on the variance in the dataset as described previously [9]. This variance normalization [21] eliminates the arbitrary factor linked to differences in expression levels of enzyme variants. A possible drawback could be that global substrate selectivity features of the whole enzyme collec-

tion are not corrected and could introduce some arbitrary weighting of data. Two dimensionality reduction statistical approaches: principal component analysis (PCA) and multidimensional scaling (MDS), were performed on the variance-normalized dataset to obtain projections of the data in two-dimensional diagrams. The two statistical approaches, PCA and MDS, are complementary and somewhat slightly affected by unavoidable data bias. PCA is a linear approach fairly sensitive to weighting and by the way the "distances" between activities are calculated (the metrics). MDS is in contrast a nonlinear projection, frequently less sensitive to these factors which, in contrast to PCA, do not necessarily yield a unique solution.

3.3. Structural Features Important for a Steroid to Be a CYP1A Substrate. Correlations existing between substrate structural elements and the activities were first investigated using a subset of the dataset for the sake of clarity. This subset includes only the wild-type CYP1A enzymes and the activities corresponding to the principal metabolite produced for each steroidal substrate. The analysis was limited to wild-type enzymes since the process of chimera formation by sequence shuffling may alter the typical profile of the protein structural elements that fit the substrate compared to the original profile in the wild-type enzyme. Similarly, the analysis was limited to the main activities for each steroidal substrate because the main product being produced at a high rate implies an optimal fitting (in the general acceptance of this term) of the substrate molecule within the catalytic cavity of the enzyme.

The correlation matrix (i.e., dissimilarity matrix) was deduced from this subset and used to extract a MDS configuration plot in which the objects shown are steroid substrates scattered throughout the plot depending on their behavior towards wild-type CYP1A enzymes taken globally (Figure 3, left panel). Two steroid substrates will be found close together in this MDS configuration plot if and only if their behaviors toward the considered set of wild-type CYP1A enzymes are similar; the closer, the more similar. On the contrary, steroid substrates which are dissimilar for the CYP1A enzymes are found to be plotted far from each other.

Activities toward steroid substrates fall within three distinct groups, one of them involving two subgroups. The mifepristone N-demethylase activity appears at a well-separated position on the graph. A surprising observation is that the same substrate belongs to a distantly related group when steroid hydroxylation is considered. A likely explanation of this fact would be that the N-demethylation reaction takes place at a side of the mifepristone molecule opposite to the side at which hydroxylation reactions occur. This strongly suggests that hydroxylation regioselectivity plays an important role in the graph structure, hence in CYP1A metabolism of steroids.

A second group contains four steroids, two androsterones (*cis* and *trans*), DHEA, and 19-norandrostenedione. The fact that both *cis*- and *trans*-androsterones (resp., 3α-hydroxy-5α-androstan-17-one and 3β-hydroxy-5α-androstan-17-one) both fall in the same group suggests that orientation of the hydroxyl-group at position 3 of the steroid molecule is not determining for steroid recognition by CYP1A enzymes. All these steroids are characterized by no or small side groups on their molecular scaffold, particularly at C17 position which is always a keto group in this group. However, this feature is not fully characteristic of the group and a contrasting example can be found in the third group.

This third group contains most of the steroid substrates, including the mifepristone hydroxylase activity. Most of these steroids present two bulky groups at the C-17 position. Estrone, which is also a simple steroid molecule, falls in this group but its *A* ring is aromatic contrary to all other steroidal substrates tested. Due to *A* ring aromaticity, estrone is more testosterone-like and this could explain its classification in this third group. The terpene nootkatone, a steroid analog, is also found in this third group and not as an outlier, as is observed for mifepristone.

A dendrogram built by the Ward method calculated from the correlation matrix used for MDS analysis is shown in Figure 3, right panel. The chemical structure of corresponding steroid substrate is indicated in front of each branch of the resulting tree. This helps visualizing the main structural determinants on steroidal substrates that are differentiated by wild-type CYP1A enzymes and, therefore, are critical for a steroid molecule to be a substrate well recognized by CYP1A enzymes.

A closer examination of the large upper group on the dendrogram illustrates that this group could be split into several subgroups since the main branch of the dendrogram gives rise to three individual branches of low dissimilarity. All steroidal substrates found in the upper branch have in common to present both a bulky group found at C17 position and a keto group at C3 with a 4-androstene skeleton. The second subgroup contains two molecules, one being dehydroepiandrosterone (DHEA), with a modified unsaturation, and a 3-hydroxy function compared to the other 3-keto-4-androstene molecules of this group. This suggests that some modulating effects could be played by the position of the unsaturation. The third subgroup is more diverse since it includes estrone which has a small keto group at position C17 and an aromatic *A* ring. In this case, a modulating effect seems to be played by *A* ring aromaticity of the steroid molecule. It can be concluded that both the hindrance at the C17 position and the presence of unsaturated C-C bond within the *A* ring of steroid substrate play significant role for metabolism by CYP1A enzymes. The different features deduced from this analysis are summarized in Figure 4.

3.4. CYP1A Structural Elements Influencing Steroid Metabolism. When statistical analysis is no longer applied to the dataset columns (i.e., activities) but to rows (i.e., enzymes), it is possible to classify by MDS analysis the different wild-type and chimeric enzyme structures for their specificity toward the steroidal substrates assayed (Figure 5(a)). Human CYP1A2 enzyme exhibits a fairly narrow substrate specificity for steroids compared to the other tested wild-type CYP1A, even compared to its rabbit CYP1A2 ortholog. Consequently, the duplicated assays of this enzyme appear relatively isolated from its rabbit counterpart on the left border of the plot. Moreover, rabbit CYP1A2 presents a J helix which is of the

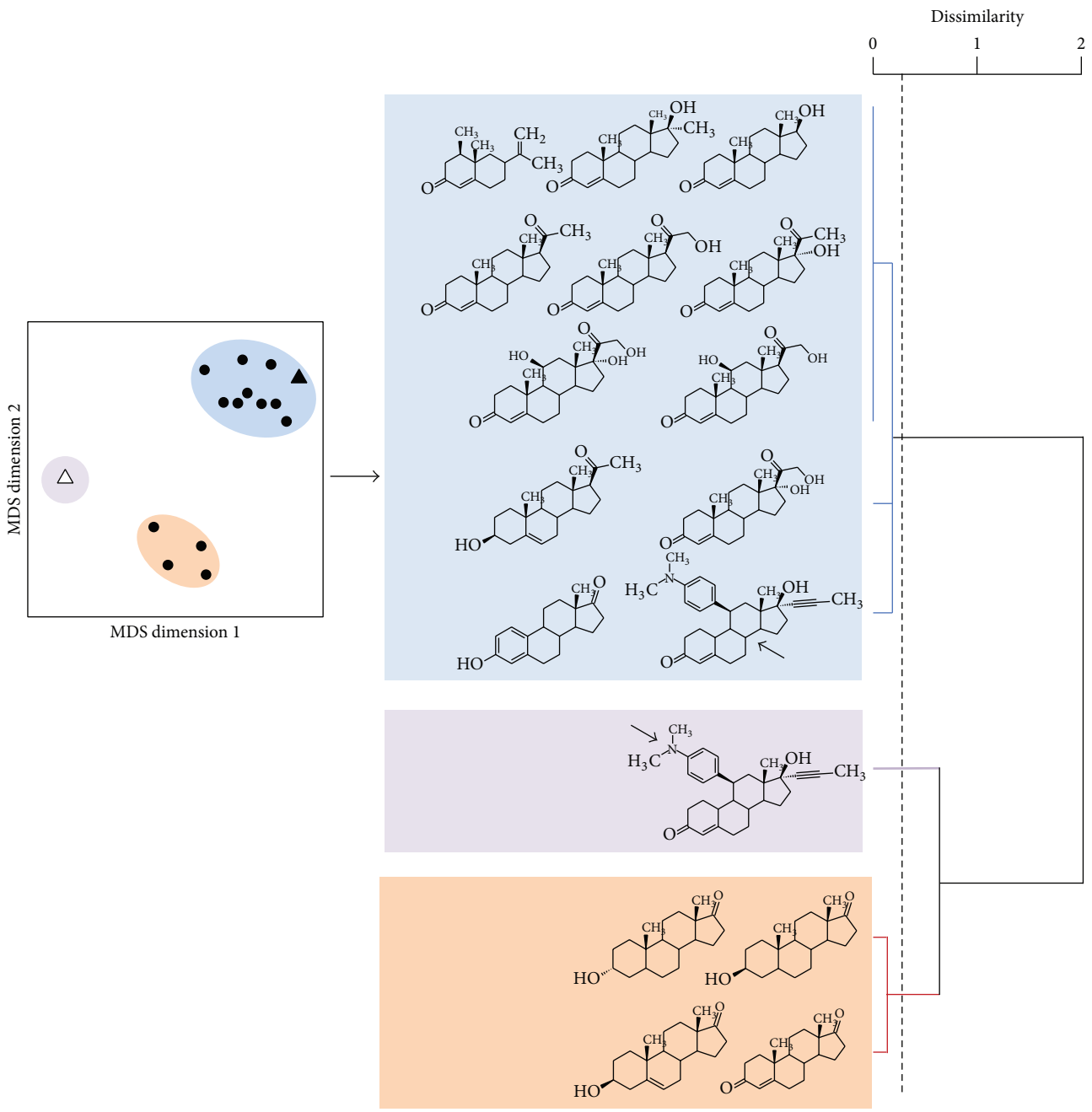


Figure 3: Visualization of how wild-type CYP1A enzymes differentiate steroidal substrates. Left panel, MDS configuration plot based on the ratio model and characterized by a stress index of 0.016. Each point represents the main activity observed for each one of the steroidal substrates tested in this work except for mifepristone which is represented in this plot by two points (triangles). The solid triangle represents the main mifepristone (RU-486) hydroxylase activity with all CYP1A enzymes; the open triangle represents mifepristone N-demethylase activity. Right panel, dendrogram deduced from the correlation matrix used to build the MDS configuration plot. The dendrogram was built by the Ward method based on the dissimilarity matrix.

1A1-type in its sequence, contrary to what is seen for human CYP1A2 (see Table 1), and this helix has been proposed to be involved in the interaction with NADPH-P450 reductase [22, 23]. Notably, a systematic inversion of two charged residues is observed between CYP1A1s and CYP1A2s in this helix; namely, Arg 338 and Glu345 in human CYP1A1 correspond to Glu338 and Lys345 in human CYP1A2. In mouse CYP1A1, these two residues are conserved with human 1A1, namely, Arg342 and Glu349. But, when looking at rabbit CYP1A2 sequence, the situation observed is typical of CYP1A1 enzymes rather than CYP1A2 ones, with the combination of Arg338 and Glu345. This suggests that rabbit CYP1A2 is rather of the 1A1-type than of the 1A2-type and it could explain why rabbit CYP1A2 duplicate is found more closely aggregated with the two CYP1A1 clusters than with human CYP1A2 duplicate (Figure 5(a)).

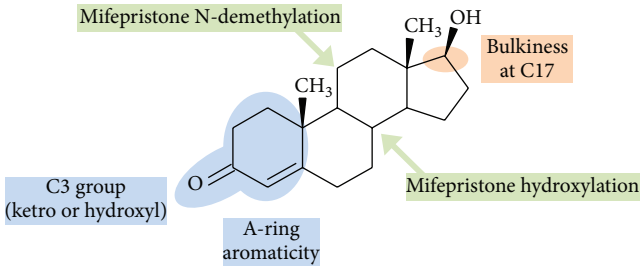


Figure 4: Features that were found important for efficient recognition of a steroidal substrate by CYP1A enzymes.

Table 1: Alignment of amino acid sequences of the J-helix of the four parental wild-type CYP1A enzymes studied in this work.

Mm CYP1A1	341-PRVQRKIQEEL-351
Hs CYP1A1	337-..........-347
Hs CYP1A2	337-.EI.....K..-347
Oc CYP1A2	337-.RR.....E..-347
	▲ ▲

HS: *Homo sapiens*, Mm: *Mus musculus*, and Oc: *Oryctolagus cuniculus*. The solid triangles indicate the charged residue discussed in the text; a dot represents an amino acid residue identical to that of Mm CYP1A1 sequence.

Interestingly, the behaviors of chimeric enzymes do not appear as a linear combination of that of parental forms, a large part being outside of the boundary of polygon drawn from parental enzyme positions on the graph (Figure 5, salmon-colored area). Therefore, some novel steroid specificity is revealed within the collection of chimeric enzymes. Moreover, the chimeras escaping boundaries of wild-type specificities belong both to the Chim and the ChiMo series, indicating the absence of any clear correlation between the complexity of shuffling (number of crossovers) and the development of new activities. Independent replicas appear always tightly grouped on the graph, thus indicating that differences of behaviors between chimeras are not significantly influenced by experimental "noise." A hidden triplicate of the same chimera (Chim4, Chim6, and Chim12) appears also as a tight aggregate on the MDS plot (red ellipse named "1").

Several groups of enzymes can be drawn from this MDS plot. For the sake of clarity, this comparison was limited to wild-type enzymes and chimeric variants of the Chim series, since these variants exhibit a simpler shuffling in sequences than that of ChiMo series. On a MDS plot, if two objects (i.e., the enzymes in this work) are found close to each other, it means that their global behaviors toward the properties analyzed (i.e., the substrates in this work) are similar.

A first group, shown by a red ellipse denoted by "1" on the plot, contains wild-type mouse CYP1A1 and is located quite closely to the group of wild-type rabbit CYP1A2. One so-called "chimera," Chim14, was found to be a hidden replicate of rabbit CYP1A2 and is thus aggregated to the rabbit CYP1A2 cluster. The two parental wild-type enzymes are found closely located. This indicates that their behaviors towards the steroidal substrates assayed in this work are quite similar.

Another, clearly outlying, group of chimeras is found in a quite remote area in the upper left part of the graph and is drawn in a red ellipse denoted by "2" in Figure 5(a). This group corresponds to a hidden triplicate of the same chimeric sequence and belongs to a larger group of more complex ChiMo chimeric structures whose representative points span the upper left quadrant. These sequences therefore encode CYP1A chimeric variants whose specificity profile for steroid substrates is very different from those of the parental wild-type enzymes.

The third group, found to be more slightly resolved from the group of wild-type enzymes, stands in the lower-right quadrant of the plot including three chimeras (Chim5, Chim8, and Chim13) of different sequences which is shown by the red ellipse denoted by "3." Steroid specificity of group "3" enzymes moderately differs from the combination of the ones of parental enzymes (group "2").

In a preliminary step to further analyze the relationships between global specificity towards steroids and sequence shuffling, Figure 5(b) illustrates a low resolution map of amino acid sequences of discussed chimera as compared to parental sequences. A simple segmentation identifies two sequence segments, designated as "A" and "B" (purple and salmon boxes on Figure 5(b)), whose sequence type varies accordingly to the global steroid specificity of the corresponding CYP1A enzyme. The sequence segment "A" encompasses amino acids ranging from residue 142 (mouse CYP1A1 numbering) to residue 247. The segment "B" encompasses amino acids from residue 377 to the C-terminus. It is apparent that the combination of the parental type of these two segments is determining of the functional groups, notably when segment "A" is of the 1A2-type and segment "B" of the 1A1-type. Moreover, the observation that chimeras in group 1 show a segment "B" either of the 1A1- or of the 1A2-type suggests that this segment is not critical for steroid metabolism but rather has modulating effect.

Altogether, sequence segments "A" and "B" represent almost 35% of the sequence, a rather large domain which needs further refinement to define the exact contributors. Segment "A" encompasses only one of the several SRSs, defined by Gotoh as P450 Substrate Recognition Sites [24], namely SRS2 which is highly divergent between 1A1 and 1A2 subtypes. SRS2 has also been highlighted frequently in controlling CYP1A enzymes functioning elsewhere [25–28]. Segment "B" encompasses two SRSs, namely, SRS5 and SRS6. Contrasting with segment "A," these two SRSs are almost identical between mouse CYP1A1 and rabbit CYP1A2. Within SRS6, a threonine in mouse CYP1A1 is replaced by an isoleucine in rabbit CYP1A2 close to the C-terminus. This suggests that combinations of amino acids residues critically controlling steroid specificity in CYP1A enzymes should be more likely localized in segment "A" rather than in segment "B."

When placed on the crystal 3D structure of human CYP1A1, the "A" segment appears to be mostly located at the periphery of the protein molecule (Figure 6), whereas the helix I and its surrounding remain not affected by the shuffling taking place in Chim enzymes. It is consistent with the literature that helix I is crucial in P450 catalysis [11, 29]

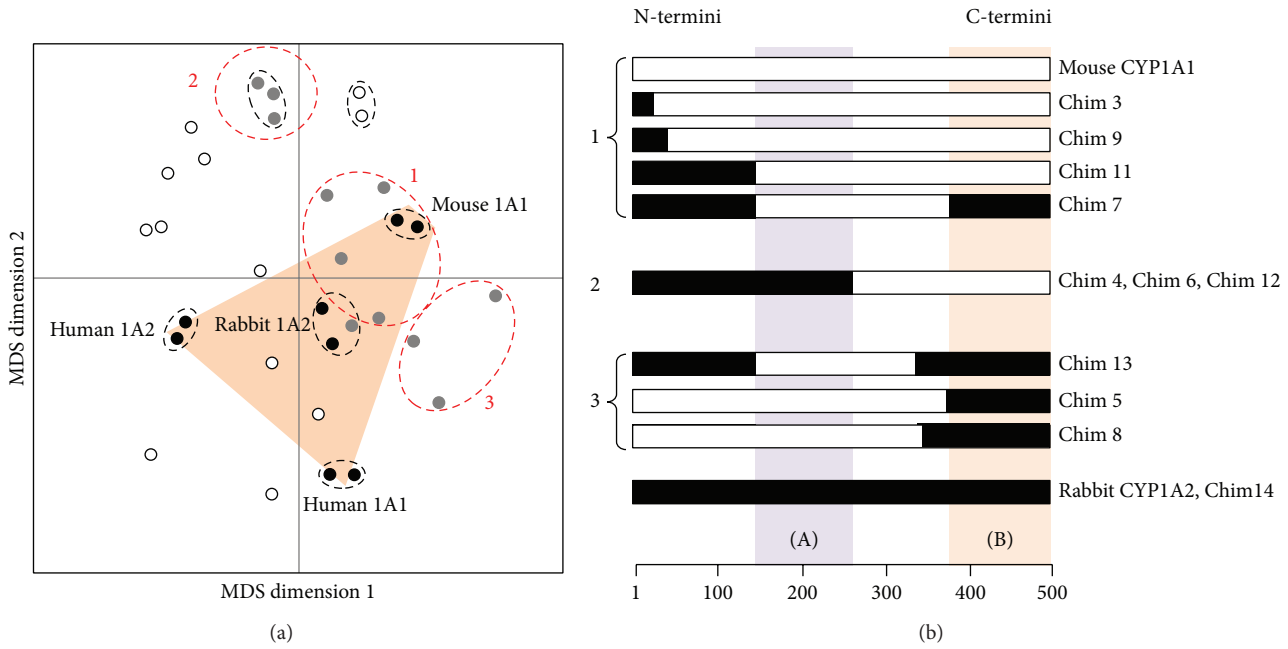


Figure 5: Panel (a), MDS configuration plot of the 31 CYP1A enzymes tested against all steroid activities. Some chimeric CYP1A enzymes are hidden replicates; that is, the cluster named "1" contains three chimeras of the Chim series which are triplicate of the same chimeric sequence. Solid circles correspond to wild-type enzymes, grey-filled circles to bi- or tripartite chimeras (Chim series), and open circles to mosaic chimeras (ChiMo series). The red ellipses correspond to cluster of enzymes which are discussed text based on their highly similar behaviour towards the steroids tested in this work. Panel (b), phylofunctional analysis of mouse-rabbit CYP1A chimeric enzymes assayed for hydroxylase activity with sixteen steroidal substrates. The open and solid bars represent, respectively, mouse CYP1A1 and rabbit CYP1A2 amino acid sequence segments. The purple and salmon boxes highlight the two elements of primary structure which are found to be correlated with the groups on the MDS plot.

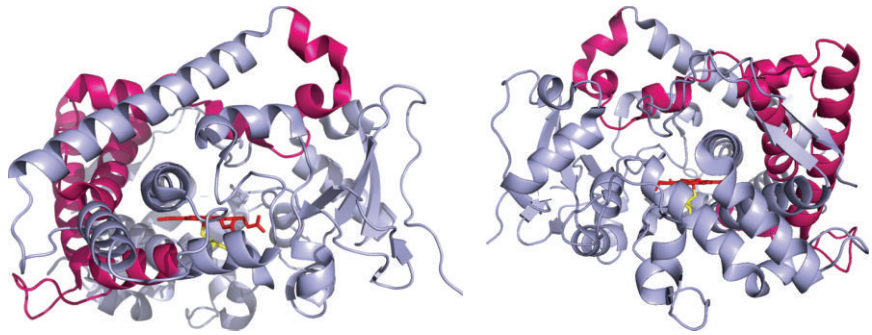


Figure 6: Human CYP1A1 shown in grey-blue ribbon structure (pdb entry 4i8v) highlighting sequence segment A (colored hot pink) and the heme (colored red). The two panels represent the structure rotated horizontally by 180° horizontally, with the I-helix viewed through its axis.

and that the distal cavity is not expected to be affected by sequence covering segment "A." It is also consistent that sequence change in segment "B" could affect the lining of the proximal side of the heme and, thus, the heme-thiolate bond because it encompasses the cysteine that ligates the heme iron.

However, the way by which these structural factors can directly impact enzyme activity remains unclear based on structural data. Obviously, activity maps associated with all chimeras likely contain much more hidden information and the way to decipher this information in conjunction with structural data remains to be established on a more thorough theoretical ground. A potential approach would be solving experimentally the 3D structures of chimeric enzymes to understand how a novel activity pattern can be generated. This approach could be hampered by the fact that the comparison of the crystal structures of human CYP1A1 and human CYP1A2 parental enzymes was unfortunately poorly explanatory of functional differences since the two enzymes are highly similar in structure.

One interesting working hypothesis, supported by some preliminary data (not shown), would be that the apparent

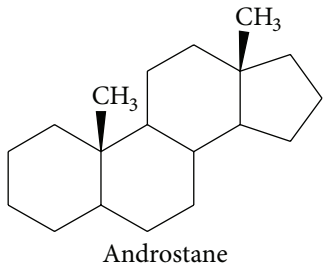


Figure 7

difference of specificity could more specifically result from differential control mechanisms of the substrate dependent initiation of oxygen activation cycle between CYP1A1s and CYP1A2s rather than from detailed features of substrate binding in the active site. In this hypothesis, the catalytic cycle of CYP1A1 would appear constitutively initiated even in the absence of substrate/ligand bound into the active site, whereas CYP1A2 would exhibit a more tight substrate dependent control of the initiation of the catalytic cycle. Such idea is suggested by preliminary data indicating that activity of chimeric structures supported by redox partners (NADPH-P450 reductase and cytochrome b_5) or supported by hydroperoxide compounds, such as cumene hydroperoxide, can significantly differ for the same substrate depending on the presence of specific CYP1A2 sequence segments.

4. Conclusions

The approach initiated in this work takes profit of both a combinatorial mutagenesis to yield functional chimeric enzymes of increasing sequence shuffling complexity and of a combinatorial library of structurally related substrates having lateral groups of increasing complexity decorating a common simple chemical skeleton.

QSAR correlations may bring information on both the structural elements of the enzyme and the structural elements of the substrate that together govern recognition and regioselectivity in a crosstalk that remains to be understood. Several studies have illustrated the fact that CYP3A4 is one of the major players involved in the oxidation of steroid hormones in human liver microsomes [30], together with CYP2Cs [31]. Our work illustrated that members of the CYP1A family might play a complex role in hormonal regulation taking into account that CYP1A2 expression is mostly constitutive while that of CYP1A1 is highly inducible by a large range of xenobiotics, some of them being environmental pollutants and others being drugs [32].

Two important steroid structural features were characterized in this study, the substituent at the C17 position and the possibility for steroid molecules to bind in two opposite orientations within the catalytic site of CYP1A enzymes as evidenced by mifepristone activities. The nature and bulkiness of side groups at the C17 position clearly affect the catalytic efficiency of CYP1A enzymes.

Finally, the structural elements of the CYP1A protein that govern the parental phenotypes appear multiples, some combinations leading to the appearance of a novel substrate regiospecificity for steroids and related molecules. These determinants are mostly located at or near the protein surface and not close to the buried catalytic cavity suggesting some indirect mechanism controlling CYP1A activity. This result confirms previous works on P450 enzymes [33, 34] and opens new pathways for looking at sequence-activity relationships more deeply.

Conflict of Interests

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

References

[1] M. J. Coon, "Cytochrome P450: nature's most versatile biological catalyst," *Annual Review of Pharmacology and Toxicology*, vol. 45, pp. 1–25, 2005.

[2] D. C. Lamb, M. R. Waterman, S. L. Kelly, and F. P. Guengerich, "Cytochromes P450 and drug discovery," *Current Opinion in Biotechnology*, vol. 18, no. 6, pp. 504–512, 2007.

[3] F. P. Guengerich, "Cytochrome P450 and chemical toxicology," *Chemical Research in Toxicology*, vol. 21, no. 1, pp. 70–83, 2008.

[4] J. McGraw and D. Waller, "Cytochrome P450 variations in different ethnic populations," *Expert Opinion on Drug Metabolism and Toxicology*, vol. 8, no. 3, pp. 371–382, 2012.

[5] D. J. Waxman, "Regulation of liver-specific steroid metabolizing cytochromes P450: cholesterol 7α-hydroxylase, bile acid 6β-hydroxylase, and growth hormone-responsive steroid hormone hydroxylases," *Journal of Steroid Biochemistry and Molecular Biology*, vol. 43, no. 8, pp. 1055–1072, 1992.

[6] D. J. Waxman, C. Attisano, F. P. Guengerich, and D. P. Lapenson, "Human liver microsomal steroid metabolism: identification of the major microsomal steroid hormone 6β-hydroxylase cytochrome P-450 enzyme," *Archives of Biochemistry and Biophysics*, vol. 263, no. 2, pp. 424–436, 1988.

[7] F. P. Guengerich, "Oxidation of 17β-ethynylestradiol by human liver cytochrome P-450," *Molecular Pharmacology*, vol. 33, no. 5, pp. 500–508, 1988.

[8] H. Yamazaki and T. Shimada, "Progesterone and testosterone hydroxylation by cytochromes P450 2C19, 2C9, and 3A4 in human liver microsomes," *Archives of Biochemistry and Biophysics*, vol. 346, no. 1, pp. 161–169, 1997.

[9] P. Urban, G. Truan, and D. Pompon, "High-througput enzymology and combinatorial mutagenesis for mining cytochrome P450 functions," *Expert Opinion on Drug Metabolism and Toxicology*, vol. 4, no. 6, pp. 733–747, 2008.

[10] S. Sansen, J. K. Yano, R. L. Reynald et al., "Adaptations for the oxidation of polycyclic aromatic hydrocarbons exhibited by the structure of human P450 1A2," *The Journal of Biological Chemistry*, vol. 282, no. 19, pp. 14348–14355, 2007.

[11] A. A. Walsh, G. D. Szklarz, and E. E. Scott, "Human cytochrome P450 1A1 structure and utility in understanding drug and xenobiotic metabolism," *The Journal of Biological Chemistry*, vol. 288, no. 18, pp. 12932–12943, 2013.

[12] J. Gautier, S. Lecoeur, J. Cosme et al., "Contribution of human cytochrome P450 to benzo [a]pyrene and benzo [a]pyrene-7,8-dihydrodiol metabolism, as predicted from heterologous expression in yeast," *Pharmacogenetics*, vol. 6, no. 6, pp. 489–499, 1996.

[13] D. Pompon, "cDNA cloning and functional expression in yeast *Saccharomyces cerevisiae* of β-naphthoflavone-induced rabbit liver P-450 LM4 and LM6," *European Journal of Biochemistry*, vol. 177, no. 2, pp. 285–293, 1988.

[14] D. Pompon, B. Louerat, A. Bronine, and P. Urban, "Yeast expression of animal and plant P450s in optimized redox environments," *Methods in Enzymology*, vol. 272, pp. 51–64, 1996.

[15] D. Pompon and A. Nicolas, "Protein engineering by cDNA recombination in yeasts: shuffling of mammalian cytochrome P-450 functions," *Gene*, vol. 83, no. 1, pp. 15–24, 1989.

[16] V. Abécassis, D. Pompon, and G. Truan, "High efficiency family shuffling based on multi-step PCR and in vivo DNA recombination in yeast: statistical and functional analysis of a combinatorial library between human cytochrome P450 1A1 and 1A2," *Nucleic Acids Research*, vol. 28, no. 20, p. E88, 2000.

[17] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, vol. 24, no. 6, pp. 417–441, 1933.

[18] W. S. Torgerson, *Theory and Methods of Scaling*, Academic Press, New York, NY, USA, 1958.

[19] D. E. Almonacid and P. C. Babbitt, "Toward mechanistic classification of enzyme functions," *Current Opinion in Chemical Biology*, vol. 15, no. 3, pp. 435–442, 2011.

[20] M. Landwehr, M. Carbone, C. R. Otey, Y. Li, and F. H. Arnold, "Diversification of catalytic function in a synthetic family of chimeric cytochrome P450s," *Chemistry and Biology*, vol. 14, no. 3, pp. 269–278, 2007.

[21] R. A. Fisher, "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.

[22] V. Cojocaru, P. J. Winn, and R. C. Wade, "The ins and outs of cytochrome P450s," *Biochimica et Biophysica Acta*, vol. 1770, no. 3, pp. 390–401, 2007.

[23] L. Nikfarjam, S. Izumi, T. Yamazaki, and S. Kominami, "The interaction of cytochrome P450 17α with NADPH-cytochrome P450 reductase, investigated using chemical modification and MALDI-TOF mass spectrometry," *Biochimica et Biophysica Acta*, vol. 1764, no. 6, pp. 1126–1131, 2006.

[24] O. Gotoh, "Substrate recognition sites in cytochrome P450 family 2 (CYP2) proteins inferred from comparative analyses of amino acid and coding nucleotide sequences," *The Journal of Biological Chemistry*, vol. 267, no. 1, pp. 83–90, 1992.

[25] A. Parikh, P. D. Josephy, and F. P. Guengerich, "Selection and characterization of human cytochrome P450 1A2 mutants with altered catalytic properties," *Biochemistry*, vol. 38, no. 17, pp. 5283–5289, 1999.

[26] C.-H. Yun, G. P. Miller, and F. P. Guengerich, "Rate-determining steps in phenacetin oxidations by human cytochrome P450 1A2 and selected mutants," *Biochemistry*, vol. 39, no. 37, pp. 11319–11329, 2000.

[27] D. Kim and F. P. Guengerich, "Selection of human cytochrome P450 1A2 mutants with enhanced catalytic activity for heterocyclic amine N-hydroxylation," *Biochemistry*, vol. 43, no. 4, pp. 981–988, 2004.

[28] J. Liu, S. S. Ericksen, M. Sivaneri, D. Besspiata, C. W. Fisher, and G. D. Szklarz, "The effect of reciprocal active site mutations in human cytochromes P450 1A1 and 1A2 on alkoxyresorufin metabolism," *Archives of Biochemistry and Biophysics*, vol. 424, no. 1, pp. 33–43, 2004.

[29] I. F. Sevrioukova and T. L. Poulos, "Dissecting cytochrome P450 3A4-ligand interactions using ritonavir analogues," *Biochemistry*, vol. 52, no. 26, pp. 4474–4481, 2013.

[30] H. Yamazaki, M. Nakano, Y. Imai, Y. Ueng, F. P. Guengerich, and T. Shimada, "Roles of cytochrome b_5 in the oxidation of testosterone and nifedipine by recombinant cytochrome P450 3A4 and by human liver microsomes," *Archives of Biochemistry and Biophysics*, vol. 325, no. 2, pp. 174–182, 1996.

[31] H. Yamazaki and T. Shimada, "Progesterone and testosterone hydroxylation by cytochromes P450 2C19, 2C9, and 34 in human liver microsomes," *Archives of Biochemistry and Biophysics*, vol. 346, no. 1, pp. 161–169, 1997.

[32] Q. Ma and A. Y. H. Lu, "CYP1A induction and human risk assessment: an evolving tale of in vitro and in vivo studies," *Drug Metabolism and Disposition*, vol. 35, no. 7, pp. 1009–1016, 2007.

[33] M. Lewinska, U. Zelenko, F. Merzel, S. Golic Grdadolnik, J. C. Murray, and D. Rozman, "Polymorphisms of CYP51A1 from cholesterol synthesis: associations with bird weight and maternal lipid levels and impact on CYP51 protein structure," *PLoS ONE*, vol. 8, no. 12, Article ID e82554, 2013.

[34] L. N. Ma, Z. Z. Du, P. Lian, and D. Q. Wei, "A theoretical study on the mechanism of a superficial mutation inhibiting the enzyme activity of CYP1A2," *Interdisciplinary Sciences: Computational Life Sciences*, vol. 6, no. 1, pp. 25–31, 2013.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 906168, 10 pages
http://dx.doi.org/10.1155/2014/906168

Research Article

Systematic Analysis of the Association between Gut Flora and Obesity through High-Throughput Sequencing and Bioinformatics Approaches

Chih-Min Chiu,[1] Wei-Chih Huang,[1] Shun-Long Weng,[1,2,3,4] Han-Chi Tseng,[5] Chao Liang,[6] Wei-Chi Wang,[6] Ting Yang,[1] Tzu-Ling Yang,[1] Chen-Tsung Weng,[6] Tzu-Hao Chang,[7] and Hsien-Da Huang[1,8,9,10]

[1] *Institute of Bioinformatics and Systems Biology, National Chiao Tung University, Hsinchu 300, Taiwan*
[2] *Department of Obstetrics and Gynecology, Hsinchu Mackay Memorial Hospital, Hsinchu 300, Taiwan*
[3] *Mackay Medicine, Nursing and Management College, Taipei 104, Taiwan*
[4] *Department of Medicine, Mackay Medical College, New Taipei City 251, Taiwan*
[5] *Tseng Han-Chi General Hospital, Nantou 542, Taiwan*
[6] *Health GeneTech Corporation, Taoyuan 330, Taiwan*
[7] *Graduate Institute of Biomedical Informatics, Taipei Medical University, Taipei 110, Taiwan*
[8] *Department of Biological Science and Technology, National Chiao Tung University, Hsinchu 300, Taiwan*
[9] *Center for Bioinformatics Research, National Chiao Tung University, Hsinchu 300, Taiwan*
[10] *Department of Biomedical Science and Environmental Biology, Kaohsiung Medical University, Kaohsiung 807, Taiwan*

Correspondence should be addressed to Tzu-Hao Chang; kevinchang@tmu.edu.tw and Hsien-Da Huang; bryan@mail.nctu.edu.tw

Received 11 April 2014; Accepted 27 June 2014; Published 14 August 2014

Academic Editor: Li-Ching Wu

Eighty-one stool samples from Taiwanese were collected for analysis of the association between the gut flora and obesity. The supervised analysis showed that the most, abundant genera of bacteria in normal samples (from people with a body mass index (BMI) ≤ 24) were *Bacteroides* (27.7%), *Prevotella* (19.4%), *Escherichia* (12%), *Phascolarctobacterium* (3.9%), and *Eubacterium* (3.5%). The most abundant genera of bacteria in case samples (with a BMI ≥ 27) were *Bacteroides* (29%), *Prevotella* (21%), *Escherichia* (7.4%), *Megamonas* (5.1%), and *Phascolarctobacterium* (3.8%). A principal coordinate analysis (PCoA) demonstrated that normal samples were clustered more compactly than case samples. An unsupervised analysis demonstrated that bacterial communities in the gut were clustered into two main groups: N-like and OB-like groups. Remarkably, most normal samples (78%) were clustered in the N-like group, and most case samples (81%) were clustered in the OB-like group (Fisher's P value $= 1.61E-07$). The results showed that bacterial communities in the gut were highly associated with obesity. This is the first study in Taiwan to investigate the association between human gut flora and obesity, and the results provide new insights into the correlation of bacteria with the rising trend in obesity.

1. Background

Enterobacteria, or gut microbiota, in the human gastrointestinal (GI) tract play important roles in the body's functions. For example, they can regulate immune responses and metabolic functions of the host [1–4]. The gut microbiota is frequently used to study the association between human health and an individual's lifestyle. In 2011, three major kinds of enterotypes consisting of *Bacteroides*, *Prevotella*, and *Ruminococcus* [5] provided new perspectives to classify individuals. *Ruminococcus* was reported to be an ambiguous enterotype, and the *Bacteroides* and *Prevotella* enterotypes

were associated with dietary habits [6]. Animal protein and saturated fats were highly correlated with the *Bacteroides* enterotype. Low meat intake and plant-based nutrition with high carbohydrates were correlated with the *Prevotella* enterotype.

The Human Microbiome Project (HMP) [7] was launched by the US National Institutes of Health in 2008, and understanding the relationship between human health and microbiota that live in or on the human body was recognized as an important concept. To decipher this relationship, high-throughput sequencing, also called next-generation sequencing (NGS), supporting a large number of sequences, can be used to sequence 16S ribosomal (r)RNA to construct complex microbial community profiles. 16S rRNA is considered the standard for studying microbial communities and assigning taxonomy to bacteria. Compared to conventional polymerase chain reaction- (PCR-) based or culture-based methods, 16S rRNA sequencing by NGS can detect hundreds to thousands of bacteria at one time and offer relative quantification of the bacteria. Interactions between different bacterial communities and their environments can be comprehensively analyzed by metagenomics research. Associations between diseases and specific bacteria have been described in previous studies, for example, type 2 diabetes [8–11], irritable bowel syndrome (IBS) [12–14], and colorectal cancer (CRC) [15–17]. Moreover, some bacteria which are significantly associated with specific diseases were thought to be biomarkers for construction of a disease risk prediction model [8, 18].

Obesity is a major public health problem worldwide, and its prevalence is rapidly increasing [19]. Obesity is related to several disorders, including type 2 diabetes [20–23], cardiovascular disease [24–26], and cancer [26–28]. Recently, obesity was shown to be associated with an alteration of the gut microbiota, both in human [29–32] and animal models [30, 33, 34]. It was observed that a reduced proportion of the Bacteroidetes and increased proportion of the Firmicutes were associated with human obesity [30, 35, 36]. Also, an increase of *Actinobacteria* in obese individuals was reported [35]. In another study, amounts of Archaea and Methanobacteriales were positively correlated with obesity [37], and their amounts in obesity samples decreased or disappeared after gastric bypass surgery. In addition, another study also mentioned that the amounts of *Bifidobacterium* and *Ruminococcus* decreased in obesity samples [38]. However, some studies indicated that the ratio of the proportions of Bacteroidetes and Firmicutesis contradictory [39] or not associated [5, 40] with obesity.

With different ethnicities and regions, dietary habits and environmental factors can widely vary, and there is a lack of studies focusing on Taiwanese samples. Therefore, herein we collected 81 stool samples from Taiwanese for analysis of the association between gut flora and obesity. According to a study by Pan et al. [41], Taiwan adopted body mass index (BMI) values of 24 and 27 as the cutoff points for being overweight and obese, respectively. In this study, the stools of 36 obese (BMI ≥ 27) and 45 normal persons (BMI ≤ 24) were collected, and 16S rRNA sequencing was used to assess the association between obesity and the taxonomic composition of the gut microbiota.

Table 1: Study participant characteristics and demographics.

Participant characteristic	
Gender (number of samples)	
Male	30
Female	51
Age (years)	
Range	20~89
Mean	41.2
Height (cm)	
Range	148.5~181
Mean	164
Weight (kg)	
Range	45~110
Mean	69.7
Body mass index (kg/m^2) (number of samples)	
≤24	45
≥27	36

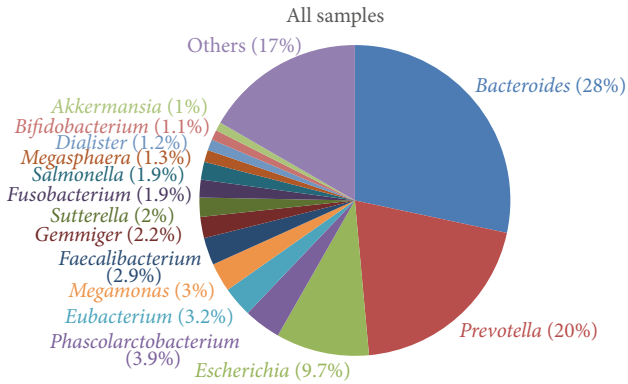


Figure 1: The distribution of genera among all samples.

2. Results

Participant metadata are summarized in Table 1, and detailed sample profiles are given in Table S1 available online in Supplementary material at http://dx.doi.org/10.1155/2014/906168, including the number of reads, gender, age, height, weight, and BMI. In total, 4,152,740 sequence reads were obtained from the 81 samples, and a mean of 51,268 reads with a median read length of 125 bp was obtained per study participant. Sequence reads were processed through our taxonomic mapping process, and the distribution of genera in samples is depicted in Figure 1. The sequencing results showed that the most abundant genera in all samples were *Bacteroides* (28%), *Prevotella* (20%), *Escherichia* (9.7%), *Phascolarctobacterium* (3.9%), *Eubacterium* (3.2%), *Megamonas* (3%), *Faecalibacterium* (2.9%), *Gemmiger* (2.2%), and *Sutterella* (2%).

2.1. Unsupervised Clustering Analysis. Hierarchical clustering was performed using the UniFrac unweighted distance, and gut bacterial communities and clinical values of each sample are shown in Figure 2. The results demonstrate that the bacterial communities in the gut were clustered into two main groups: an N-like group (including the N1 and N2 subgroups) and an OB-like group (including the OB1, OB2, OB3, and OB4 subgroups). Figure 3 shows that the most abundant genera in N-like samples were *Bacteroides* (27.8%), *Prevotella* (18.6%), *Escherichia* (12.7%), *Phascolarctobacterium* (4%), and *Eubacterium* (3.5%). The most abundant genera in OB-like samples were *Bacteroides* (28.8%), *Prevotella* (21.7%), *Escherichia* (7.1%), *Megamonas* (4.4%), and *Phascolarctobacterium* (3.7%). Remarkably, most normal samples (78%) were clustered in the N-like group, and most case samples (81%) were clustered in the OB-like group (Fisher's P value = $1.61E{-}07$). The results showed that gut bacterial community types were highly associated with obesity. The genera diversity analysis showed that the bacterial communities in the N-like group exhibited significantly higher alpha diversity and lower beta diversity than those in the OB-like group (Figure 4).

2.2. Supervised Clustering Analysis. To investigate the association between gut bacterial communities and obesity, 45 stool samples of participants with a BMI of ≤ 24 were defined as normal samples, and 36 samples of participants with a BMI ≥ 27 were used as case samples. Figure 5 shows that the most abundant bacteria in normal samples were *Bacteroides* (27.7%), *Prevotella* (19.4%), *Escherichia* (12%), *Phascolarctobacterium* (3.9%), and *Eubacterium* (3.5%). The most abundant bacteria in case samples were *Bacteroides* (29%), *Prevotella* (21%), *Escherichia* (7.4%), *Megamonas* (5.1%), and *Phascolarctobacterium* (3.8%). Normal samples had a significantly higher proportion of *Escherichia*, while case samples had a higher proportion of *Megamonas*.

Genera with significantly different proportions between normal and case samples are listed in Table 2. Additionally, genera with a significantly different presence between normal and case samples are listed in Table 3, and significantly different species are also provided in Table S2. The genera of *Shewanella, Citrobacter, Cronobacter, Leclercia, Tatumella, and Acinetobacter* exhibited significant differences in both proportions and presence. Unweighted alpha and beta diversities of genera in the normal and case samples are shown in Figures 6(a) and 6(b), respectively. The results showed that the bacterial communities in normal samples exhibited significantly higher alpha diversity and lower beta diversity than those in case
samples.

A PCoA of gut bacterial communities is shown in Figure 7. The results showed that most normal samples (green nodes) were located in the bottom left area, and case samples (red nodes) were spread in other areas (Figure 7(a)). Samples in the N1, N2, OB1, OB2, OB3, and OB4 subgroups are depicted in Figure 7(b). The results show that bacterial communities of N1 and N2 were highly associated with normal-weight individuals, and others were associated with obese individuals.

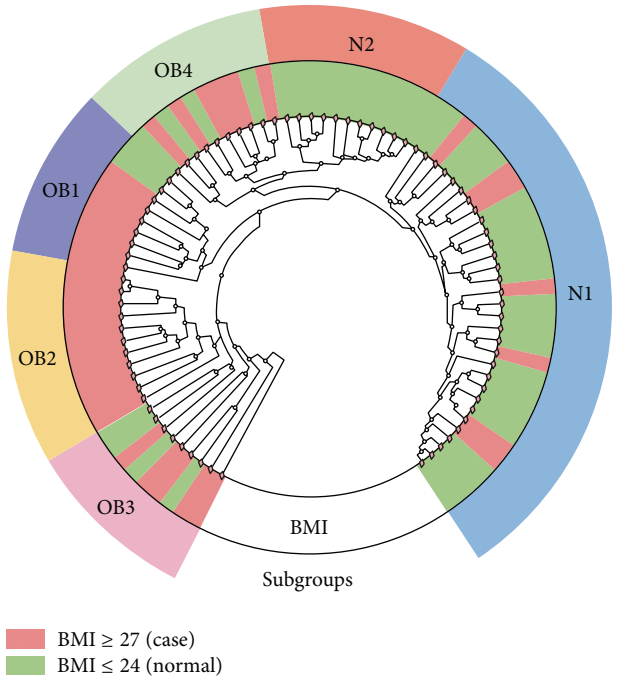


Figure 2: Bacterial communities in the samples.

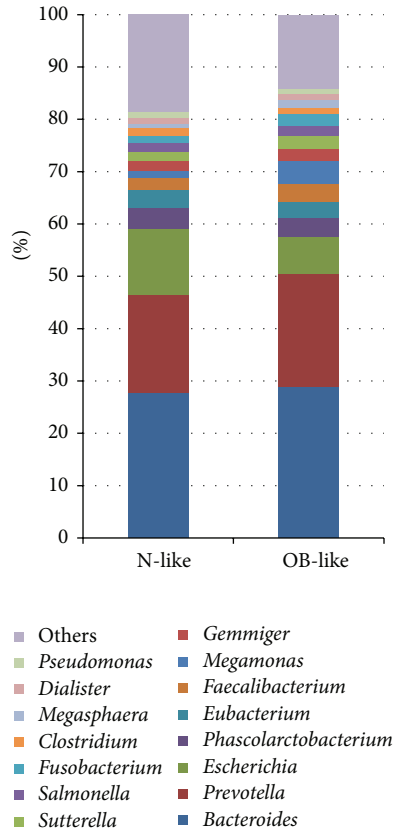


Figure 3: Relatively abundant genera in the N-like and OB-like groups.

Table 2: Genera with significantly different proportions between normal and case samples.

Genus	Fold change (case/control)	KS test *P* value	ANOVA *P* value	Case mean	Control mean
Collinsella	0.56	0.01	0.37	0.001	0.002
Barnesiella	0.61	0.01	0.26	0.002	0.003
Clostridium	0.52	0.01	0.01	0.009	0.017
Coprococcus	0.54	0.04	0.37	0.001	0.001
Lachnospira	1.68	0.02	0.27	0.003	0.002
Oscillibacter	0.55	0.01	0.02	0.001	0.001
Megamonas	5.70	0.08	0.01	0.051	0.009
Veillonella	0.47	0.01	0.27	0.002	0.004
Shewanella	4.69	0.00	0.04	0.001	<0.001
Citrobacter	0.26	0.00	0.11	0.002	0.006
Cronobacter	0.08	0.00	0.00	<0.001	0.002
Enterobacter	0.38	0.00	0.07	0.001	0.002
Erwinia	0.18	0.00	0.00	<0.001	0.002
Escherichia	0.62	0.00	0.04	0.074	0.120
Leclercia	0.09	0.00	0.06	0.001	0.005
Morganella	0.05	0.01	0.07	<0.001	0.001
Serratia	0.02	0.00	0.00	<0.001	0.018
Tatumella	0.06	0.00	0.00	<0.001	0.001
Halomonas	3.51	0.00	0.08	0.005	0.002
Acinetobacter	116.08	0.00	0.03	0.002	<0.001

KS: Kolmogorov-Smirnov; ANOVA: analysis of variance.

Table 3: Genera with a significantly different presence between normal and case samples.

Genus	Presence Case/normal	Absence Case/normal	Fisher's test *P* value	Odds ratio (95% CI)
Butyricimonas	28/44	8/1	0.009	0.082 (0.002~0.664)
Butyrivibrio	0/11	36/34	0.001	0.088 (0.002~0.663)
Lachnobacterium	0/16	36/29	<0.001	0.052 (0.001~0.371)
Lachnospira	22/43	14/2	<0.001	0.076 (0.008~0.373)
Syntrophococcus	0/10	36/35	0.002	0.099 (0.002~0.764)
Pectinatus	0/14	36/31	<0.001	0.063 (0.001~0.460)
Comamonas	0/10	36/35	0.002	0.099 (0.002~0.764)
Pseudoalteromonas	12/1	24/44	<0.001	21.261 (2.837~956.295)
Shewanella	15/3	21/42	<0.001	9.703 (2.381~58.040)
Citrobacter	17/43	19/2	<0.001	0.043 (0.004~0.210)
Cronobacter	6/34	30/11	<0.001	0.068 (0.018~0.218)
Leclercia	8/36	28/9	<0.001	0.075 (0.021~0.233)
Rahnella	0/13	36/32	<0.001	0.070 (0.002~0.516)
Shigella	5/31	31/14	<0.001	0.076 (0.019~0.250)
Tatumella	2/23	34/22	<0.001	0.058 (0.006~0.273)
Marinomonas	12/1	24/44	<0.001	21.261 (2.837~956.295)
Acinetobacter	12/2	24/43	0.001	10.443 (2.070~103.809)
Aliivibrio	8/1	28/44	0.009	12.231 (1.505~568.466)

CI: confidence interval.

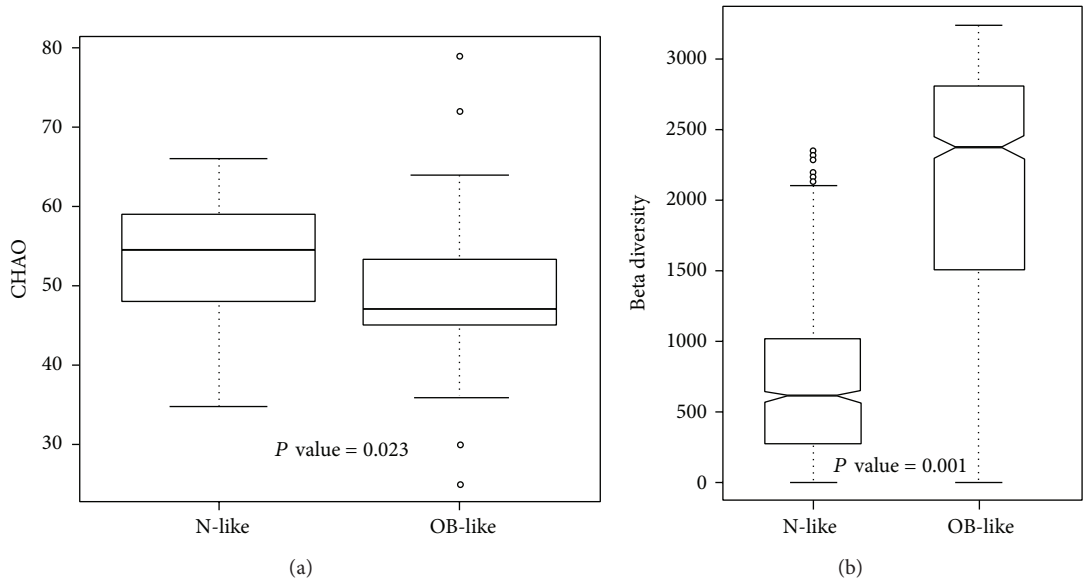


Figure 4: Unweighted (a) alpha diversity and (b) beta diversity of bacterial communities in the N-like and OB-like groups.

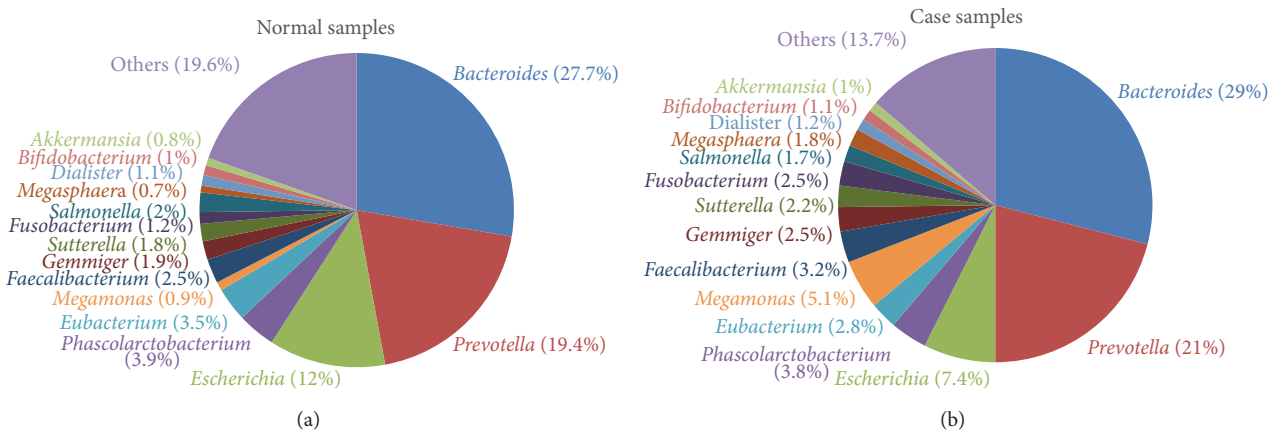


Figure 5: Relatively abundant genera in the normal and case samples.

2.3. Potential Markers for Classification of Normal Weight and Obesity. The identified bacteria with statistical significance were used for rule-based clustering. Threefold cross-validation was used to evaluate the performance of the classification model. Two out of the significant species in Table S2, *Parabacteroides distasonis* and *Serratia* sp. *DAP4*, were selected as discriminating factors in the J48 decision tree. As shown in Figure 8, the classification rules are described as follows. (1) a sample is classified as normal if *Parabacteroides distasonis* was absent. (2) A sample with the presence of *Parabacteroides distasonis* and absence of *Serratia* sp. *DAP4* was classified as a case; otherwise, it was classified as normal. As shown in Table S3, the classifier performed well, and the area under the receiver operating characteristic curve (AUC) was 0.813. The results showed that *Parabacteroides distasonis* and *Serratia* sp. *DAP4* might be potential markers for further clinical analysis and investigation of obesity.

3. Discussion

Several relatively abundant genera were identified in samples (Figure 5), including *Bacteroides*, *Prevotella*, *Escherichia*, *Phascolarctobacterium*, *Eubacterium*, *Megamonas*, *Faecalibacterium*, *Gemmiger*, *Sutterella*, *Fusobacterium*, *Salmonella*, *Megasphaera*, *Dialister*, *Bifidobacterium*, and *Akkermansia*. In previous studies, the presence of *Bacteroides*, *Prevotella*, and *Sutterella* was negatively associated with obesity [36, 42,

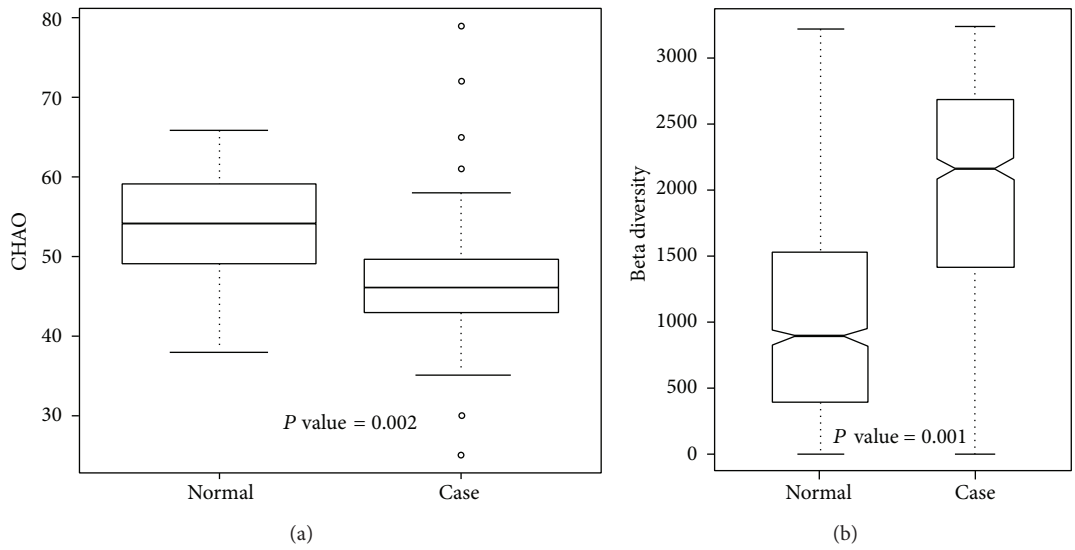


Figure 6: Unweighted (a) alpha diversity and (b) beta diversity of bacterial communities in case and control samples.

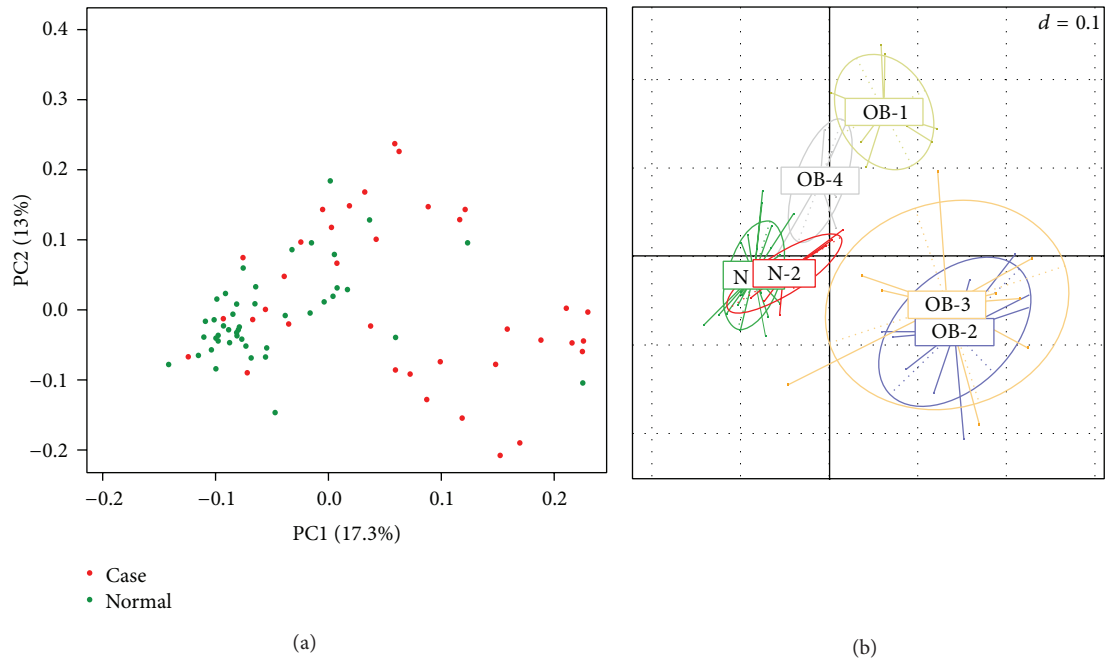


Figure 7: Unweighted principal coordinate analysis plot of (a) case and normal samples, (b) samples in N1, N2, OB1, OB2, OB3, and OB4 subgroups.

43]. In other related gastrointestinal diseases, *Prevotella* was increased in children diagnosed with IBS [44]. *Bacteroides*, *Eubacterium,* and *Prevotella* were increased, and *Faecalibacterium* was reduced in CRC patients [45, 46]. Increased *Bacteroides* and reduced *Eubacterium* and *Prevotella* were also found in a rat model of CRC [47].

At the genus level, the presence of *Acinetobacter*, *Aliivibrio*, *Marinomonas*, *Pseudoalteromonas*, and *Shewanella* had positive associations with obesity (Table 3). *Acinetobacter* is a genus of Gram-negative bacteria, and the species *Acinetobacter baumannii* is a key pathogen of infections in hospitals [48]. *Aliivibrio* is a reclassified genus from the "*Vibrio fischeri* species group" [49], and species of *Aliivibrio* are symbiotic with marine animals or are described as fish pathogens [50–52]. *Shewanella* is a genus of marine bacteria, and some species can cause infections [53, 54]. *Lachnospira*

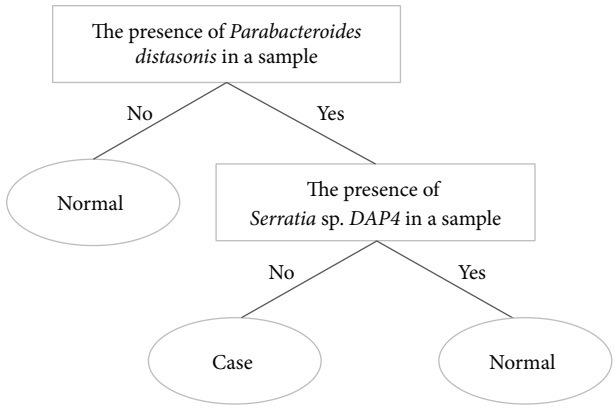


Figure 8: Classification rule and potential markers for discriminating between obese and normal-weight individuals.

[37], *Citrobacter* [43], and *Shigella* [43] were reported to be positively associated with obesity. *Lachnobacterium* [18] showed a negative association with obesity.

At the species level, the presence of *Parabacteroides distasonis*, *Lactobacillus kunkeei*, *Pseudoalteromonas piscicida*, *Shewanella algae*, *Marinomonas posidonica*, and *Aliivibrio fischeri* was positively associated with obesity (Table S2). Species of *Bacteroides* and *Parabacteroides* represent opportunistic pathogens in infectious diseases, and they are able to develop antimicrobial drug resistance [55]. *Parabacteroides distasonis*, previously known as *Bacteroides distasonis* [56], is prominently found in the gut of healthy individuals [57]. It is also related to improved human bowel health release [58] and negatively associated with celiac disease [59]. Our results revealed a positive association between *Parabacteroides distasonis* and obesity. *Blautia producta* [60] and *Enterobacter cloacae* [61] were suggested to be related to a high-fat diet causing obesity in a mouse model. *Serratia* is a genus of Gram-negative, facultatively anaerobic, rod-shaped bacteria. In hospitals, *Serratia* species tend to colonize the respiratory and urinary tracts causing nosocomial infections [62, 63]. In related studies of the GI tract, *Serratia* increased in formula-fed mice [64] and was positively correlated with infants with colic [65].

In the alpha diversity analysis (Figure 6(a)), the Chao richness index between normal and case groups exhibited a significant difference ($P = 0.002$). This shows that bacterial communities in normal samples had a greater genera richness than those in case samples. Results of the beta diversity analysis (Figure 6(b)) showed that bacterial communities in normal samples were more similar than those in case samples. The unweighted PCoA plot (Figure 7) showed that bacterial communities in the N-like group (including N1 and N2) were highly associated with normal individuals, and bacterial communities in the OB-like group (including OB1, OB2, OB3, and OB4) were more associated with obese individuals. The unsupervised clustering heatmap of all samples (Figure S1(A)) was generated using Spearman correlations, and the results showed that most normal samples and most case samples were, respectively, clustered together, when all genera were used for clustering. However, when only relatively abundant genera were used for clustering, normal, and case samples were interwoven with each other (Figure S1(B)). This indicates that some genera found in small proportions might be important for distinguishing obese from normal individuals.

4. Conclusions

This is the first study in Taiwan to investigate the association between human gut microbiota and obesity using metagenomic sequencing. The results showed that bacterial communities in the gut were clustered into N-like and OB-like groups which were highly associated with normal and obese subjects, respectively. Several relatively abundant bacteria with significantly different distributions between normal and case samples were identified and used to establish a rule-based classification model. Although detailed functional roles or mechanisms of these bacteria are needed for further validation, the results provide new insights about bacterial communities in the gut with a rising trend of obesity.

5. Methods

5.1. Sample Collection and DNA Extraction. Eighty-one stool samples were collected by Sigma-transwab (Medical Wire) into a tube with Liquid Amies Transport Medium and stored at 4°C until being processed. Fresh faeces were obtained from participants, and DNA was directly extracted from stool samples using a QIAamp DNA Stool Mini Kit (Qiagen). A swab was vigorously vortexed and incubated at room temperature for 1 min. The sample was transferred to microcentrifuge tubes containing 560 μL of Buffer ASL, then vortexed, and incubated at 37°C for 30 min. In addition, the suspension was incubated at 95°C for 15 min, vortexed, and centrifuged at 14,000 rpm for 1 min to obtain pelletized stool particles. Extraction was performed following the protocol of the QIAamp DNA Stool Mini Kit. DNA was eluted with 50 μL Buffer AE, centrifuged at 14,000 rpm for 1 min, and then the DNA extract was stored at −20°C until being further analyzed.

5.2. Library Construction and Sequencing of the V4 Region of 16S rDNA. The PCR primers, F515 (5′-GTGCCAGCMGCCGCGGTAA-3′) and R806 (5′-GGACTACHVGGGTWTCTAAT-3′), were designed to amplify the V4 region of bacterial 16S rDNA as described previously [66]. Polymerase chain reaction (PCR) amplification was performed in a 50 μL reaction volume containing 25 μL 2x Taq Master Mix (Thermo Scientific), 0.2 μM of each forward and reverse primer, and 20 ng of a DNA template. The reaction conditions included an initial temperature of 95°C for 5 min, followed by 30 cycles of 95°C for 30 s, 54°C for 1 min, and 72°C for 1 min, with a final extension of 72°C for 5 min. Next, amplified products were checked by 2% agarose gel electrophoresis

and ethidium bromide staining. Amplicons were purified using the AMPure XP PCR Purification Kit (Agencourt) and quantified using the Qubit dsDNA HS Assay Kit (Qubit) on a Qubit 2.0 Fluorometer (Qubit), all according to the respective manufacturer's instructions. For V4 library preparation, Illumina adapters were attached to the amplicons using the Illumina TruSeq DNA Sample Preparation v2 Kit. Purified libraries were processed for cluster generation and sequencing using the MiSeq system.

5.3. Filtering 16S rRNA (rDNA) Sequencing Data for Quality. The FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit) was used to process the raw fastq read data files from Illumina Miseq. The sequence quality criteria were as follows: (1) the minimum acceptable phred quality score of sequences was 30 with a score of >70% of sequence bases of ≥20; (2) after quality trimming from the sequence tail, sequences of >100 bp were retained, and they also had an acceptable phred quality score of 30; and (3) both forward and reverse sequencing reads which met the first and second requirements were retained for subsequent analysis. Sequencing reads from different samples were identified and separated according to specific barcodes in the 5′ end of the sequence (with two mismatches allowed).

5.4. Taxonomic Assignments of Bacterial 16S rRNA Sequences. Paired-end sequences were obtained, and their qualities were assessed using the FASTX-Toolkit. To generate taxonomic assignments, Bowtie2 was used to align sequencing reads against the collection of a 16S rRNA sequences database. A standard of 97% similarity against the database was applied. 16S rRNA sequences of bacteria were retrieved from the SILVA ribosomal RNA sequence database [67]. Following sequence data collection, sequences were extracted using V4 forward and reverse primers. To prevent repetitive sequence assignments, V4 sequences from SILVA were then clustered into several clusters by 97% similarity using UCLUST [68]. Results of the taxonomic assignment were filtered to retain assignments with >10 sequences.

5.5. Bacterial Community Analysis. After taxonomic assignment, an operational taxonomic unit (OTU) table was generated. To normalize the sample size of all samples, a rarefaction process was performed on the OTU table. Alpha and beta diversities were calculated based on a rarified OTU table. The Kolmogorov-Smirnov test and an analysis of variance (ANOVA) test with the Bonferroni correction were used to investigate significant differences between different sample groups. To observe relationships between samples and explore taxonomic associations, weighted and unweighted UniFrac [69] distance metrics were also generated based on the rarified OTU table. A principal coordinate analysis (PCoA) and unsupervised clustering were performed based on the UniFac distance matrix. To explore relationships between clinical features and different sample groups, Spearman's correlation coefficient and regression analysis were performed. The statistical analytical process was done in R language. The J48 machine learning method in Weka 3.6.7 [70] was used to construct a classification rule for discriminating between obese and normal individuals.

Conflict of Interests

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

Authors' Contribution

Chih-Min Chiu, Wei-Chih Huang, and Shun-Long Weng contributed equally to this work.

Acknowledgments

The authors would like to thank the National Science Council of the Republic of China for financially supporting this research under Grant nos. NSC 101-2311-B-009-005-MY3 and NSC 102-2627-B-009-001. This work was supported in part by the Ministry of Science and Technology under Grant no. MOST 103-2221-E-038-013-MY2. This work was supported in part by UST-UCSD International Center of Excellence in Advanced Bioengineering sponsored by the Ministry of Science and Technology I-RiCE Program under Grant no. NSC-102-2911-I-009-101-, and Veterans General Hospitals and University System of Taiwan (VGHUST) Joint Research Program under Grant no. VGHUST103-G5-1-2. This work was also partially supported by MOE ATU.

References

[1] J. R. Goldsmith and R. B. Sartor, "The role of diet on intestinal microbiota metabolism: downstream impacts on host immune function and health, and therapeutic implications," *Journal of Gastroenterology*, vol. 49, no. 5, pp. 785–798, 2014.

[2] J. Chen, X. He, and J. Huang, "Diet effects in gut microbiome and obesity," *Journal of Food Science*, vol. 79, no. 4, pp. R442–R451, 2014.

[3] O. O. Erejuwa, S. A. Sulaiman, and M. S. Wahab, "Modulation of gut microbiota in the management of metabolic disorders: the prospects and challenges," *International Journal of Molecular Sciences*, vol. 15, no. 3, pp. 4158–4188, 2014.

[4] G. L. Hold, M. Smith, C. Grange, E. R. Watt, E. M. El-Omar, and I. Mukhopadhya, "Role of the gut microbiota in inflammatory bowel disease pathogenesis: what have we learnt in the past 10 years?" *World Journal of Gastroenterology*, vol. 20, no. 5, pp. 1192–1210, 2014.

[5] M. Arumugam, J. Raes, E. Pelletier et al., "Enterotypes of the human gut microbiome," *Nature*, vol. 473, no. 7346, pp. 174–180, 2011.

[6] G. D. Wu, J. Chen, C. Hoffmann et al., "Linking long-term dietary patterns with gut microbial enterotypes," *Science*, vol. 334, no. 6052, pp. 105–108, 2011.

[7] J. Peterson, S. Garges, M. Giovanni et al., "The NIH human microbiome project," *Genome Research*, vol. 19, no. 12, pp. 2317–2323, 2009.

[8] J. Qin, Y. Li, Z. Cai et al., "A metagenome-wide association study of gut microbiota in type 2 diabetes," *Nature*, vol. 490, no. 7418, pp. 55–60, 2012.

[9] F. H. Karlsson, V. Tremaroli, I. Nookaew et al., "Gut metagenome in European women with normal, impaired and diabetic glucose control," *Nature*, vol. 498, no. 7452, pp. 99–103, 2013.

[10] X. Zhang, D. Shen, Z. Fang et al., "Human gut microbiota changes reveal the progression of glucose intolerance," *PLoS ONE*, vol. 8, no. 8, Article ID e71108, 2013.

[11] F. Karlsson, V. Tremaroli, J. Nielsen, and F. Backhed, "Assessing the human gut microbiota in metabolic diseases," *Diabetes*, vol. 62, no. 10, pp. 3341–3349, 2013.

[12] A. Lyra, T. Rinttilä, J. Nikkilä et al., "Diarrhoea-predominant irritable bowel syndrome distinguishable by 16S rRNA gene phylotype quantification," *World Journal of Gastroenterology*, vol. 15, no. 47, pp. 5936–5945, 2009.

[13] S. O. Noor, K. Ridgway, L. Scovell et al., "Ulcerative colitis and irritable bowel patients exhibit distinct abnormalities of the gut microbiota," *BMC Gastroenterology*, vol. 10, article 134, 2010.

[14] M. Rajilić-Stojanović, E. Biagi, H. G. H. J. Heilig et al., "Global and deep molecular analysis of microbiota signatures in fecal samples from patients with irritable bowel syndrome," *Gastroenterology*, vol. 141, no. 5, pp. 1792–1801, 2011.

[15] Y. Zhu, T. Michelle Luo, C. Jobin, and H. A. Young, "Gut microbiota and probiotics in colon tumorigenesis," *Cancer Letters*, vol. 309, no. 2, pp. 119–127, 2011.

[16] M. A. Hullar, A. N. Burnett-Hartman, and J. W. Lampe, "Gut microbes, diet, and cancer," *Cancer Treatment and Research*, vol. 159, pp. 377–399, 2014.

[17] H. R. Hagland and K. Soreide, "Cellular metabolism in colorectal carcinogenesis: influence of lifestyle, gut microbiome and metabolic pathways," *Cancer Letters*, 2014.

[18] K. Korpela, H. J. Flint, A. M. Johnstone et al., "Gut microbiota signatures predict host and microbiota responses to dietary interventions in obese individuals," *PLoS One*, vol. 9, no. 3, Article ID e90702, 2014.

[19] B. A. Swinburn, G. Sacks, K. D. Hall et al., "The global obesity pandemic: shaped by global drivers and local environments," *The Lancet*, vol. 378, no. 9793, pp. 804–814, 2011.

[20] K. C. Portero McLellan, K. Wyne, E. T. Villagomez, and W. A. Hsueh, "Therapeutic interventions to reduce the risk of progression from prediabetes to type 2 diabetes mellitus," *Therapeutics and Clinical Risk Management*, vol. 10, pp. 173–188, 2014.

[21] A. Everard and P. D. Cani, "Diabetes, obesity and gut microbiota," *Best Practice and Research: Clinical Gastroenterology*, vol. 27, no. 1, pp. 73–83, 2013.

[22] J. A. Bell, M. Kivimaki, and M. Hamer, "Metabolically healthy obesity and risk of incident type 2 diabetes: a meta-analysis of prospective cohort studies," *Obesity Reviews*, vol. 15, no. 6, pp. 504–515, 2014.

[23] D. Nagakubo, M. Shirai, Y. Nakamura et al., "Prophylactic effects of the glucagon-like Peptide-1 analog liraglutide on hyperglycemia in a rat model of type 2 diabetes mellitus associated with chronic pancreatitis and obesity," *Comparative Medicine*, vol. 64, no. 2, pp. 121–127, 2014.

[24] M. Kratz, T. Baars, and S. Guyenet, "The relationship between high-fat dairy consumption and obesity, cardiovascular, and metabolic disease," *European Journal of Nutrition*, vol. 52, no. 1, pp. 1–24, 2013.

[25] G. M. Hinnouho, S. Czernichow, A. Dugravot et al., "Metabolically healthy obesity and the risk of cardiovascular disease and type 2 diabetes: The Whitehall II Cohort Study," *European Heart Journal*, 2014.

[26] K. A. Britton, J. M. Massaro, J. M. Murabito, B. E. Kreger, U. Hoffmann, and C. S. Fox, "Body fat distribution, incident cardiovascular disease, cancer, and all-cause mortality," *Journal of the American College of Cardiology*, vol. 62, no. 10, pp. 921–925, 2013.

[27] E. E. Frezza, M. S. Wachtel, and M. Chiriva-Internati, "Influence of obesity on the risk of developing colon cancer," *Gut*, vol. 55, no. 2, pp. 285–291, 2006.

[28] K. Nakamura, A. Hongo, J. Kodama, and Y. Hiramatsu, "Fat accumulation in adipose tissues as a risk factor for the development of endometrial cancer," *Oncology Reports*, vol. 26, no. 1, pp. 65–71, 2011.

[29] M. Remely, E. Aumueller, D. Jahn, B. Hippe, H. Brath, and A. G. Haslberger, "Microbiota and epigenetic regulation of inflammatory mediators in type 2 diabetes and obesity," *Beneficial Microbes*, vol. 5, no. 1, pp. 33–43, 2014.

[30] R. E. Ley, P. J. Turnbaugh, S. Klein, and J. I. Gordon, "Microbial ecology: human gut microbes associated with obesity," *Nature*, vol. 444, no. 7122, pp. 1022–1023, 2006.

[31] J. Shen, M. S. Obin, and L. Zhao, "The gut microbiota, obesity and insulin resistance," *Molecular Aspects of Medicine*, vol. 34, no. 1, pp. 39–58, 2013.

[32] M. Nieuwdorp, P. W. Gilijamse, N. Pai, and L. M. Kaplan, "Role of the microbiome in energy regulation and metabolism," *Gastroenterology*, vol. 146, no. 6, pp. 1525–1533, 2014.

[33] P. J. Turnbaugh, F. Bäckhed, L. Fulton, and J. I. Gordon, "Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome," *Cell Host and Microbe*, vol. 3, no. 4, pp. 213–223, 2008.

[34] F. Bäckhed, H. Ding, T. Wang et al., "The gut microbiota as an environmental factor that regulates fat storage," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 44, pp. 15718–15723, 2004.

[35] P. J. Turnbaugh, M. Hamady, T. Yatsunenko et al., "A core gut microbiome in obese and lean twins," *Nature*, vol. 457, no. 7228, pp. 480–484, 2009.

[36] J. Furet, L. Kong, J. Tap et al., "Differential adaptation of human gut microbiota to bariatric surgery-induced weight loss: links with metabolic and low-grade inflammation markers," *Diabetes*, vol. 59, no. 12, pp. 3049–3057, 2010.

[37] H. Zhang, J. K. DiBaise, A. Zuccolo et al., "Human gut microbiota in obesity and after gastric bypass," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 7, pp. 2365–2370, 2009.

[38] Y. N. Yin, Q. F. Yu, N. Fu, X. W. Liu, and F. G. Lu, "Effects of four Bifidobacteria on obesity in high-fat diet induced rats," *World Journal of Gastroenterology*, vol. 16, no. 27, pp. 3394–3401, 2010.

[39] R. E. Ley, "Obesity and the human microbiome," *Current Opinion in Gastroenterology*, vol. 26, no. 1, pp. 5–11, 2010.

[40] The Human Microbiome Project Consortium, "Structure, function and diversity of the healthy human microbiome," *Nature*, vol. 486, no. 7402, pp. 207–214, 2012.

[41] W. Pan, K. M. Flegal, H. Chang, W. Yeh, C. Yeh, and W. Lee, "Body mass index and obesity-related metabolic disorders in Taiwanese and US whites and blacks: Implications for definitions of overweight and obesity for Asians," *The American Journal of Clinical Nutrition*, vol. 79, no. 1, pp. 31–39, 2004.

[42] A. Santacruz, M. C. Collado, L. García-Valdés et al., "Gut microbiota composition is associated with body weight, weight gain and biochemical parameters in pregnant women," *British Journal of Nutrition*, vol. 104, no. 1, pp. 83–92, 2010.

[43] S. Xiao, N. Fei, X. Pang et al., "A gut microbiota-targeted dietary intervention for amelioration of chronic inflammation underlying metabolic syndrome," *FEMS Microbiology Ecology*, vol. 87, no. 2, pp. 357–367, 2014.

[44] L. Rigsbee, R. Agans, V. Shankar et al., "Quantitative profiling of gut microbiota of children with diarrhea-predominant irritable bowel syndrome," *The American Journal of Gastroenterology*, vol. 107, no. 11, pp. 1740–1751, 2012.

[45] N. Wu, X. Yang, R. Zhang et al., "Dysbiosis signature of fecal microbiota in colorectal cancer patients," *Microbial Ecology*, vol. 66, no. 2, pp. 462–470, 2013.

[46] W. Chen, F. Liu, Z. Ling, X. Tong, and C. Xiang, "Human intestinal lumen and mucosa-associated microbiota in patients with colorectal cancer," *PLoS ONE*, vol. 7, no. 6, Article ID e39743, 2012.

[47] Q. Zhu, Z. Jin, W. Wu et al., "Analysis of the intestinal lumen microbiota in an animal model of colorectal cancer," *PLoS ONE*, vol. 9, no. 3, Article ID e90849, 2014.

[48] D. Wang, D. Yan, W. Hou, X. Zeng, Y. Qi, and J. Chen, "Characterization of bla_{OxA-23}gene regions in isolates of *Acinetobacter baumannii*," *Journal of Microbiology, Immunology and Infection*, 2014.

[49] H. Urbanczyk, J. C. Ast, M. J. Higgins, J. Carson, and P. V. Dunlap, "Reclassification of *Vibrio fischeri*, *Vibrio logei*, *Vibrio salmonicida* and *Vibrio wodanis* as *Aliivibrio fischeri* gen. nov., comb. nov., *Aliivibrio logei* comb. nov., *Aliivibrio salmonicida* comb. nov. and *Aliivibrio wodanis* comb. nov," *International Journal of Systematic and Evolutionary Microbiology*, vol. 57, part 12, pp. 2823–2829, 2007.

[50] J. C. Ast, H. Urbanczyk, and P. V. Dunlap, "Multi-gene analysis reveals previously unrecognized phylogenetic diversity in *Aliivibrio*," *Systematic and Applied Microbiology*, vol. 32, no. 6, pp. 379–386, 2009.

[51] R. Beaz-Hidalgo, A. Doce, S. Balboa, J. L. Barja, and J. L. Romalde, "*Aliivibrio finisterrensis* sp. nov., isolated from Manila clam, *Ruditapes philippinarum* and emended description of the genus Aliivibrio," *International Journal of Systematic and Evolutionary Microbiology*, vol. 60, no. 1, pp. 223–228, 2010.

[52] S. Yoshizawa, H. Karatani, M. Wada, A. Yokota, and K. Kogure, "*Aliivibrio sifiae* sp. nov., luminous marine bacteria isolated from seawater," *Journal of General and Applied Microbiology*, vol. 56, no. 6, pp. 509–518, 2010.

[53] Y. S. Chen, Y. C. Liu, M. Y. Yen, J. H. Wang, S. R. Wann, and D. L. Cheng, "Skin and soft-tissue manifestations of Shewanella putrefaciens infection," *Clinical Infectious Diseases*, vol. 25, no. 2, pp. 225–229, 1997.

[54] N. Vignier, M. Barreau, C. Olive et al., "Human infection with *Shewanella putrefaciens* and *S. algae*: report of 16 cases in Martinique and review of the literature," *American Journal of Tropical Medicine and Hygiene*, vol. 89, no. 1, pp. 151–156, 2013.

[55] R. F. Boente, L. Q. Ferreira, L. S. Falcão et al., "Detection of resistance genes and susceptibility patterns in Bacteroides and Parabacteroides strains," *Anaerobe*, vol. 16, no. 3, pp. 190–194, 2010.

[56] M. Sakamoto and Y. Benno, "Reclassification of Bacteroides distasonis, Bacteroides goldsteinii and Bacteroides merdae as Parabacteroides distasonis gen. nov., comb. nov., Parabacteroides goldsteinii comb. nov and Parabacteroides merdae comb. nov," *International Journal of Systematic and Evolutionary Microbiology*, vol. 56, part 7, pp. 1599–1605, 2006.

[57] J. Xu, M. A. Mahowald, R. E. Ley et al., "Evolution of symbiotic bacteria in the distal human intestine.," *PLoS Biology*, vol. 5, no. 7, Article ID e156, 2007.

[58] J. M. Clarke, D. L. Topping, C. T. Christophersen et al., "Butyrate esterified to starch is released in the human gastrointestinal tract," *The American Journal of Clinical Nutrition*, vol. 94, no. 5, pp. 1276–1283, 2011.

[59] E. Sánchez, E. Donat, C. Ribes-Koninckx, M. Calabuig, and Y. Sanz, "Intestinal *Bacteroides* species associated with coeliac disease," *Journal of Clinical Pathology*, vol. 63, no. 12, pp. 1105–1111, 2010.

[60] N. Becker, J. Kunath, G. Loh, and M. Blaut, "Human intestinal microbiota: characterization of a simplified and stable gnotobiotic rat model," *Gut Microbes*, vol. 2, no. 1, 2011.

[61] N. Fei and L. Zhao, "An opportunistic pathogen isolated from the gut of an obese human causes obesity in germfree mice," *ISME Journal*, vol. 7, no. 4, pp. 880–884, 2013.

[62] A. Hejazi and F. R. Falkiner, "Serratia marcescens," *Journal of Medical Microbiology*, vol. 46, no. 11, pp. 903–912, 1997.

[63] M. Patankar, S. Sukumaran, A. Chhibba, U. Nayak, and L. Sequeira, "Comparative in-vitro activity of cefoperazone-tazobactam and cefoperazone-sulbactam combinations against ESBL pathogens in respiratory and urinary infections," *Journal of Association of Physicians of India*, vol. 60, no. 11, pp. 22–24, 2012.

[64] E. M. Carlisle, V. Poroyko, M. S. Caplan, J. Alverdy, M. J. Morowitz, and D. Liu, "Murine gut microbiota and transcriptome are diet dependent," *Annals of Surgery*, vol. 257, no. 2, pp. 287–294, 2013.

[65] C. de Weerth, S. Fuentes, P. Puylaert, and W. M. de vos, "Intestinal microbiota of infants with colic: development and specific signatures," *Pediatrics*, vol. 131, no. 2, pp. e550–e558, 2013.

[66] J. G. Caporaso, C. L. Lauber, W. A. Walters et al., "Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 1, pp. 4516–4522, 2011.

[67] E. Pruesse, C. Quast, K. Knittel et al., "SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB," *Nucleic Acids Research*, vol. 35, no. 21, pp. 7188–7196, 2007.

[68] R. C. Edgar, "Search and clustering orders of magnitude faster than BLAST," *Bioinformatics*, vol. 26, no. 19, pp. 2460–2461, 2010.

[69] C. Lozupone, M. E. Lladser, D. Knights, J. Stombaugh, and R. Knight, "UniFrac: an effective distance metric for microbial community comparison," *ISME Journal*, vol. 5, no. 2, pp. 169–172, 2011.

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

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 940821, 9 pages
http://dx.doi.org/10.1155/2014/940821

Research Article

Studying the Complex Expression Dependences between Sets of Coexpressed Genes

Mario Huerta,[1] Oriol Casanova,[2] Roberto Barchino,[2] Jose Flores,[2] Enrique Querol,[1] and Juan Cedano[3]

[1] *Institut de Biotecnologia i Biomedicina, Universitat Autònoma de Barcelona, Bellaterra, 08193 Barcelona, Spain*
[2] *Escola Tècnica Superior de Ingenieria, Universitat Autònoma de Barcelona, Bellaterra, 08193 Barcelona, Spain*
[3] *Laboratorio de Inmunología, CENUR del Noroeste, Universidad de la República-Salto, Rivera 1350, 50000 Salto, Uruguay*

Correspondence should be addressed to Juan Cedano; jcedano@unorte.edu.uy

Received 8 April 2014; Revised 13 June 2014; Accepted 6 July 2014; Published 24 July 2014

Academic Editor: Julia Tzu-Ya Weng

Organisms simplify the orchestration of gene expression by coregulating genes whose products function together in the cell. The use of clustering methods to obtain sets of coexpressed genes from expression arrays is very common; nevertheless there are no appropriate tools to study the expression networks among these sets of coexpressed genes. The aim of the developed tools is to allow studying the complex expression dependences that exist between sets of coexpressed genes. For this purpose, we start detecting the nonlinear expression relationships between pairs of genes, plus the coexpressed genes. Next, we form networks among sets of coexpressed genes that maintain nonlinear expression dependences between all of them. The expression relationship between the sets of coexpressed genes is defined by the expression relationship between the *skeletons* of these sets, where this *skeleton* represents the coexpressed genes with a well-defined nonlinear expression relationship with the *skeleton* of the other sets. As a result, we can study the nonlinear expression relationships between a target gene and other sets of coexpressed genes, or start the study from the *skeleton* of the sets, to study the complex relationships of activation and deactivation between the sets of coexpressed genes that carry out the different cellular processes present in the expression experiments.

1. Introduction

Organisms have evolved to vary internal and external cell environments by carefully controlling the abundance and activity of these proteins to suit their conditions. To simplify this task, genes whose products function together are often under common regulatory control. This regulatory control is such that these genes are coordinately expressed under the appropriate conditions. The experimental observation that a set of genes is coexpressed frequently implies that the genes share a biological function and are under common regulatory control [1]. These regulators that govern the expression of sets of coexpressed genes that carry out the appropriate cell functions are also regulated and synchronized among them. Nevertheless, the regulation and synchronization among the regulatory mechanisms is much more complex. These regulatory mechanisms are not directly regulated by the other regulatory mechanisms, but by the coexpressed genes product of their activation cascade. These coexpressed genes switch from the inhibition to the allowance of a regulatory mechanism depending on if they reach or lose certain expression levels. Furthermore, these regulatory mechanisms are multiregulated. The final activation or deactivation of a regulatory process will depend not only on internal and external factors to the cell but also on the expression level of a set of coexpressed genes. For this reason, to study the synchronization and regulation among the regulatory mechanisms based on gene expression is really difficult. Furthermore, many proteins have multiple roles in the cell and act with distinct sets of cooperating proteins to fulfil each role. The genes that synthesize these proteins are therefore coexpressed with different sets of genes, each one governed by a distinct regulatory mechanism, in response to the varying demands of the cell. The experimental conditions will

determine the regulatory mechanism activated in each case and thus the set of coexpressed genes that will be activated. An increased number of different experimental conditions for the same genes will provide less genes in each set of coexpressed genes, more different sets, and higher alternation of the activation and deactivation of the coexpressed genes. But irrespective of the conditions, how can we study the effect of their activation and deactivation on the rest of sets of coexpressed genes?

Microarray technology, as well as the new techniques of next generation sequencing (NGS), allows us to obtain large size gene expression arrays [2]. These gene expression arrays contain the expression levels of thousands of genes for tens of sample conditions. The usual analyses of these data are focused on the differentially expressed genes [3] as well as on the use of clustering methods to obtain the sets of coexpressed genes. It is of vital importance to detect these clusters of coexpressed genes, among other reasons, because, as mentioned before, these clusters of coexpressed genes carry out the different cellular functions [1]. There are powerful coexpression analysis tools for this purpose [4, 5]. As expression arrays allow simultaneous analyses of thousands of genes, we can study genes responsible for very diverse cellular functions. Therefore, it makes it easier for the researcher to understand the cellular behaviour in the performed experiments from a holistic point of view, that is, involving the largest number of cellular processes possible. Without this holistic point of view it is very difficult to deal with the multiple functions, phenotypes, or states of the living beings, in which a large amount of genes are collaborating. This holistic point of view can be useful to characterize phenotypes previously unknown, like for instance the description of the "fish fever" in zebrafish [6]. Nevertheless, even though clustering methods allow us to obtain coexpressed genes and thus differentiate diverse cellular processes, clustering methods explain very little of the expression relationships between the different sets of coexpressed genes. As a consequence, the researcher is constrained to study each one of the coexpressed gene sets individually, losing much of the potential that the technologies for obtaining gene expression arrays offer.

Current statistical technologies allow us to study inclusion relationships between clusters of coexpressed genes, that is, to study which clusters of coexpressed genes would be more correlated and which others would be more uncorrelated [7]. But they do not go much beyond that. The main obstacle to the tools that attempt to study the regulation of coexpressed genes is that the low number of copies of the regulatory genes impedes the correct capture of their expression by technologies to obtain gene expression. Furthermore, the coexpression of genes is strongly linked to the chromatin structure, especially in multicellular organisms. This chromatin structure depends on a complex network of cellular signalling and posttranscriptional modifications of proteins (phosphorylation, acetylation, ubiquitination, etc.). So, even having detected the regulatory genes, we would have an incomplete puzzle. All of this makes it enormously difficult for the tools to be able to obtain information about the regulation among the sets of coexpressed genes and the processes they carry out. Then, without these regulatory elements, how can we describe a regulatory network among the processes performed by the sets of coexpressed genes? Based on the dependences between the gene expressions product of this complex regulation. Furthermore, these final genes present expression ranges wide enough so the fluctuations of their expression dependences can be analysed.

With the developed tools, we expect to detect the complex expression dependences between the different sets of coexpressed genes, synthesize them, and make them easier for the researcher to interpret. With this purpose we will provide the researcher with networks that show the expression dependences between all the coexpressed gene sets of the network. In this way, the researcher will be able to study the alternation or synchronism among all these sets of coexpressed genes.

Our methodology is based on the following three principles. First, the interdependence between sets of coexpressed genes cannot be described by linear expression relationships. Second, if two genes maintain an expression relationship with a certain type of curve, the genes coexpressed with these two genes maintain expression relationships of the same type between them. Third, the curve type of the intergroup expression relationships will describe the dependence of activation and deactivation between these sets of coexpressed genes. Thus, the strategy proposed here is focused on the detection of nonlinear expression relationships between sets of coexpressed genes.

Activation and deactivation dependences between sets of coexpressed genes can be very complex; some of the most common ones are those in which a set of coexpressed genes act as a trigger of another set of coexpressed genes; the case of antagonist processes, where the coexpressed gene set that carries out each process needs to be totally deactivated so the other set can express; or sets of coexpressed genes that activate or deactivate another set of coexpressed genes when losing their basal values of expression. In any case, the system does not anticipate any type of expression relationship. Since the system is able to recognize curves of very different shapes, it can process unknown activation and deactivation relationships as reliably as when processing the best known relationships.

There are multiple works that highlight the relevance of the analysis of nonlinear expression relationships [8–10]. In the last cited work [10] the difference of considering the nonlinear relationships with respect to considering only the linear relationships can be observed very clearly.

As it is shown in the mentioned paper [10], nonlinear expression relationships allow detecting new hubs in gene networks, because they allow relating genes by complex expression relationships and to discover new relationships that otherwise would not be possible to detect. Now, the next step in the analysis based on nonlinear expression relationships is to deepen in the study of complex expression dependences between sets of coexpressed genes. In order to perform this task we need to detect the types of curve of the expression relationships, since the type of curve describes the type of relationship between the sets of coexpressed genes and thus between the processes that these sets of genes carry out.

The ultimate goal of our approach is that researchers are able to know the networks of processes hidden in their

experimental data, as well as the activation and deactivation relationships between all of these processes. Furthermore, if the researcher is particularly interested in specific genes, the system will allow him/her to study the way the expression of a gene activates and deactivates different processes from the process that this gene, and those genes coexpressed with it, carry out.

2. Methodology

2.1. The Suitable Expression Data to Be Analysed. The appropriate data to be analysed by our methodology must come from large sample series (i.e., expression matrices with a high number of sample conditions). This large sample series will not consist of repetitions of the same sample condition; on the contrary, it must include the highest number of different sample conditions. A sample series with few experiments or with repetitions of the same experiment will not allow detecting coexpressed genes and even less to detect complex expression relationships. Note that de-noise, normalization, and similar procedures should be considered before using our tools.

The examples provided in the paper and supplementary materials use the data from AT_matrix. This matrix is the correlation between survival (A_matrix) and expression (T_matrix). A_matrix contains the growth inhibitory activities of 118 compounds tested on 60 tumour cell lines. This compound set includes most of the drugs currently in clinical use for tumour treatment. The microarray data (T_matrix) reflect the level of expression of 1376 genes, plus 40 individually assessed targets (proteins) and 40 other targets in the previous 60 tumour cell lines. AT_matrix links both matrices using the 60 tumour cell lines as a sample space to generate a correlation matrix of 1416 rows (genes and targets) by 118 columns (substances) [13]. These expression data were chosen because they cover a wide range of phenotypes shared by many different human tumour tissues.

2.2. PCOP. The mathematics behind this system uses the principal curves of oriented points (PCOP) calculation [10] to describe the expression relationships inner pattern. The principal curves is a nonlinear and nonvariable-dependent analysis technique, which is very suitable for the nonlinear analysis of expression relationships [14]. PCOP is defined by the generalization, at local level, of the following principal-component property: for a normal multivariable distribution X, if X is projected over a hyperplane, the total variance of the projection is minimised when the hyperplane is orthogonal to the first principal component. In this way, a principal oriented point (POP) is found for each local area, and the curve that goes throughout all of the POPs is the PCOP. The PCOP method provides the uncorrelation factor f. The f calculation considers not only the data dispersion around the curve, but also how well the curve pattern describes the morphology of the data cloud for any continuous curve type [10]. Thus, this f value provides an excellent measure for the nonlinear correlations.

2.3. Genes Coexpressed with a Pair of Genes with a Nonlinear Expression Relationship between Them. All the nonlinear expression relationships are detected for each gene of the expression array. These nonlinear expression relationships are classified by the type of curve. The *curvature points* of the PCOP are used to identify and classify the nonlinear expression relationships. *Curvature points* are those POPs in the PCOP in which a change in slope occurs. The detection of *curvature points* in expression relationships identifies the nonlinear expression relationships. The type of curve is described by the function of the curve: $y = e^x$, $y = -e^x$, $y = -\ln(x)$, $y = x^2$, $y = -x^2$, $y = x^3$, $1 = x^2 + y^2$, and so on. The genes coexpressed with each gene are also detected (none *curvature points* are detected in the PCOP of two coexpressed genes). This will allow us to study the nonlinear expression relationships between a user's gene of interest and different sets of coexpressed genes.

The correlation degree provided by the PCOP calculus is what guarantees us that the linear expression relationships (coexpressed genes) as well as the nonlinear expression relationships (intergroup expression relationships) are not a product of chance and have a biological meaning. For this reason, we require a high correlation degree for the linear expression relationships as well as for the nonlinear expression relationships. We are also restrictive in the classification of the expression relationships as linear expression relationships and the consequent consideration of two genes as coexpressed genes. Even a small curvature in the relationship of two coexpressed genes can cause a diversity in the typology of the expression relationships of these two genes with the genes of another set of coexpressed genes, more concretely, being A and B, two coexpressed genes whose linear expression relationship has a small curvature, and being C, a set of coexpressed genes that maintain nonlinear expression relationships with A and B. The expression relationships of gene A with set C may describe a different typology with respect to the expression relationships of gene B with set C.

2.4. Cliques of Nonlinear Expression Relationships between Genes. A clique in an undirected graph is a subset of its vertices such that every two vertices in the subset are connected by an edge. If we consider a graph of all the nonlinear expression relationships with a high correlation, we obtain its cliques. These cliques will not relate sets of coexpressed genes yet, but genes individually. Nevertheless, as the cliques are not relating pairs of genes but several genes in a network of nonlinear expression relationships, the cliques will be the seed to relate the sets of coexpressed genes between all of them. The genes of a clique must be at least three and they must maintain nonlinear expression relationships between all the genes of the clique.

2.5. Pairs of Isomorphic and Linear Cliques of Nonlinear Expression Relationships. Once the cliques are detected, they are grouped in pairs by relating genes that belong to the same set of coexpressed genes.

The cliques that will form each pair will meet two conditions.

(1) Each one of the genes of a clique will be coexpressed with a different gene of the other clique, forming pairs of coexpressed genes.

(2) The type of curve that relates two pairs of coexpressed genes will be the same in both cliques.

This provides us with pairs of cliques. Each gene of a clique will be coexpressed with a different gene of the other clique forming pairs of coexpressed genes. Then, the expression relationships that relate genes from different pairs of coexpressed genes will maintain the same type of curve for the two genes of the pair.

2.6. Cliques of Isomorphic and Linear Cliques of Nonlinear Expression Relationships among Genes. On the previous section we obtained the nonlinear expression relationships between pairs of coexpressed genes. Now, we obtain sets of coexpressed genes that maintain nonlinear expression relationships between them by grouping these pairs of coexpressed genes into sets of coexpressed genes. If we consider a graph where the vertices are the cliques of nonlinear relationships between genes and the edges link the pairs of linear isomorphic cliques, now we will calculate the cliques of this new graph obtaining the cliques of cliques. Thereby we obtain the *skeleton* of each set of coexpressed genes and the nonlinear expression relationships between the *skeletons*. The different networks among sets of coexpressed genes will be formed from the relationships between the *skeletons* of each set of coexpressed genes.

The genes of the *skeleton* of a set of coexpressed genes are those genes of the set that maintain a high-correlated nonlinear expression relationship with the *skeleton* of the other sets. The genes of the set of coexpressed genes that are not part of the *skeleton* will be coexpressed with the genes of the *skeleton*. These genes coexpressed with the *skeleton* will maintain the same type of nonlinear expression relationship with the other sets of coexpressed genes the closer they are to a $y = x$ relationship with respect to the genes of the *skeleton*. Deformations of this $y = x$ relationship between the genes of the set and its *skeleton* genes will produce a distortion of the expected curve. A higher variance in the coexpression with respect to the *skeleton* genes will also imply a higher variability in the expected type of curve.

The higher the number of genes of the *skeleton* is, the more representative the processes carried out by the coexpressed gene sets are. Thanks to the second condition of the linear-isomorphic-cliques definition we can make sure that the relationships between the genes of the *skeleton* of different sets of coexpressed genes maintain the same type of curve for all the genes of the *skeleton*. In the same way, we can also make sure that the genes coexpressed with the *skeleton* maintain expression relationships also of the same type (although it will depend on the correlation degree and how close to $y = x$ the coexpressed gene is).

The correlation degree to consider the expression relationships coexpressed enough will depend on the number of genes of the expression array. It is useful for small expression matrices, where nonlinear expression relationships between sets of coexpressed genes can be detected although these expression relationships have high entropy. The aim is to always detect enough nonlinear expression relationships to be able to find the *skeletons* that relate the sets of coexpressed genes.

The threshold to consider an expression relationship as linear or nonlinear will also depend on the number of genes, being (this threshold) more restrictive for the linear ones in matrices with less genes. Thus, large sets of coexpressed genes with very sharp curves between them can be formed for large expression matrices, whereas smaller sets of coexpressed genes, as well as more subtle nonlinear expression relationships between the sets, will be considered for small matrices.

The expression relationships have been filtered by the uncorrelation factor provided by the PCOP calculation to be considered correlated enough [10]. The threshold formula is

$$0.12 \times \frac{1600}{\text{num genes}} - \left(\frac{\text{num genes}}{40000}\right)^{18}. \quad (1)$$

The threshold formula for the curvature to consider whether the relationships are linear or nonlinear is

$$160 - \left(\frac{(15.0/20000) + (14.0/18400)}{2 \times \text{num genes}}\right). \quad (2)$$

However, a formula that depends only on the number of genes is not enough to guarantee the quality of the analysis extracted from the data because the data correspond to very different experiments with different nature. The diverse nature of the experiments is a qualitative variable that cannot be quantified with a formula. For this reason the correlation threshold calculation uses an online correction: from the relationships already analysed and the number of relationships pending to be analysed, the system makes an estimation of the number of relationships that would finally pass the threshold. From the calculated estimation, the system modifies automatically the threshold.

A higher number of genes in the expression array increase the number of coexpressed genes and nonlinear expression relationships, which facilitates finding *skeletons*. But in any case, the number of expression relationships with high correlation, as well as the number of linear expression relationships with respect to the nonlinear ones, will always depend on the nature of the experiments of the sample series.

3. Results and Discussion

The system allows to study sets of coexpressed genes that maintain nonlinear expression relationships among them, as well as to study the nonlinear expression relationships that a concrete gene of interest maintains with different sets of coexpressed genes. This can be studied for this target gene as well as for the genes coexpressed with it. There have been found 4573 nonlinear relationships and 20269 pairs of coexpressed genes (all highly correlated) from the microarray of 1416 genes used in the examples.

3.1. Searching for the Complex Expression Relationships between a Target Gene and Sets of Coexpressed Genes. The

study of the expression relationships between sets of coexpressed genes can start from the researcher's genes of interest. All the nonlinear expression relationships that a gene of interest maintains with different sets of coexpressed genes will be shown. These relationships will be shown classified by curve type, because each curve type implies a different activation/deactivation relationship. There will be shown only the nonlinear expression relationships that maintain a sufficient correlation degree.

We will study the activation and deactivation relationship between our gene of interest and different sets of coexpressed genes starting from these high-correlated nonlinear expression relationships. Two lists of coexpressed genes will be shown in a new view for each high-correlated nonlinear expression relationship of the gene of interest (Figure 1). The first list will show the genes coexpressed with the gene of interest. The second one will show the genes coexpressed with the gene that maintains the high-correlated expression relationship with the gene of interest. Then, the user can study the expression relationships between the two sets of coexpressed genes. The user can select genes from both lists of coexpressed genes to study in detail their expression relationship using a different interface [11, 12] (Figure 3). The first list of coexpressed genes is ordered by their correlation degree with the gene of interest and the second list is ordered by their correlation with respect to the gene nonlinearly related to the gene of interest.

An icon shows the type of nonlinear relationship between the two main genes and, by extension, between the two sets of coexpressed genes. The curve type is very important, since it determines the role of the genes in each expression dependence.

3.2. The Curve Type Indicates the Type of Activation and Deactivation Relationship between Sets of Coexpressed Genes. The system obtains the inner pattern of the curve for any type of expression relationship and classifies it. The only requirement is that the data cloud must be continuous. A $y = e^x$ relationship will provide an activation relationship between a set of genes and the other set; in other words, the first set of coexpressed genes must overexpress so the second set starts to express. In a $y = -e^x$ relationship, instead of an activation of the second set, there will be a deactivation. A $y = -\ln(x)$ relationship indicates a mutual-exclusion dependence; that is, one of the two sets of genes must be deactivated so the other set of coexpressed genes expresses. Note that these types of relationships are different from the positive and inverse coexpression relationships. This difference is precisely what allows us to detect different sets of coexpressed genes as well as the complex expression dependences between them. On the other hand, a $y = -e^x$ relationship, a $y = -x$ relationship, and a $y = -\ln(x)$ relationship explain completely different activation/deactivation dependences.

A $y = x^2$ relationship would indicate a deactivation of the second set of coexpressed genes by the overexpression as well as the underexpression of the first set. A $y = -x^2$ relationship would indicate an activation of the second set of genes, for the overexpression as well as the underexpression of the first set. Whereas in the relationships of type $|e^x|$, the overexpression of the set of coexpressed genes affects the other set. In the relationships of type $|x^2|$, the overexpression as well as the underexpression of the genes has an inhibitory or activatory effect on the other set of coexpressed genes.

Other relationships, such as those of type $y = x^3$ or $1 = x^2 + y^2$, will indicate other complex expression dependences between different sets of coexpressed genes. Real examples of each type of curve are shown in the supplementary material available at http://platypus.uab.es/nlnet.

As pointed out in the introduction, one of the three principles of our methodology is as follows. The type of curve between two genes is also maintained between the genes coexpressed with each one of them. Let us see an example: HLA genes are indicative of cell maturation marking the cell so it is recognised by the immune system [15]. HLA genes mark the cell making possible the inflammation of the tissue and the activation of the immune system. GRAMD1A is not a well-known membrane receptor that inhibits programmed cell death and it is linked to disease resistance [16]. HLA-F and GRAMD1A maintain a nonlinear expression relationship of type $y = e^x$ (Figure 3(a)). This points out that, possibly, the function associated with GRAMD1A can only be performed once the cell is marked by HLA genes. HLA-F and HLA-A are coexpressed genes (Figure 3(b)), and GRAMD1A is coexpressed with NREP (Figure 3(c)). Thus, HLA-A and NREP will also maintain a nonlinear expression relationship of type $y = e^x$ (Figure 3(d)). C5orf13 (NREP) expression is linked to hypertrophic scar [17]. This points out that hypertrophic scar and the function associated with NREP can only be performed once the cell is marked by HLA genes [18]. Even though the relationship of these genes with hypertrophic scar was already known [17, 18], there was no knowledge about how it was regulated.

Hypertrophic scarring (HS) is a result of increased fibrogenesis, which is thought to be caused by an exaggerated inflammatory response [19–21]. There is a clear association between specific HLA alleles and cutaneous fibrosis. Specific examples of cutaneous fibrosis include hypertrophic scars (HS) among others [18].

The relation of HLA and hypertrophic scar is already documented, but using our tool we found that this relation is mediated by NREP, because HLA genes must be overexpressed to activate NREP (a gene directly linked to HS [17]).

In this way, it is valuable that even though the technologies to obtain gene-expression arrays do not capture regulatory genes because of the low variability in their gene expression, these technologies do allow studying the regulation between processes through the genes that perform these processes (coexpressed genes that result from the activation cascade started by regulatory genes). This is because these final genes do maintain wide enough expression ranges, which allows our high-throughput tool to analyse the expression dependence between the sets of coexpressed genes.

3.3. Studying the Complex Expression Relationships between Sets of Coexpressed Genes. The different networks of nonlinear expression relationships among sets of coexpressed genes are classified by the number of sets and the curve types of the expression relationships between the sets. Once a

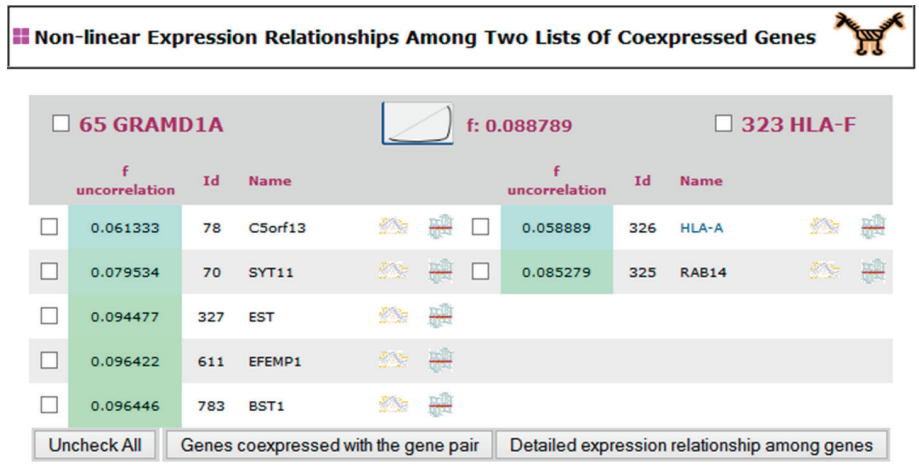


Figure 1: This view shows a nonlinear expression relationship where a researcher's gene of interest participates. The gene of interest is displayed on the top of the view on the left side. The column on the left side displays the genes coexpressed with the gene of interest, while the column on the right displays a set of coexpressed genes that maintain a nonlinear expression relationship with the gene of interest. The coexpressed genes are ordered by their correlation degree with their respective gene at the top (the f value obtained by the PCOP calculation). The icon shows the curve type of the nonlinear expression relationship. Each curve type implies a different expression dependence: mutual exclusion, trigger, double trigger, and so on. All the expression relationships relating genes from the two sets of coexpressed genes should be of the type shown by the icon. By selecting genes from the two sets, their expression relationship can be studied in detail in a new interface (Figure 3).

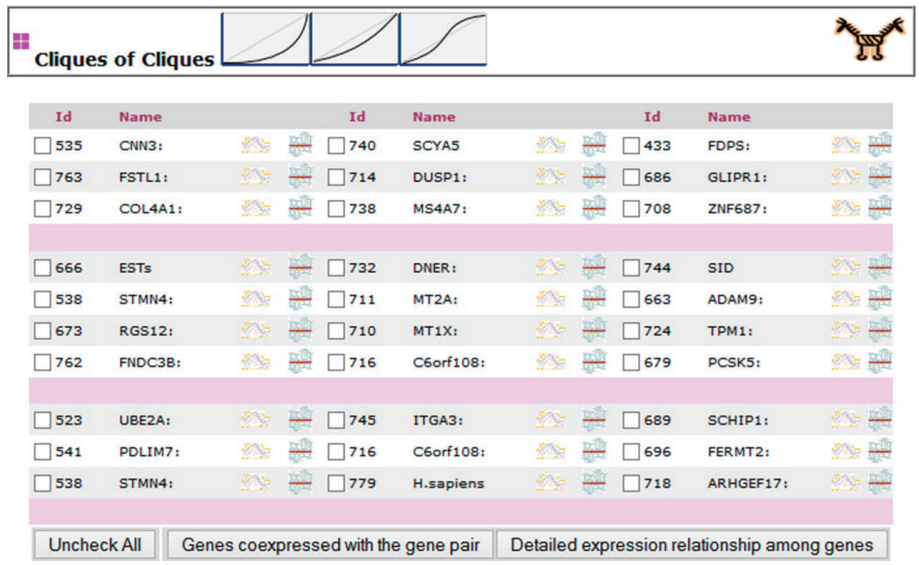


Figure 2: This view shows networks of concrete types of nonlinear expression relationships between sets of coexpressed genes. The icons at the top show the curve-type pattern of the networks listed. The networks will always form a complete graph. The pink line separates the networks found for the curve-type pattern. The columns contain the genes of the *skeleton* of each set of coexpressed genes. The genes of the different *skeletons* can be selected to study their expression relationship in detail [11, 12] (Figure 3). The genes of the different *skeletons* can be selected to study the expression relationships between the rest of coexpressed genes of the two sets, opening the view of Figure 1 for the two *skeleton* genes.

network type is selected, the networks found in the expression matrix that maintain this pattern in the intergroup expression relationships are displayed.

In the view that shows the networks (Figure 2), the genes that belong to the *skeleton* of each set of coexpressed genes are displayed in different columns. Each column displays the genes of the *skeleton* of each set of coexpressed genes. By selecting genes from the different *skeletons*, the expression relationships between them can be studied. Starting from the genes of the *skeleton*, the expression relationships between the rest of coexpressed genes of the sets can also be studied. By selecting one gene from the *skeleton* of two different sets of coexpressed genes, the genes coexpressed with each one of the genes of the two *skeletons* will be shown (in the way shown

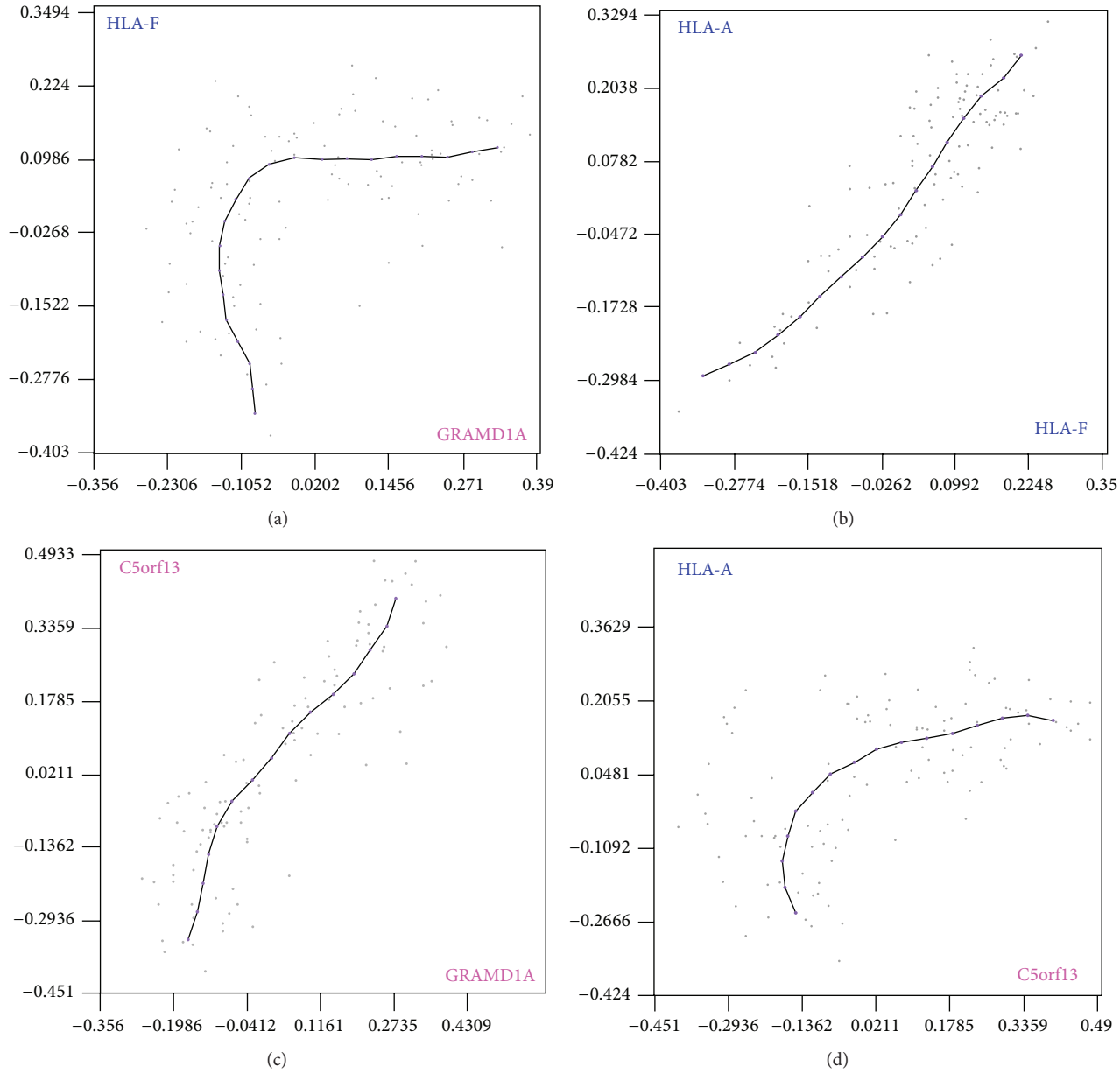


Figure 3: Four expression relationships are shown. The different sample conditions of the expression matrix (the sample series) constitute the data cloud. The PCOP describes the expression-relationship inner pattern. (b) and (c) show coexpressed genes. HLA-A and HLA-F are coexpressed genes (b) and GRAMD1A and NREP are also coexpressed genes (c). (a) and (d) show nonlinear expression relationships of $y = e^x$ type, an activation relationship. Since HLA-F and GRAMD1A maintain a nonlinear expression relationship of type $y = e^x$ (a), HLA-F is coexpressed with HLA-A (b), and GRAMD1A is coexpressed with NREP (c), therefore HLA-A and NREP (C5orf13) maintain a nonlinear expression relationship of the same type (d). HLA-A and GRAMD1A would also maintain a nonlinear relationship of type $y = e^x$, and HLA-F and NREP would maintain a nonlinear relationship of the same type. This is the key point of our approach: all the nonlinear expression relationships that relate genes from two sets of coexpressed genes will have the same type of curve.

in Figure 1). All of them should maintain the same type of nonlinear expression relationship between them. These listed genes can also be selected to study their nonlinear expression relationships in detail [11, 12]. In this way, it can be studied whether the genes coexpressed with the *skeleton* maintain the type of curve or whether it is distorted or lost. The genes coexpressed with the genes of the *skeleton* of the two sets appear ordered by their correlation degree with the gene of the *skeleton* of each set. The higher the correlation between a coexpressed gene and the gene of its *skeleton* is, the lower the variations of the curve types between this gene and the genes of the other set with respect to the curve type between the genes of both *skeletons* are.

In Figure 3 we can see the key point of our approach: all the nonlinear expression relationships that relate genes from two sets of coexpressed genes will describe the same

type of curve. In this way we can start the analyses from the relationships between all the pairs of genes of the microarray, but we construct the relationships between the sets of coexpressed genes from the *skeletons*. Since each set carries out a different cellular process, using our tool we can study the relationships between these independent cellular processes.

4. Conclusions

To respond to diverse and frequently changing conditions, cells must precisely mediate the synthesis and function of the proteins in the cell. This is controlled in part by the overall genomic expression program that results from the combined action of different regulatory factors, each of which responds to specific extra- and intracellular signals. These regulators govern the expression of sets of coexpressed genes that perform the appropriate cell functions. The variations in the expression of these coexpressed genes can be captured by high-throughput technologies to obtain gene expression arrays. In this way, the researcher is able to know which processes are carried out in the conditions he/she wishes to study, by knowing the different genes coexpressed in them. But what if the researcher wishes to know more? What if he/she wishes to know which relations have those different processes between them? In the case of working with large sample series, how do we know how these processes are activating or deactivating and activating again among them? If the researcher suspects that certain target genes can be a therapeutic target, how can he/she know the effect of their expression on the rest of the processes that this target gene does not belong, since it expresses with a different set of coexpressed genes? To know this could be implied from discovering unknown side effects to finding new ways to manipulate the expression of this gene.

In order to solve all these issues, we perform our high-throughput analysis. We obtain the coexpressed genes and the high-correlated nonlinear expression relationships and from them we obtain cliques (complete graphs) between coexpressed gene sets that maintain nonlinear expression relationships between them. In these networks, all the sets of coexpressed genes maintain a nonlinear expression relationship with each and every one of the other sets of coexpressed genes of the network. So, anytime, you know how to move from one process to another, passing by any other intermediate process. As a result, multiple networks are provided by a dynamic system that allows detecting sets of coexpressed genes related between them by complex activation dependences. The networks found will always depend on the analysed microarray. Large enough sample series and wide enough gene expression ranges facilitate the detection of coexpressed genes as well as the detection of nonlinear expression relationships between them. It is important to note that to detect a nonlinear expression dependence between two sets of coexpressed genes, this nonlinear expression dependence must exist, even though it affects only two small sets of coexpressed genes. In case these dependences exist, the developed tools allow to obtain very relevant information for the researcher since it makes possible to observe how the sets of coexpressed genes of his/her experiment interact between all of them. We think this approach could be a useful complement to other computational methods commonly used to analyse gene expression data.

As we present in the introduction, the expression dependences between sets of coexpressed genes, as well as between the processes these sets of coexpressed genes carry out, would never be linear. This is why new tools like the presented one are necessary.

Conflict of Interests

The authors declare no conflict of interests.

Acknowledgments

This research was supported by Grant nos. BFU2010-22209-C02-01 and BIO2013-48704-R from MCYT (Ministerio de Ciencia y Tecnología, Spain), from the Centre de Referència de R + D de Biotecnologia de la Generalitat de Catalunya, and by Comisión Coordinadora del Interior (Uruguay).

References

[1] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 25, pp. 14863–14868, 1998.

[2] T. Barrett, S. E. Wilhite, P. Ledoux et al., "NCBI GEO: archive for functional genomics data sets—update," *Nucleic Acids Research*, vol. 41, no. 1, pp. D991–D995, 2013.

[3] M. Huerta, M. Munyi, D. Expósito, E. Querol, and J. Cedano, "MGDB: crossing the marker genes of a user microarray with a database of public-microarrays marker genes," *Bioinformatics*, vol. 30, no. 12, pp. 1780–1781, 2014.

[4] S. van Dam, R. Cordeiro, T. Craig, J. van Dam, S. H. Wood, and J. P. de Magalhães, "GeneFriends: An online co-expression analysis tool to identify novel gene targets for aging and complex diseases," *BMC Genomics*, vol. 13, no. 1, article 535, 2012.

[5] J. M. Stuart, E. Segal, D. Koller, and S. K. Kim, "A gene-coexpression network for global discovery of conserved genetic modules," *Science*, vol. 302, no. 5643, pp. 249–255, 2003.

[6] S. Boltaña, S. Rey, N. Roher et al., "Behavioural fever is a synergic signal amplifying the innate immune response," *Proceedings of the Royal Society B: Biological Sciences*, vol. 280, no. 1766, 2013.

[7] A. Thalamuthu, I. Mukhopadhyay, X. Zheng, and G. C. Tseng, "Evaluation and comparison of gene clustering methods in microarray analysis," *Bioinformatics*, vol. 22, no. 19, pp. 2405–2412, 2006.

[8] N. R. Street, S. Jansson, and T. R. Hvidsten, "A systems biology model of the regulatory network in Populus leaves reveals interacting regulators and conserved regulation," *BMC Plant Biology*, vol. 11, article 13, 2011.

[9] X. Guo, Y. Zhang, W. Hu, H. Tan, and X. Wang, "Inferring nonlinear gene regulatory networks from gene expression data based on distance correlation," *PLoS ONE*, vol. 9, no. 2, Article ID e87446, 2014.

[10] P. Delicado and M. Huerta, "Principal curves of oriented points: theoretical and computational improvements," *Computational Statistics*, vol. 18, no. 2, pp. 293–315, 2003.

[11] J. Cedano, M. Huerta, and E. Querol, "NCR-PCOPGene: an exploratory tool for analysis of sample-classes effect on gene-expression relationships," *Advances in Bioinformatics*, vol. 2008, Article ID 789026, 2008.

[12] M. Huerta, J. Cedano, D. Peña, A. Rodriguez, and E. Querol, "PCOPGene-Net: holistic characterisation of cellular states from microarray data based on continuous and non-continuous analysis of gene-expression relationships," *BMC Bioinformatics*, vol. 10, article 138, 2009.

[13] U. Scherf, D. T. Ross, M. Waltham et al., "A gene expression database for the molecular pharmacology of cancer," *Nature Genetics*, vol. 24, no. 3, pp. 236–244, 2000.

[14] M. Huerta, J. Cedano, and E. Querol, "Analysis of nonlinear relations between expression profiles by the principal curves of oriented-points approach," *Journal of Bioinformatics and Computational Biology*, vol. 6, no. 2, pp. 367–386, 2008.

[15] M. Parkes, A. Cortes, D. A. van Heel, and M. A. Brown, "Genetic insights into common pathways and complex relationships among immune-mediated diseases," *Nature Reviews Genetics*, vol. 14, no. 9, pp. 661–673, 2013.

[16] S. Lorrain, B. Lin, M. C. Auriac et al., "Vascular Associated Death1, a novel GRAM domain-containing protein, is a regulator of cell death and defence responses in vascular tissues," *Plant Cell*, vol. 16, no. 8, pp. 2217–2232, 2004.

[17] J. Tan, X. Peng, G. Luo et al., "Investigating the role of P311 in the hypertrophic scar," *PLoS ONE*, vol. 5, no. 4, Article ID e9995, 2010.

[18] S. M. Mccarty, F. Syed, and A. Bayat, "Influence of the human leukocyte antigen complex on the development of cutaneous fibrosis: an immunogenetic perspective," *Acta Dermato-Venereologica*, vol. 90, no. 6, pp. 563–574, 2010.

[19] J. W. Lawrence, S. T. Mason, K. Schomer, and M. B. Klein, "Epidemiology and impact of scarring after burn injury: a systematic review of the literature," *Journal of Burn Care and Research*, vol. 33, no. 1, pp. 136–146, 2012.

[20] N. E. E. van Loey and M. J. M. van Son, "Psychopathology and psychological problems in patients with burn scars: epidemiology and management," *The American Journal of Clinical Dermatology*, vol. 4, no. 4, pp. 245–272, 2003.

[21] S. Liu, L. Jiang, H. Li et al., "Mesenchymal stem cells prevent hypertrophic scar formation via inflammatory regulation when undergoing apoptosis," *Journal of Investigative Dermatology*, 2014.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 563016, 6 pages
http://dx.doi.org/10.1155/2014/563016

Research Article

An Efficient Parallel Algorithm for Multiple Sequence Similarities Calculation Using a Low Complexity Method

Evandro A. Marucci,[1] Geraldo F. D. Zafalon,[1] Julio C. Momente,[1] Leandro A. Neves,[1] Carlo R. Valêncio,[1] Alex R. Pinto,[2] Adriano M. Cansian,[1] Rogeria C. G. de Souza,[1] Yang Shiyou,[3] and José M. Machado[1]

[1] *Department of Computer Science and Statistics, Sao Paulo State University, Rua Cristóvão Colombo 2265, 15054-000 São José do Rio Preto, SP, Brazil*
[2] *Department of Control Engineering and Automation, Federal University of Santa Catarina, Rua Pomerode 710, 89065-300 Blumenau, SC, Brazil*
[3] *College of Electrical Engineering, Zhejiang University, Hangzhou 310027, China*

Correspondence should be addressed to Geraldo F. D. Zafalon; zafalon@sjrp.unesp.br

Received 12 February 2014; Accepted 2 July 2014; Published 22 July 2014

Academic Editor: Tzong-Yi Lee

With the advance of genomic researches, the number of sequences involved in comparative methods has grown immensely. Among them, there are methods for similarities calculation, which are used by many bioinformatics applications. Due the huge amount of data, the union of low complexity methods with the use of parallel computing is becoming desirable. The *k-mers* counting is a very efficient method with good biological results. In this work, the development of a parallel algorithm for multiple sequence similarities calculation using the *k-mers* counting method is proposed. Tests show that the algorithm presents a very good scalability and a nearly linear speedup. For 14 nodes was obtained 12x speedup. This algorithm can be used in the parallelization of some multiple sequence alignment tools, such as MAFFT and MUSCLE.

1. Introduction

The use of sequence comparison methods has been remarkably growing in recent years, in response to the data expansion of genomic research. Consequently, methods that reduce the execution time are fundamental to the progress of this area. Many efforts have been made concerning their optimization [1, 2].

With the increase in number of multiple sequences alignments problems, the development of methods which have a lower computational complexity also increases. Nevertheless, just the creation of low complexity methods was not enough to work with high volume of data [3, 4]. The interest in parallel computing has grown and, hence, several parallel methods were created and embedded in many sequence comparison tools [5].

In general, the sequence comparison starts with the similarity calculation between pairs of sequences. It occurs through methods which require or not a prior alignment, varying the algorithm complexity and the level of biological accuracy. The alignment free methods correspond to a class of low complexity methods, which have a low temporal and spatial complexity in relation to the methods which requires a prior alignment. Moreover, they also keep a high level of biological accuracy for divergent sequences. For this reason, they are becoming increasingly important in computational biology.

The similarity calculation methods shall be classified into two main categories: methods based on the counting of words and methods that do not involve such a counting. Among the methods based on the counting of words there is the *k-mers* counting method, which was proposed by Katoh et al. [6]. This method counts the number of *k-mers* (words of size k) shared by a pair of sequences, using it as an approximation of the similarity level. It also uses a different alphabet, built

on the basis of statistical data, maximizing the biological accuracy in relation to the previously proposed methods.

The *k-mers* counting method was implemented in some important multiple sequences alignment tools, such as MAFFT [6, 7] and MUSCLE [1], and to the extent of our knowledge there is no other work in the literature that has developed and tested a parallel approach, specifically of this method. Once they are important tools, giving excellent results, in both biological accuracy and computational complexity [8], the development of a parallel algorithm for multiple sequence similarities calculation using the *k-mers* counting method is very useful. Although a parallel MUSCLE exists for shared memory systems [3], it does not include the parallelization of this stage. The parallel algorithm can be used in the development of any parallel tool that requires it in one of your stages. Tree construction or progressive multiple alignment tools, in general, are some examples.

This paper presents the *k-mers* counting method with the proposed similarity calculation parallel algorithm for multiple sequences through this method. The algorithm was developed for distributed memory parallel systems, using the library MPI [9].

2. Materials and Methods

2.1. Word Counting Methods. In general, word counting methods start with the mapping from sequences to vectors which store the length of each word. These words are subsequences of length k, also known as a k-tuple.

In order to understand the behavior of these methods, it is interesting to perform a review about some words statistical concepts. Then, consider a sequence X, of length n, defined as a segment of n symbols of a finite alphabet A, of length r.

A segment of k symbols, with $k \leq n$, is a k-tuple. The set W_k consists of all possible k-tuples from the alphabet set A and has N elements:

$$W_k = \{w_{k,1}, w_{k,2}, \ldots, w_{k,N}\}, \quad N = r^k. \tag{1}$$

Then, we count the number of k-tuples of W_k which appear in the sequence X. Computationally, this count is normally made moving a window of size k through the sequence, from the position 1 until the position $n - k + 1$. The vector c_k^X is responsible for the storage of the number of occurrences of k-tuples in the sequence X:

$$c_k^X = \left(c_{k,1}^X, c_{k,2}^X, \ldots, c_{k,N}^X\right). \tag{2}$$

A frequency vector f_k^X can then be gotten from the relative quantity of each k-tuple:

$$f_k^X = \frac{c_k^X}{\sum_{j=1}^{N} c_{k,j}^X} \equiv f_{k,i}^X = \frac{c_{k,i}^X}{n-k+1}. \tag{3}$$

As an example of the use of these structures, imagine a DNA sequence, where $A = \{A, T, C, G\}$ and $r = 4$. For $k = 3$, ATC and AAA are k-tuple belonging to the set W_3.

Table 1: Examples of compressed alphabets.

Alphabet (N)	Classes
SE-B (14)	A, C, D, EQ, FY, G, H, IV, KR, LM, N, P, ST, W
SE-B (10)	AST, C, DN, EQ, FY, G, HW, ILMV, KR, P
SE-V (10)	AST, C, DEN, FY, G, H, ILMV, KQR, P, W
Li-A (10)	AC, DE, FWY, G, HN, IV, KQR, LM, P, ST
Li-B (10)	AST, C, DEQ, FWY, G, HN, IV, KR, LM, P
Solis-D (10)	AM, C, DNS, EKQR, F, GP, HT, IV, LY, W
Solis-G (10)	AEFIKLMQRVW, C, D, G, H, N, P, S, T, Y
Murphy (10)	A, C, DENQ, FWY, G, H, ILMV, KR, P, ST
SE-B (8)	AST, C, DHN, EKQR, FWY, G, ILMV, P
SE-B (6)	AST, CP, DEHKNQR, FWY, G, ILMV
Dayhoff (6)	AGPST, C, DENQ, FWY, HKR, ILMV

For the sequence $X = ATATAC$, where $n = 6$, the counting and frequency vectors (c_3^X and f_3^X, resp.) are constructed as k-tuples of all W_3 which are identified in the sequence X. The sequence X is travelled in a window of size $k = 3$. The word within each window is compared with the words of W_3. In this case, $n - k + 1 = 4$ comparisons are necessary (ATA, TAT, ATA, TAC):

$$W_3 = \{ATA, TAT, TAC, AAA, \ldots\}, \quad c_3^X = (2, 1, 1, 0, \ldots), \quad f_3^X = (0.5, 0.25, 0.25, 0, \ldots). \tag{4}$$

2.2. The k-mers Counting Method. In the *k-mers* counting method, we use the term *k-mer* to represent the words, or k-tuples. This method presents a considerably greater speed in relation to conventional methods, which require alignment [10]. Its algorithm, implemented to determine the number of *k-mers* shared by two sequences, is $O(n)$, for sequences of size n. Differently, the conventional algorithms, which require alignment, are $O(n^2)$.

This algorithm uses, in general, a little different alphabet. In most cases, the alphabet used is a variation of the default alphabet. Known by compressed alphabets, these alphabets contain symbols that denote classes that correspond to two or more different types of residues (each residue is represented by a letter).

For amino acids sequences, a compressed alphabet C of size N is a partition of the default amino acids alphabet A, which contains 20 letters, in N disjoined classes containing similar amino acids. Table 1, extracted from [10], shows some examples of compressed alphabets.

With the use of compressed alphabets the identity is highly conserved. Pairs of related sequences have always a greater or equal identity and, therefore, more *k-mers* in common in an alphabet smaller than the default alphabet. An example of this characteristic can be seen in Table 2.

In Table 2, the upper and lower alignment are the same. The difference is that the first uses A, as default amino acids alphabet, while the second uses the compressed alphabet SE-V(10), whose members of the classes are represented by their

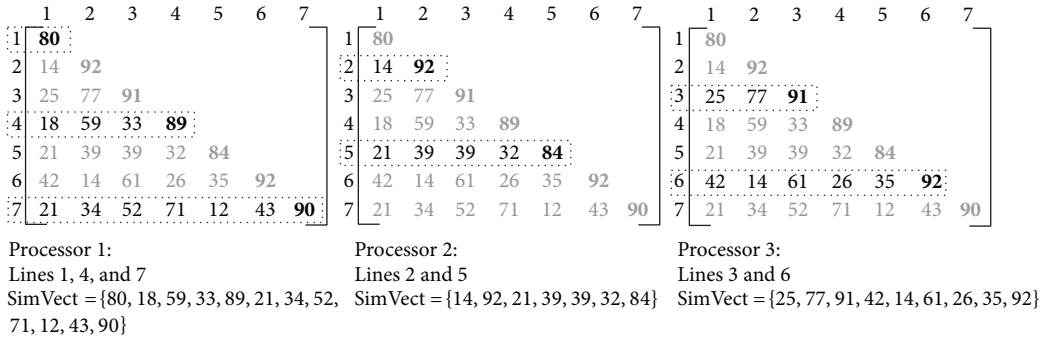


Figure 1: Example of how the similarities matrix calculation is distributed between the slaves.

Table 2: Comparison between the use of the alphabet A and the compressed alphabet SE-V (10).

Seq1: sAaNiLvGEnlvcKvaDFGLARl
Seq2: aArNiLvGEnyicKvaDFGLARl
Seq3: aArNvLiGEdnvaKicDFGLARv
Using the default amino acids alphabet
Seq1: AAaNILIGENlIcKIaDFGLARI
Seq2: AArNILIGENyIcKIaDFGLARI
Seq3: AArNILIGENnIaKIcDFGLARI
Using the compressed alphabet SE-V (10)
(each class member is represented by the first letter in alphabetical order)

first letters in alphabetical order—I, L, M, and V are shown as I, for example. The columns which are fully conserved are indicated with capital letters and with an asterisk below. The number of conserved columns ($k = 1$) increased from 12 in A to 19 in SE-V(10). For $k = 3$, the number of fully conserved *k-mers* increased from 4 in A to 10 in SE-V(10) and, for $k = 4$, from 3 to 8.

The choice of the alphabet and the value of k is based on statistics and has strong impact on the number of conserved identities. If the alphabet is selected, which means that there is a high probability of residues replacements of the same class and a low probability of residues replacements from distinct classes, then we probably have an increase in the number of identities detected. Moreover, the value of k confines this increasing in regions of continuous identity. As the sequences differ, the number of conserved *k-mers* is reduced, reaching a limit compared to the expected number of no related sequences. The use of compressed alphabets increases the likelihood of this limit be reached at a greater evolutionary distance. Subtle choices of the size of this alphabet and the value of k can provide a better measure of similarity.

The following equation shows how we calculate the similarity between sequences X and Y by the *k-mers* counting method:

$$F(X,Y) = \frac{\sum_{\tau} \min\left[n_X(\tau), n_Y(\tau)\right]}{\left[\min\left(L_X, L_Y\right) - k + 1\right]}. \quad (5)$$

Here τ is a *k-mer*, L_X and L_Y are the sequences lengths, and $n_X(\tau)$ and $n_Y(\tau)$ are the number of times τ appears in X and Y, respectively.

2.3. Multiple Sequence Similarities Calculation. In any application that performs a comparison between multiple sequences, either by multiple alignment or just by the construction of phylogenetic trees, we perform the similarity calculation in many independent sequences. Thus, the F value is obtained for all pairs of sequences involved in the processing. As $F(X, Y)$ equals $F(Y, X)$, this value is calculated once, by two nested loops, as

```
for (i = 1; i < num_seq; ++i)
for (j = 0; j < i; ++i)
M[i,j] = F(i,j)
```

All obtained values of F are, therefore, stored in a triangular matrix M.

2.4. Parallel Algorithm. In an application, the amount of sequences in comparison can be huge, making its implementation unfeasible in a single machine, even with the use of low computational complexity methods, as the *k-mers* counting method. For this reason, we propose a parallel algorithm that calculates the similarities of multiple pairs of sequences using the *k-mers* counting method.

This algorithm dynamically divides the computation among existing processors through a master-slave approach. This parallelism is performed distributing the similarity calculation of sequence pairs to available slaves. It is possible because each similarity pair calculation is independent of other calculations of similarity pair.

Initially, the master sends all sequences by broadcasting to all slaves. Using broadcast the master can reduce the overhead with the messages exchange, decreasing the communication time between processors.

The tasks distribution is based on the processor identifier, assigning to each slave the calculation of specific lines of the triangular matrix of similarities. The way this matrix is obtained is exemplified in Figure 1. Each slave initially calculates the corresponding line of the identifier, in a p-step loop, where p is the number of processors. In Figure 1, we have seven sequences and, hence, seven lines in the matrix. The first slave is responsible for the calculation of the lines 1,

	1	2	3	4	5	6	7
1	**80**						
2	14	**92**					
3	25	77	**91**				
4	18	59	33	**89**			
5	21	39	39	32	**84**		
6	42	14	61	26	35	**92**	
7	21	34	52	71	12	43	**90**

Figure 2: Similarities matrix that is built by the master processor.

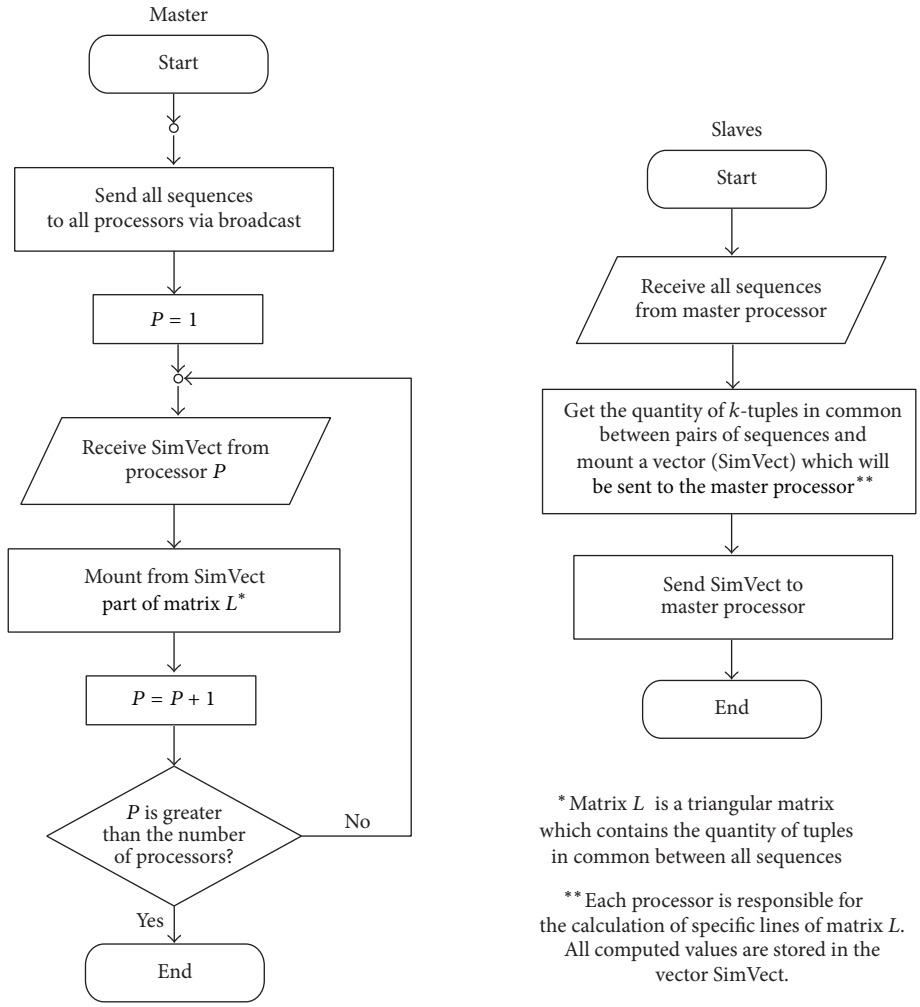


Figure 3: Flowchart of the parallel algorithm for multiple sequence similarities calculation.

4, and 7, the second slave by lines 2 and 5 and the third slave by lines 3 and 6. The gray values are the obtained results in other slaves which will only be joined in the master for the creation of the similarities matrix.

During the matrix lines calculation, each slave stores the results of all lines in a single vector (`SimVect`) which is sent to master. The master receives all slave vectors and, from these vectors and the identifier the slave has sent, it builds the similarities matrix. From the previous example, the obtained similarities matrix is shown in Figure 2.

In Figure 3 is showed the flowchart of the algorithm which performs the task in Figures 1 and 2.

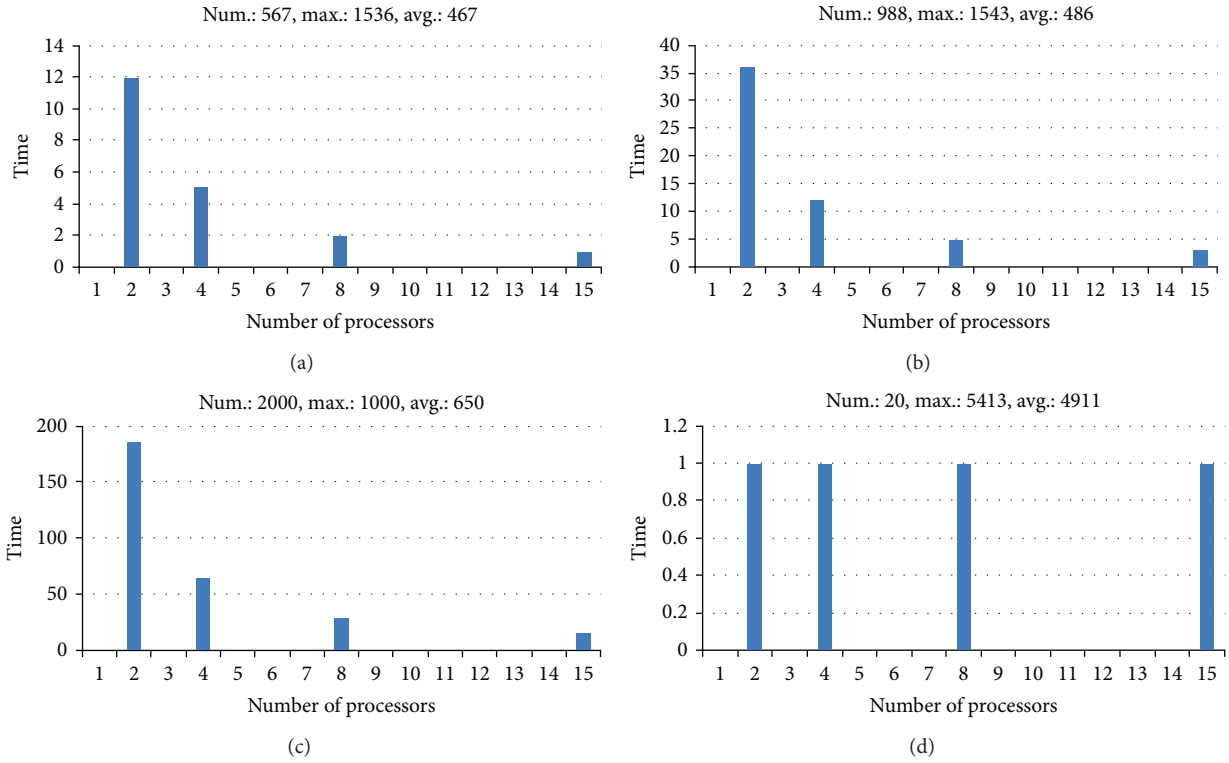


Figure 4: Processing time, in seconds, of the parallel algorithm for 2, 4, 8, and 15 nodes executing with four different datasets.

3. Results

3.1. Performance Evaluation. Many tests were performed to measure the performance of the proposed algorithms. The datasets used in these experiments have differences only on the number of sequences to be aligned and their lengths. The sequences used were extracted from the NCBI database (http://www.ncbi.nlm.nih.gov/). For each performed test, we describe specific information of the dataset used, as the number of sequences and the average and maximum length of the sequences.

Tests were performed on a Beowulf cluster running under Linux Debian. The Beowulf cluster consists of 15 nodes, each one composed by one AMD Athlon XP 2100+ processor with 1 GB of RAM memory. The nodes are connected with a dedicated 10/100 Fast Ethernet switch. The tests were performed with 1, 2, 4, 8, and 15 nodes. The run times were measured by executing them in stand alone mode, to ensure exclusive use of the communication and processors CPU and memory.

In order to verify the proposed algorithm performance, we executed tests with four different datasets. For each dataset, we verified the algorithm scalability when executed in a crescent machine number.

The first three graphics ((a), (b) and (c)) illustrated in Figure 4 show an almost linear speedup for tests with more than 500 sequences. Notice that the minimum system set for the parallel algorithm execution consists of two nodes, because such algorithm has a master-slave model. One machine (the master) is responsible for data management. The performance difference between the sequential algorithm and the parallel one, when run on a minimum system, is almost zero. For this reason, we only showed the tests performed with the parallel algorithm from the minimum system of two nodes until the maximum number of nodes in the cluster.

The graphic (d), also in Figure 4, illustrates also a constant performance, regardless of the number of machines, for an entry with few sequences. In this case, the sequential implementation of the algorithm is so fast for that input (approximately 1 second) that the speedup achieved with tasks division is close to the time spent with messages exchanges.

Comparing the minimum system (2 nodes) with the maximum system (15 nodes) we have an increase of 14 nodes and, approximately, a 12x speedup. Therefore, it can be noticed that the parallel algorithm presents excellent results with a nearly linear speedup.

4. Discussion

In this work, we presented a parallel strategy to calculate similarities between multiple pairs of sequences using the *k-mers* couting method, a low computational complexity method with good biological results. This calculation is used, for example, in multiple sequence alignment tools. The proposed parallel algorithm has been implemented for distributed memory systems, due to the wide use of Beowulf clusters in genomic research laboratories.

The tests performed show that the algorithm presents a good scalability and a nearly linear speedup. With the use of 14 processing nodes (slaves), the system achieved a 12x speedup. This speedup is justified by the total independence

of the scheduled tasks and by the good load balancing obtained with the triangular matrix lines distribution.

Additionally, the communication cost is minimized because the computed data in slaves are sent to master in a single message. Considering the 12x speedup achieved and using Amdahl's law [11] we can estimate that about 1.5% of the processes in the system are unparallelized, which reinforces the need of obtained optimization with the concentrated communication, avoiding message overload.

Conflict of Interests

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

Acknowledgments

The authors would like to thank all of their collaborators and institutions for the support to the development of the present work. This work was partially supported by the São Paulo Research Foundation (FAPESP, Brazil) under Grant no. 06/59592-0.

References

[1] R. C. Edgar, "MUSCLE: a multiple sequence alignment method with reduced time and space complexity," *BMC Bioinformatics*, vol. 5, article 113, 2004.

[2] G. F. Zafalon, E. A. Marucci, J. C. Momente, J. R. Amazonas, L. M. Sato, and J. M. Machado, "Improvements in the score matrix calculation method using parallel score estimating algorithm," *Journal of Biophysical Chemistry*, vol. 4, no. 2, pp. 47–51, 2013.

[3] X. Deng, E. Li, J. Shan, and W. Chen, "Parallel implementation and performance characterization of muscle," in *Proceedings of the 20th IEEE Parallel and Distributed Processing Symposium (IPDPS '06)*, 2006.

[4] H. A. Arcuri, G. F. D. Zafalon, E. A. Marucci et al., "SKPDB: a structural database of shikimate pathway enzymes," *BMC Bioinformatics*, vol. 11, article 12, pp. 1–7, 2010.

[5] E. A. Marucci, G. F. Zafalon, J. C. Momente et al., "Using threads to overcome synchronization delays in parallel multiple progressive alignment algorithms," *American Journal of Bioinformatics*, vol. 1, no. 1, pp. 50–63, 2012.

[6] K. Katoh, K. Misawa, K. Kuma, and T. Miyata, "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform," *Nucleic Acids Research*, vol. 30, no. 14, pp. 3059–3066, 2002.

[7] K. Katoh and H. Toh, "Parallelization of the MAFFT multiple sequence alignment program," *Bioinformatics*, vol. 26, no. 15, Article ID btq224, pp. 1899–1900, 2010.

[8] R. C. Edgar and S. Batzoglou, "Multiple sequence alignment," *Current Opinion in Structural Biology*, vol. 16, no. 3, pp. 368–373, 2006.

[9] M. J. Quinn, *Parallel Programming in C with MPI and OpenMP*, McGraw-Hill Education, New York, NY, USA, 1st edition, 2003.

[10] R. C. Edgar, "Local homology recognition and distance measures in linear time using compressed amino acid alphabets," *Nucleic Acids Research*, vol. 32, no. 1, pp. 380–385, 2004.

[11] G. M. Amdahl, "Validity of the single processor approach to achieving large scale computing capabilities," in *Proceedings of the Spring Joint Computer Conference (AFIPS '67)*, 1967.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 891725, 9 pages
http://dx.doi.org/10.1155/2014/891725

Research Article

The Definition of a Prolonged Intensive Care Unit Stay for Spontaneous Intracerebral Hemorrhage Patients: An Application with National Health Insurance Research Database

Chien-Lung Chan,[1] Hsien-Wei Ting,[1,2] and Hsin-Tsung Huang[3]

[1] *Department of Information Management, Yuan Ze University, 135 Yuan-Tung Road, Chungli, Taoyuan 32003, Taiwan*
[2] *Department of Neurosurgery, Taipei Hospital, Ministry of Health and Welfare, No. 127, Su-Yuan Road, Hsin-Chuang District, New Taipei City 24213, Taiwan*
[3] *Medical Affairs Division, National Health Insurance Administration, Ministry of Health and Welfare, No. 140, Section 3, Hsin-Yi Road, Taipei 10634, Taiwan*

Correspondence should be addressed to Hsien-Wei Ting; ting.ns@gmail.com

Received 11 April 2014; Revised 18 June 2014; Accepted 24 June 2014; Published 14 July 2014

Academic Editor: Tzong-Yi Lee

Introduction. Length of stay (LOS) in the intensive care unit (ICU) of spontaneous intracerebral hemorrhage (sICH) patients is one of the most important issues. The disease severity, psychosocial factors, and institutional factors will influence the length of ICU stay. This study is used in the Taiwan National Health Insurance Research Database (NHIRD) to define the threshold of a prolonged ICU stay in sICH patients. *Methods.* This research collected the demographic data of sICH patients in the NHIRD from 2005 to 2009. The threshold of prolonged ICU stay was calculated using change point analysis. *Results.* There were 1599 sICH patients included. A prolonged ICU stay was defined as being equal to or longer than 10 days. There were 436 prolonged ICU stay cases and 1163 nonprolonged cases. *Conclusion.* This study showed that the threshold of a prolonged ICU stay is a good indicator of hospital utilization in ICH patients. Different hospitals have their own different care strategies that can be identified with a prolonged ICU stay. This indicator can be improved using quality control methods such as complications prevention and efficiency of ICU bed management. Patients' stay in ICUs and in hospitals will be shorter if integrated care systems are established.

1. Introduction

Length of stay (LOS) in the intensive care unit (ICU) is one of the most important factors that influence health management. There are several factors that influence the ICU LOS: medical severity factors, psychosocial factors, and institutional factors [1]. Some studies have treated LOS as a hospital cost [2]. There are also many methods and management strategies that can influence the LOS. A higher nurse-patient ratio results in a lower hospital LOS [3]. The distributions of LOS analyzed by disease or institute are frequently skewed to the right. The mean, median, and range require more in-depth analysis for LOS studies [4]. However, the threshold of a prolonged LOS is useful for managers for analysis of the quality of care and hospital costs for diseases or institutes. The definition of a prolonged ICU stay varies by hospital type, ICU type, and also different diseases [5–10]. The severity of stroke, medical complications, and degree of disability are factors that influence the length of stay of patients [11–15].

Spontaneous intracerebral hemorrhage (sICH) is one of the most important diseases. The incidence of sICH varies by sex, age, and ethnic group [16–22]. The annual mortality rates of sICH vary in different countries but overall are around 50% [20, 23–25]. It represents around 10~35% of stroke patients and causes a higher mortality and morbidity than other strokes and costs much more in terms of medical facilities [16, 22, 26–29]. Koton et al. [30] constructed a prolonged length of stay score. They defined a prolonged LOS of stroke patients in acute wards as 7 days and suggested

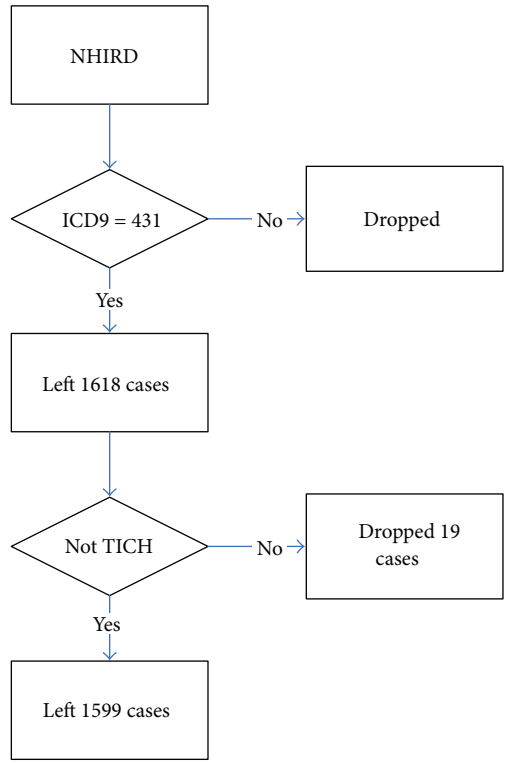


Figure 1: Flow chart of data management of the NHIRD.

that the severity of stroke, comorbidities, prior disabilities, and sICH are all important factors affecting the length of stay in stroke patients. Luker et al. [15] also concluded that stroke severity, previous independence, comorbidities, day of admission, stroke unit admission, and LOS are outcome factors for stroke patients. Actually, sICH and LOS are part of the outcome evaluation factors. According to previous research, the quality of intensive care can be a good indicator of the outcome of sICH. However, few studies have discussed the definition of a prolonged ICU stay in sICH patients.

Most sICH patients will be admitted to ICU. However, the resources of ICUs are limited. These patients have one of the most important diseases and will utilize the ICU frequently. Therefore, evaluation of ICU resource allocation of sICH patients is very important. This study analyzed the Taiwan National Health Insurance Research Database (NHIRD) to define the threshold of a prolonged ICU stay in sICH patients. Other than the severity of the disease, the causes of a prolonged ICU stay arise from the management methods of hospitals or institutes. The result will be a good indicator of facilities allocation and outcome prediction for sICH patients.

2. Method

This research analyzed the Longitudinal Health Insurance Database (LHID2005) in the NHIRD from 2005 to 2009. A dataset of one million subjects was constructed from all subjects with National Health Insurance in 2005 by random sampling in Taiwan. The dataset used in this study contained outpatients and inpatients claims data with details recorded for each visit/stay, and the registry file for beneficiaries was processed to identify the demographic data from 2001 to 2009 [31]. Because the sampled one million subjects were patients in 2005, the subjects were definitely alive before 2005. It would severely increase the mortality rate if subjects prior to 2005 were included. Therefore, this study only used the NHIRD from 2005 to 2009 as the analysis base to prevent bias of mortality.

The inclusion criteria of patients in this study were patients with a first attack of sICH whose diagnosis ICD9 code was 431. All patient data were collected at admission for a first sICH attack. There were a total of 1618 sICH cases included. Patients who were admitted due to traumatic intracranial hemorrhage (TICH) whose diagnosis ICD9 code was from 800 to 804.99, from 850 to 854.19, 959.01, and 959.09 were all excluded. There were 19 cases dropped owing to the exclusion criteria, and 1599 cases remained in total. The process of data management is shown in Figure 1. The admission cost, ICU admission and discharge dates, admission and discharge dates, and major admission comorbidities were all recorded. Multiple comorbidities were defined as patients who had 2 or more than 2 kinds of major admission comorbidities. Major admission comorbidities were defined as the first 5 major comorbidities collected in the NHIRD.

This study used change point analysis to find the threshold of a prolonged ICU stay in sICH patients. Change point analysis originated from studies of quality control [32, 33]. It can be used to solve problems of thresholds, identify the existence of any change point, and find the change point

if there is one [34]. This method is used in many fields, including thresholds of ecological and biological processes, software reliability estimations, financial problems, changes of control charts and the status of manufacturing, finding trends and growth rate functions, flood season segmentation, and also medical problems [8, 35–38]. This study used the software combination of cumulative sum charts (CUSUM) and bootstrapping to detect the definition of the threshold for a prolonged ICU stay [39]. We usually construct a CUSUM chart and find the cumulative sum of the differences between individual data values and the mean. If there is no shift in the mean of the data, the chart will be relatively flat with no pronounced changes in slope. The range will also be small. A dataset with a shift in the mean will have a slope change at the data point at which the change occurred, and the range will be relatively large [40, 41].

After the threshold of a prolonged ICU stay was defined, the percentages of prolonged ICU stays for sICH patients in hospitals with different training capacities, those with different ownerships, and those in different regions were also compared. Training hospitals were divided into medical centers, regional hospitals, and local hospitals. Medical centers performed the most staff training. Local hospitals provide medical care for the local area and do not have a great training burden. Hospital ownership was divided into government hospitals, public medical school hospitals, military hospitals, veterans hospitals, religious hospitals, private medical school hospitals, and private hospitals. There were 6 divisions: the Taipei division, Northern division, Central division, Southern division, Kaoping division, and Eastern division. The Taipei division is the area of the capital of Taiwan. The Eastern division is a rural area in Taiwan.

Student's t-test was used for the continuous data. The χ^2 test was used for the categorical data. Kaplan-Meier survival analysis was used for the survival analysis of sICH patients. Student's t-test, χ^2 test, and Kaplan-Meier survival analysis were calculated using SPSS version 12.0 (SPSS Inc., Chicago, IL, USA). The change points were analyzed using Change-Point Analyzer version 2.3 (Taylor Enterprises, Inc., Libertyville, IL, USA). Statistical significance was defined as $P < 0.05$.

3. Results

With the cumulative sum control chart (CUSUM chart) of days standard deviation, standard deviation changes were found for the 11th day and the 23rd day (Figures 2 and 3 and Table 1). A prolonged ICU stay of sICH patients was defined as more than 10 days (Figure 4). According to this definition, 436 cases (27.3%) had a prolonged ICU stay and the remaining 1163 cases (72.7%) did not have a prolonged ICU stay. There are 36.5% of the patients who were female. There was no significant difference in the gender ratio between the prolonged and nonprolonged ICU stay patients. The mean age of the sICH patients was 62.8 years (SD = 15.0). The prolonged ICU stay patients (64.7 years, SD = 14.0) were significantly older than the nonprolonged ICU stay patients (62.1 years, SD = 15.3) ($P < 0.01$). The mean admission

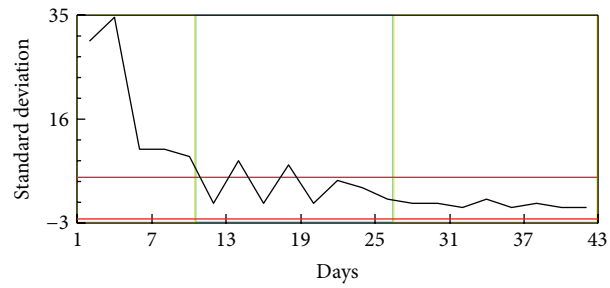


Figure 2: Plot of days against standard deviation.

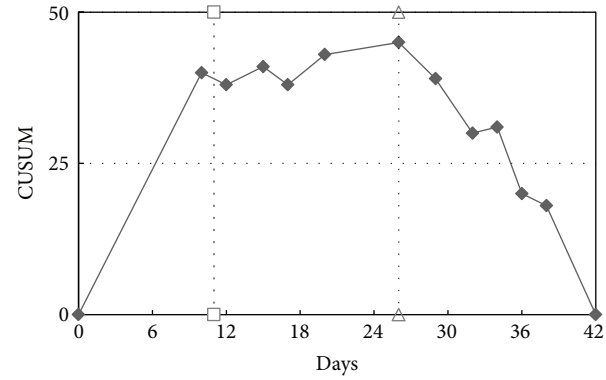


Figure 3: CUSUM (cumulative sum control) chart of days standard deviation.

Table 1: Significant change points of days standard deviations.

Row	Confidence interval	Confidence level	From	To
11	(11, 15)	94%	15.725	3.6693
23	(23, 31)	94%	3.6693	0.52418

Confidence level for candidate changes = 50%; confidence level for inclusion in table = 90%; confidence interval = 95%; bootstraps = 1000.

LOS and ICU stay were 16.7 (SD = 12.0) and 7.8 (SD = 7.7), respectively. Both the hospital LOS (27.1 days, SD = 11.1) and the ICU stay (18.2 days, SD = 7.2) of the prolonged ICU stay patients were longer than those of the nonprolonged ICU stay patients (13.1 days, SD = 10.3 and 3.9 days; SD = 2.5) ($P < 0.001$). The surgical intervention rate for the sICH patients was 25.5%, and the rate for the prolonged ICU stay patients (67.6%) was higher than that for the nonprolonged ICU stay patients (25.3%). Although the patient-day cost ratio of the nonprolonged ICU stay patients (56.4%) was higher than that of the prolonged ICU stay patients (43.6%) ($P < 0.001$), the ICU patient-day cost ratio of the prolonged ICU stay patients (63.3%) was higher than that of the nonprolonged ICU stay patients (36.7%) (Table 2).

The hospital cost of the prolonged ICU stay cases (US$11,036, SD = 4808) was significantly higher than that of the nonprolonged ICU stay cases (US$3,155, SD = 2510) ($P < 0.001$). If the total hospital fees cost is discussed, the prolonged ICU stay patients represented only 436 cases (27.3%) but cost more than half (56.7%) of the sICH patients'

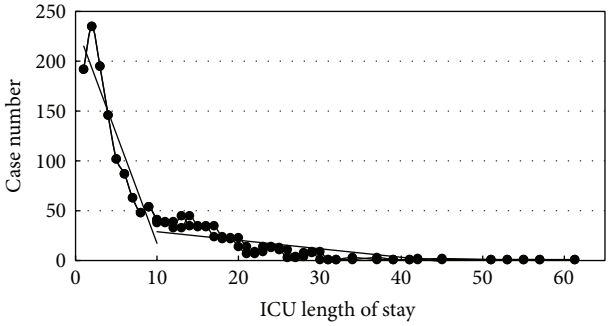


Figure 4: Case distribution of ICU days and change point analysis results.

hospital care. The highest fees cost is the room fee, which represents around 34.5%, followed by medications fees (14.3%), surgical fees (13.1%), diagnostic fees (8.5%), doctor care fees (5.4%), and other treatment fees (24.2%). The percentage of diagnostic fees of the nonprolonged ICU stay cases (10.6%) was higher than that of the prolonged ICU stay cases (7.0%) ($P < 0.01$). The percentage of medication fees of the nonprolonged ICU stay cases (12.6%) was lower than that of the prolonged ICU stay cases (15.6%) ($P < 0.05$). There were no significant differences in the percentages of the other types of cost between the prolonged and nonprolonged ICU stay patients (Table 2).

The mean survival time (months) of the sICH patients was 36.0 months (95% CI = 34.5~37.5). That of the nonprolonged ICU stay patients (38.0, 95% CI = 36.2~39.7) was longer than that of the prolonged ICU stay patients (31.0, 95% CI = 28.2~33.7) ($P < 0.01$). The total mortality rate of the sICH patients was 41.8%. The mortality rate of the prolonged ICU stay cases (50.7%) was significantly higher than that of the nonprolonged ICU stay cases (38.4%), with the odds ratio being 1.674 (95% CI = 1.342~2.087) ($P < 0.001$). According to Figure 3, the patients with a nonprolonged ICU stay will die very quickly in the start time and slow down after surviving for more than one month. The survival lines of the prolonged ICU stay and nonprolonged ICU stay patients crossed at the 5th month. Finally, the patients with a nonprolonged ICU stay had a higher survival rate than those with a prolonged ICU stay (Figure 5).

This study also studied the major admission comorbidities of sICH patients. 30.6% of the sICH patients had multiple comorbidities. The patients with a prolonged ICU stay (39.2%) had a higher percentage of multiple comorbidities than the patients with a nonprolonged ICU stay (27.3%), with the odds ratio being 1.715 (95% CI = 1.360~2.161) ($P < 0.001$). The most common comorbidity of sICH patients was found to be hypertension (62.0%). The patients with a prolonged ICU stay (54.1%) had a lower percentage of having hypertension than the patients with a nonprolonged ICU stay (65.0%), with the odds ratio being 0.635 (95% CI = 0.508~0.794) ($P < 0.001$). The subsequent most common diseases were pulmonary diseases (18.2%). The patients with a prolonged ICU stay (37.6%) were at higher risk of having pulmonary diseases than the patients with a nonprolonged ICU stay (10.9%), with the odds ratio being 4.918 (95% CI = 3.764~6.426). 17.3% sICH patients had diabetes mellitus (DM). There was no significant difference in the percentage of DM between sICH patients with a prolonged and a nonprolonged ICU stay. There were 15.1% sICH patients with hydrocephalus. A higher percentage of patients with a prolonged ICU stay (28.9%) had hydrocephalus than patients with a nonprolonged ICU stay (10.0%), with the odds ratio being 3.699 (95% CI = 2.792~4.902). The incidences of the other common comorbidities in sICH patients were an old CVA (5.6%), heart diseases (3.8%), renal diseases (3.4%), liver diseases (2.6%), peptic ulcer (1.8%), and dementia (1.3%). There were no significant differences in the percentages of sICH patients with a prolonged and nonprolonged ICU stay with these comorbidities (Table 3).

This study evaluated the differences between hospitals of different classifications, including training capacity, ownership of the hospital, and region of the hospital. There were no significant differences in the female ratio and mortality within 30 days between hospitals, no matter what the classification. According to the classifications of training hospitals, most ICH patients stay in regional hospitals (881 cases); the rest stay in medical centers (571 cases) and local hospitals (147 cases). There were no significant differences in the surgical intervention ratio, mortality ratio, and prolonged ICU stay ratio among medical centers, regional hospitals, and local hospitals. However, the mean age of patients of medical centers (61.4 years, SD = 15.3) was significantly younger than those of regional (63.3 years, SD = 14.8) and local hospitals (64.8 years, SD = 14.3) ($P < 0.05$). Local hospitals (9.3 days, SD = 9.6) had the longest mean ICU days and regional hospitals (7.4 days, SD = 7.1) had the shortest mean ICU days ($P < 0.05$). Medical centers (US$6327, SD = 5582) had the highest hospital expenditure as compared with regional (US$4,701, SD = 4161) and local hospitals (US$4,942, SD = 4,137) ($P < 0.001$) (Table 4).

According to hospital ownership, most patients stay in private hospitals: private hospitals (1018 cases), private medical school hospitals (142 cases), and religious hospitals (56 cases). There were no significant differences in the mean age and mortality rates between hospitals of different ownerships. Private hospitals had a significantly lower ratio of prolonged ICU stays than public hospitals ($P < 0.001$). The veterans hospitals had the highest hospital expenditure (US$8,979, SD = 7,698) as compared with the other hospitals, and the rest are military hospitals (US$6,629, SD = 6,040). The religious hospitals (US$4,245, SD = 3,755) had the lowest hospital expenditure of all the hospitals ($P < 0.001$). The surgical intervention ratios of the veterans hospitals (50.0%) were the highest among other hospitals, and the public medical school hospitals (19.7%) had the lowest ratio of surgical intervention ($P < 0.001$). Both military hospitals (20.0 days, SD = 13.2 and 9.5 days; SD = 9.1) and veterans hospitals (21.7 days, SD = 14.1 and 11.8 days; SD = 11.2) had the highest hospital admission days and ICU days ($P < 0.001$) (Table 4).

According to the regions of hospitals, more than 1/4 of all the cases were in the Taipei division (409 cases). There were no significantly different ratios of surgical interventions and mortality rates among the divisions. The patients in the Taipei

Table 2: Demographic data of sICH patients from 2005 to 2009.

	Nonprolonged ICU stay (SD)	Prolonged ICU stay (SD)	Total (SD)
Case number (*n*)	1163	436	1599
Female case percentage	36.8%	35.6%	36.5%
Mean age**	62.1 (15.3)	64.7 (14.0)	62.8 (15.0)
Mean admission days***	13.1 (10.3)	27.1 (11.1)	16.7 (12.0)
Mean ICU days***	3.9 (2.5)	18.2 (7.2)	7.8 (7.7)
Surgical intervention ratio***	25.5%	67.7%	37.0%
Patient-day spend ratio***	56.4%	43.6%	100%
ICU patient-day cost ratio***	36.7%	63.3%	100%
Total cost percentage	43.3%	56.7%	100%
†Total hospital fees***	3,155 (2510)	11,036 (4,808)	5304 (4,817)
Hospital room fees ratio	34.2%	34.7%	34.5%
Medications fees ratio*	12.6%	15.6%	14.3%
Surgical fees ratio	13.7%	12.6%	13.1%
Diagnostic fees ratio**	10.4%	7.0%	8.5%
Doctor care fees ratio	6.4%	4.7%	5.4%

*$P < 0.05$, **$P < 0.01$, and ***$P < 0.001$; †unit is US dollars.

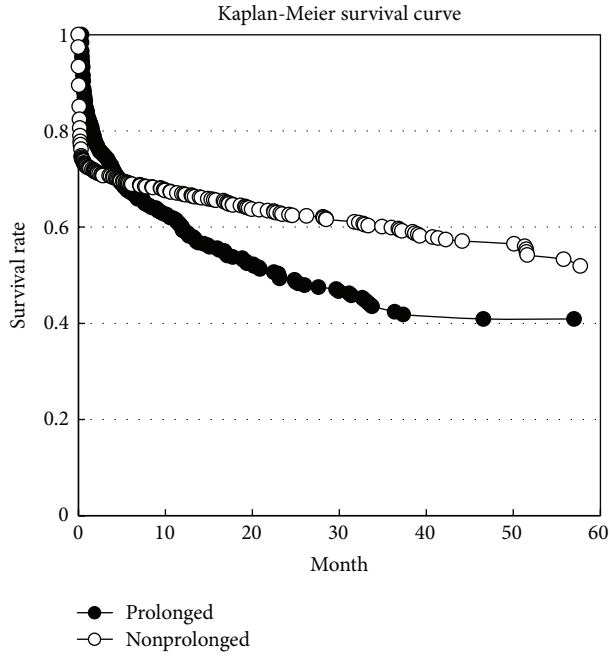


Figure 5: Survival rate curves of nonprolonged ICU stay patients and prolonged ICU stay patients ($P < 0.001$). The mortality rates of the patients with a prolonged/nonprolonged ICU stay were 50.7%/38.4%, respectively. The overall mortality rate was 41.8%.

division had a higher ratio of a prolonged ICU stay than the patients in the other divisions ($P < 0.01$). The hospital admission days (20.6 days, SD = 14.2) and ICU days (9.0 days, SD = 9.2) of the Taipei division were both higher than those of the other divisions ($P < 0.001$). The Taipei division (US$6,646, SD = 5,940) and Northern division (US$5,522, SD = 4,527) had the highest hospital expenditures as compared with the other divisions ($P < 0.001$) (Table 4).

4. Discussion

The Taiwan National Health Research Institute (NHRI) constructed the NHIRD, a nationwide research database, for medical research purposes [31]. Many researchers have used the NHIRD to study medical and epidemiological issues [42, 43]. Some researchers have also completed medical cost-related research using the NHIRD [42, 44]. The Cochrane systemic review found that some studies have treated LOS as a financial factor, which can be considered a surrogate for hospital cost [2]. The hospital LOS and the ICU stay will vary under different care policies in different countries and hospitals [1]. The care policies for sICH patients in different countries or regions differ, and ICH patients receive a high quality of care and incur a lower cost as they are covered by the Taiwan National Health Insurance [22]. A prolonged length of stay is defined as 7 days for stroke patients. One

Table 3: Comorbidity influence of patients with a prolonged ICU stay and those with a nonprolonged ICU stay.

	Nonprolonged (n = 1163)	Prolonged (n = 436)	All (n = 1599)	Odds ratio
Multiple comorbidities***	27.3%	39.2%	30.6%	1.715 (1.360~2.161)
Hypertension***	65.0%	54.1%	62.0%	0.635 (0.508~0.794)
Pulmonary diseases***	10.9%	37.6%	18.2%	4.918 (3.764~6.426)
Diabetes mellitus	16.9%	18.3%	17.3%	NP
Hydrocephalus***	10.0%	28.9%	15.1%	3.669 (2.766~4.865)
Old CVA	5.3%	6.4%	5.6%	NP
Heart diseases	4.3%	2.5%	3.8%	NP
Renal disease	2.9%	4.8%	3.4%	NP
Liver disease	3.0%	1.6%	2.6%	NP
Peptic ulcer	1.8%	1.8%	1.8%	NP
Dementia	1.5%	0.7%	1.3%	NP

* $P < 0.05$, ** $P < 0.01$, and *** $P < 0.001$.
DM: diabetes mellitus; CVA: old CVA that was not sICH; PU: peptic ulcer.

Table 4: Prolonged ICU stay ratios and other profile comparisons in different types of hospitals.

	Case number (n)	Surgical intervention ratio	Mortality within 30 days	Prolonged ICU stay ratio	†Total hospital fees
All cases	1599	37.00%	25.10%	27.30%	5304 (4817)
Training hospital (P value)		n.s.	n.s.	n.s.	0.000
Medical center	571	41.30%	25.20%	29.10%	6327 (5582)
Regional hospital	881	34.50%	24.00%	25.40%	4701 (4161)
Local hospital	147	34.70%	32.00%	31.30%	4942 (4137)
Hospital ownership (P value)		0.006	n.s.	0.002	0.000
Government hospital	180	30.60%	26.70%	24.40%	4772 (4379)
Public medical school hospital	61	19.70%	27.90%	31.10%	5213 (4160)
Military hospital	66	34.80%	19.70%	36.40%	6629 (6040)
Veterans hospital	76	50.00%	23.70%	46.10%	8979 (7698)
Religious hospital	56	37.50%	21.40%	21.40%	4245 (3755)
Private medical school hospital	142	41.50%	23.90%	29.60%	5543 (4558)
Private hospital	1018	37.60%	25.50%	25.50%	5068 (4518)
Hospital regions (P value)		n.s.	n.s.	0.005	0.000
Taipei division	409	36.20%	23.00%	32.00%	6466 (5940)
Northern division	212	42.50%	29.70%	27.80%	5522 (4527)
Central division	369	39.00%	21.70%	29.50%	5346 (4498)
Southern division	258	32.60%	26.70%	20.50%	4117 (3679)
Kaoping division	289	37.40%	26.00%	26.00%	4678 (4189)
Eastern division	62	27.40%	33.90%	14.50%	4495 (4605)

n.s.: nonsignificant; * $P < 0.05$, ** $P < 0.01$, and *** $P < 0.001$; † unit is US dollars.

type of stroke patients is those who are classified as sICH patients [30]; however, there is no definition for sICH patients. Actually, sICH patients have a more severe illness, and these patients will have a longer ICU stay and will incur a greater hospital expenditure. This study found the definition of a prolonged ICU stay using the national health insurance database of Taiwan. This definition can be a good indicator for hospital management and resources evaluation for sICH patients.

Another benefit of the national health database is that it provides the opportunity for research when some randomized control trial studies are impossible, for example, the design of beneficial evaluation studies for surgical intervention. According to the knowledge and training of the neurosurgeon and the ethical implications, decision-making regarding surgical intervention is focused on the best interests of the patients [45]. We cannot design a "randomized control" study for surgical intervention studies. Large dataset research may solve these problems. Even though there can be no truly double/triple blind randomized control trial for surgical intervention studies, the results of NHIRD data mining may still reflect the reality of sICH patients' care.

Change point analysis is a group of methods used to solve problems of change [32, 33]. These methods solve the problems of thresholds, identify the existence of any change point, and find the change point if there is one [34]. Huang et al. used change point analysis and defined 16 days as a surgical ICU prolonged stay [8]. This study used the NHIRD and defined 10 days as the threshold of a prolonged ICU stay for sICH patients using change point analysis. The result is reliable for sICH patients because this database represents the national medical care status in Taiwan, and change point analysis did find a valuable threshold for a prolonged ICU stay.

Although they represent only around 27% of patients, ICH patients with a prolonged ICU stay spend almost half of total hospital patient-day and more than 60% of total ICU patient-day for the care facilities. The hospital admission cost of the patients with a prolonged ICU stay (US$11,036) was almost 3 times that of patients with a nonprolonged ICU stay (US$3,155). A prolonged ICU stay may represent both hospital cost and mortality indicators for sICH patients. According to Figure 5, we found that more than 25% of ICU patients with a nonprolonged ICU stay died within one month, and then their mortality rate decreased gradually. This group (the nonprolonged group) consisted of two types of patients: the first is patients with an illness too severe to survive (early mortality patients); the other is patients who have a real nonprolonged ICU stay. There is no doubt that this group will cost less in terms of medical facilities than prolonged ICU stay patients. The early mortality patients will cost less than those with a prolonged ICU stay due to less ICU occupation duration or death during ICU admission. The other patients, the real nonprolonged ICU stay patients, cost less in terms of ICU facilities than patients with a prolonged ICU stay. Both types of patient cost less in terms of medical facilities than those patients with a prolonged ICU stay.

Comorbidities influence not only the outcome but also the hospital admission LOS of patients [15, 46]. Some studies have mentioned that different individual comorbidities could influence the outcome of sICH patients. For example, some studies have used hypertension or DM as outcome indicators, and other researches found that respiratory failure or pneumonia influence the LOS and mortality rate of sICH patients [14, 29, 47, 48]. However, few studies have discussed the influence of multiple comorbidities in sICH patients. This study defined multiple comorbidities as patients who have 2 or more than 2 kinds of major admission comorbidities. Only 4 major comorbidities are recorded during admission in the NHIRD. Actually, the number of major comorbidities plays an important role in outcome measurement and influences the LOS of sICH patients. This study found that patients with multiple major comorbidities will have a poorer outcome and a prolonged LOS. These patients utilize more hospital resources than those with a single comorbidity or no comorbidities, no matter whether the comorbidities arise from previous disease history or complications. Although the previous disease history of patients cannot be changed, decreasing complications using quality control methods will lead to a better outcome and a shorter LOS for sICH patients [14].

This study found that the northern area of Taiwan had higher ICU facilities usage and medical expenditure for the care of ICH patients. In contrast, the Eastern division, which is in a rural area, had both a lower ratio of prolonged ICU stay and a lower medical expenditure for ICH patients. The different training capacities also influence the medical expenditure and the ratio of a prolonged ICU stay in ICH patients. Medical centers have the highest ratio of a prolonged ICU stay because more severe patients will be transferred to medical centers. It is interesting, though, that local hospitals have a high ratio too. Families do not need to care for patients when they are admitted to the ICU. Psychosocial factors influence the ratio of a prolonged ICU stay in local hospitals. This study also found that private hospitals have a lower ratio of a prolonged ICU stay and incur a lower medical expenditure than public hospitals. It might be concluded that private hospitals are more efficient in terms of patient care than public hospitals. These results also proved that the threshold of a prolonged ICU stay is a good indicator of hospitals utility in some diseases. Different hospitals have their own different care strategies for those with a prolonged ICU stay.

There are many factors that influence both mortality and ICU stay, for example, diseases severity, psychosocial problems, less integrated care, and institutional factors [1]. Due to the disabilities of patients, the most important issue for ICH patients is care after the disease has been stabilized. Research has shown that early mobilization will decrease the LOS in ICU [49]. Proactive palliative care consultations for high-risk patients, for example, ICH patients, will significantly decrease the LOS in ICU [50]. Decision support systems for ICU staff are also good tools for decreasing the LOS in ICU [51]. The integration of care after discharge from ICU is also an important issue related to a decreased LOS in the ICU. Patients stay in ICUs and hospitals for a shorter period of time if integrated care systems have been established, and these methods will reduce the medical expenditure paid by insurance systems [52]. This study also found that private hospitals have lower prolonged ICU stay rates than public hospitals. It can be concluded that efficiency in terms of hospital management can also decrease the LOS in ICU. Further evaluations are needed to identify methods that can decrease the LOS in ICU for sICH patients.

There were some limitations in this research. First, the NHIRD is an administrative database used for health insurance. Therefore, this database did not collect information regarding the site and depth of the ICH. Some secondary outcomes, such as disabilities and sequelae after sICH, were not recorded in the NHIRD. However, according to information technologies and data mining methods, we were still able to find some hidden facts and knowledge from the medical database used. Although they are retrospective data, such results or knowledge has been proven following further study to be good models for medical care [53–56]. Second, this study presented a general definition of a prolonged ICU stay for ICH patients. Patients will tend to have a prolonged ICU stay if they have more comorbidities/complications. Different comorbidities/complications or diseases have different influences on the ICU length of stay. New prolonged

ICU stay definitions taking into account complications and comorbidities will be identified in future studies.

5. Conclusion

This study defined 10 days as a prolonged ICU stay using change point analysis. This study also showed that the threshold of a prolonged ICU stay is a good indicator of hospitals utilization in ICH patients. Different hospitals have their own different care strategies, which can be identified from a prolonged ICU stay. This definition can be a good indicator of quality control in hospital management. According to governmental strategies or health policies, the ICU and hospital lengths of stay will be shorter if integrated care systems are established. This could also reduce the medical cost paid by healthcare insurance systems.

Conflict of Interests

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

References

[1] D. A. Gruenberg, W. Shelton, S. L. Rose, A. E. Rutter, S. Socaris, and G. McGee, "Factors influencing length of stay in the intensive care unit," *The American Journal of Critical Care*, vol. 15, no. 5, pp. 502–509, 2006.

[2] T. Rotter, L. Kinsman, E. James et al., "Clinical pathways: effects on professional practice, patient outcomes, length of stay and hospital costs," *The Cochrane Database of Systematic Reviews*, vol. 3, Article ID CD006632, 2010.

[3] P. Thungjaroenkul, G. G. Cunmings, and A. Embleton, "The impact of nurse staffing on hospital costs and patient length of stay: a systematic review," *Nursing Economics*, vol. 25, no. 5, pp. 255–265, 2007.

[4] C. Weissman, "Analyzing intensive care unit length of stay data: problems and possible solutions," *Critical Care Medicine*, vol. 25, no. 9, pp. 1594–1600, 1997.

[5] C. M. Martin, A. D. Hill, K. Burns, and L. M. Chen, "Characteristics and outcomes for critically ill patients with prolonged intensive care unit stays," *Critical Care Medicine*, vol. 33, no. 9, pp. 1922–1927, 2005.

[6] K. B. Laupland, A. W. Kirkpatrick, J. B. Kortbeek, and D. J. Zuege, "Long-term mortality outcome associated with prolonged admission to the ICU," *Chest*, vol. 129, no. 4, pp. 954–959, 2006.

[7] V. Trottier, M. G. McKenney, M. Beninati, R. Manning, and C. I. Schulman, "Survival after prolonged length of stay in a trauma intensive care unit," *Journal of Trauma—Injury, Infection and Critical Care*, vol. 62, no. 1, pp. 147–150, 2007.

[8] Y. C. Huang, S. J. Huang, J. Y. Tsauo, and W. J. Ko, "Definition, risk factors and outcome of prolonged surgical intensive care unit stay," *Anaesthesia and Intensive Care*, vol. 38, no. 3, pp. 500–505, 2010.

[9] R. G. A. Ettema, L. M. Peelen, C. J. Kalkman, A. P. Nierich, K. G. M. Moons, and M. J. Schuurmans, "Predicting prolonged intensive care unit stays in older cardiac surgery patients: a validation study," *Intensive Care Medicine*, vol. 37, no. 9, pp. 1480–1487, 2011.

[10] A. W. Ong, L. A. Omert, D. Vido et al., "Characteristics and outcomes of trauma patients with ICU lengths of stay 30 days and greater: a seven-year retrospective study," *Critical Care*, vol. 13, no. 5, article R154, 2009.

[11] D. K. Sommerfeld, H. Johansson, A. Jönsson et al., "Rivermead mobility index can be used to predict length of stay for elderly persons, 5 days after stroke onset," *Journal of Geriatric Physical Therapy*, vol. 34, no. 2, pp. 64–71, 2011.

[12] M. H. Al-Eithan, M. Amin, and A. A. Robert, "The effect of hemiplegia/hemiparesis, diabetes mellitus, and hypertension on hospital length of stay after stroke," *Neurosciences*, vol. 16, no. 3, pp. 253–256, 2011.

[13] A. M. Naidech, B. R. Bendok, P. Tamul et al., "Medical complications drive length of stay after brain hemorrhage: a cohort study," *Neurocritical Care*, vol. 10, no. 1, pp. 11–19, 2009.

[14] A. Ingeman, G. Andersen, H. H. Hundborg, M. L. Svendsen, and S. P. Johnsen, "In-hospital medical complications, length of stay, and mortality among stroke unit patients," *Stroke*, vol. 42, no. 11, pp. 3214–3218, 2011.

[15] J. A. Luker, J. Bernhardt, and K. A. Grimmer-Somers, "Demographic and stroke-related factors as predictors of quality of acute stroke care provided by allied health professionals," *Journal of Multidisciplinary Healthcare*, vol. 4, pp. 247–259, 2011.

[16] A. I. Qureshi, A. D. Mendelow, and D. F. Hanley, "Intracerebral haemorrhage," *The Lancet*, vol. 373, no. 9675, pp. 1632–1644, 2009.

[17] J. Broderick, S. Connolly, E. Feldmann et al., "Guidelines for the management of spontaneous intracerebral hemorrhage in adults: 2007 update: a guideline from the American Heart Association/American Stroke Association Stroke Council, High Blood Pressure Research Council, and the Quality of Care and Outcomes in Research Interdisciplinary Working Group," *Circulation*, vol. 116, no. 16, pp. e391–e413, 2007.

[18] D. L. Labovitz, A. Halim, B. Boden-Albala, W. A. Hauser, and R. L. Sacco, "The incidence of deep and lobar intracerebral hemorrhage in whites, blacks, and Hispanics," *Neurology*, vol. 65, no. 4, pp. 518–522, 2005.

[19] S. A. Mayer and F. Rincon, "Treatment of intracerebral haemorrhage," *The Lancet Neurology*, vol. 4, no. 10, pp. 662–672, 2005.

[20] C. J. van Asch, M. J. Luitse, G. J. Rinkel, I. van der Tweel, A. Algra, and C. J. Klijn, "Incidence, case fatality, and functional outcome of intracerebral haemorrhage over time, according to age, sex, and ethnic origin: a systematic review and meta-analysis," *The Lancet Neurology*, vol. 9, no. 2, pp. 167–176, 2010.

[21] N. C. Smeeton, P. U. Heuschmann, A. G. Rudd et al., "Incidence of hemorrhagic stroke in black Caribbean, black African, and white populations: The South London Stroke Register, 1995–2004," *Stroke*, vol. 38, no. 12, pp. 3133–3138, 2007.

[22] C. L. Chan, H. W. Ting, and H. T. Huang, "The incidence, hospital expenditure, and, 30 day and 1 year mortality rates of spontaneous intracerebral hemorrhage in Taiwan," *Journal of Clinical Neuroscience*, vol. 21, no. 1, pp. 91–94, 2014.

[23] A. Damasceno, J. Gomes, A. Azevedo et al., "An epidemiological study of stroke hospitalizations in maputo, mozambique: a high burden of disease in a resource-poor country," *Stroke*, vol. 41, no. 11, pp. 2463–2469, 2010.

[24] R. Vibo, J. Kõrv, and M. Roose, "One-year outcome after first-ever stroke according to stroke subtype, severity, risk factors and pre-stroke treatment. A population-based study from Tartu, Estonia," *European Journal of Neurology*, vol. 14, no. 4, pp. 435–439, 2007.

[25] G. Gioldasis, P. Talelli, E. Chroni, J. Daouli, T. Papapetropoulos, and J. Ellul, "In-hospital direct cost of acute ischemic and hemorrhagic stroke in Greece," *Acta Neurologica Scandinavica*, vol. 118, no. 4, pp. 268–274, 2008.

[26] L. B. Morgenstern, J. C. Hemphill III, C. Anderson et al., "Guidelines for the management of spontaneous intracerebral hemorrhage: A guideline for healthcare professionals from the American Heart Association/American Stroke Association," *Stroke*, vol. 41, no. 9, pp. 2108–2129, 2010.

[27] P. Bhattacharya, L. Shankar, S. Manjila, S. Chaturvedi, and R. Madhavan, "Comparison of outcomes of nonsurgical spontaneous intracerebral hemorrhage based on risk factors and physician specialty," *Journal of Stroke and Cerebrovascular Diseases*, vol. 19, no. 5, pp. 340–346, 2010.

[28] V. L. Feigin, C. M. Lawes, D. A. Bennett, S. L. Barker-Collo, and V. Parag, "Worldwide stroke incidence and early case fatality reported in 56 population-based studies: a systematic review," *The Lancet Neurology*, vol. 8, no. 4, pp. 355–369, 2009.

[29] M. C. Christensen, I. Previgliano, F. J. Capparelli, D. Lerman, W. C. Lee, and N. A. Wainsztein, "Acute treatment costs of intracerebral hemorrhage and ischemic stroke in Argentina," *Acta Neurologica Scandinavica*, vol. 119, no. 4, pp. 246–253, 2009.

[30] S. Koton, N. M. Bornstein, R. Tsabari, and D. Tanne, "Derivation and validation of the Prolonged Length of Stay score in acute stroke patients," *Neurology*, vol. 74, no. 19, pp. 1511–1516, 2010.

[31] *Introduction to the National Health Insurance Research Database (NHIRD)*, http://w3.nhri.org.tw/nhird/.

[32] A. SenGupta and A. K. Laha, "A Bayesian analysis of the change-point problem for directional data," *Journal of Applied Statistics*, vol. 35, no. 5-6, pp. 693–700, 2008.

[33] F. Li, G. C. Runger, and E. Tuv, "Supervised learning for change-point detection," *International Journal of Production Research*, vol. 44, no. 14, pp. 2853–2868, 2006.

[34] W. Ning and A. K. Gupta, "Change point analysis for generalized lambda distribution," *Communications in Statistics. Simulation and Computation*, vol. 38, no. 8–10, pp. 1789–1802, 2009.

[35] S. Nam, J. H. Cha, and S. Cho, "A Bayesian change-point analysis for software reliability models," *Communications in Statistics. Simulation and Computation*, vol. 37, no. 8–10, pp. 1855–1869, 2008.

[36] C. T. Jose, B. Ismail, and S. Jayasekhar, "Trend, growth rate, and change point analysis—a data driven approach," *Communications in Statistics—Simulation and Computation*, vol. 37, no. 3–5, pp. 498–506, 2008.

[37] P. Liu, S. Guo, L. Xiong, and L. Chen, "Flood season segmentation based on the probability change-point analysis technique," *Hydrological Sciences Journal*, vol. 55, no. 4, pp. 540–554, 2010.

[38] D. Rudoy, S. G. Yuen, R. D. Howe, and P. J. Wolfe, "Bayesian change-point analysis for atomic force microscopy and soft material indentation," *Journal of the Royal Statistical Society C: Applied Statistics*, vol. 59, no. 4, pp. 573–593, 2010.

[39] "Change-Point Analysis: A Powerful New Tool for Detecting Changes," http://www.variation.com/cpa/tech/changepoint.html.

[40] P. Gavit, Y. Baddour, and R. Tholmer, "Use of change-point analysis for process monitoring and control: a better method for trend analysis than CUSUM and control charts," *BioPharm International*, vol. 22, no. 8, pp. 46–55, 2009.

[41] R. Boulkedid, O. Sibony, C. Bossu-Salvador, J. F. Oury, and C. Alberti, "Monitoring healthcare quality in an obstetrics and gynaecology department using a CUSUM chart," *BJOG*, vol. 117, no. 10, pp. 1225–1235, 2010.

[42] N. P. Yang, H. C. Chen, D. V. Phan et al., "Epidemiological survey of orthopedic joint dislocations based on nationwide insurance data in Taiwan, 2000–2005," *BMC Musculoskeletal Disorders*, vol. 12, article 253, 2011.

[43] N. P. Yang, C. L. Chan, I. L. Yu, C. Y. Lee, and P. Chou, "Estimated prevalence of orthopaedic fractures in Taiwan—a cross-sectional study based on nationwide insurance data," *Injury*, vol. 41, no. 12, pp. 1266–1272, 2010.

[44] D. Bin-Chia Wu, C. Chang, Y. Huang, Y. Wen, C. Wu, and C. S. Fann, "Cost-effectiveness analysis of pneumococcal conjugate vaccine in Taiwan: a transmission dynamic modeling approach," *Value in Health*, vol. 15, no. 1, pp. S15–S19, 2012.

[45] A. Leung, S. Luu, G. Regehr, M. L. Murnaghan, S. Gallinger, and C. Moulton, ""First, do no harm": balancing competing priorities in surgical practice," *Academic Medicine*, vol. 87, no. 10, pp. 1368–1374, 2012.

[46] T. Vu, C. F. Finch, and L. Day, "Patterns of comorbidity in community-dwelling older people hospitalised for fall-related injury: a cluster analysis," *BMC Geriatrics*, vol. 11, article 45, 2011.

[47] N. Andaluz and M. Zuccarello, "Recent trends in the treatment of spontaneous intracerebral hemorrhage: analysis of a nationwide inpatient database," *Journal of Neurosurgery*, vol. 110, no. 3, pp. 403–410, 2009.

[48] S. Tetri, S. Juvela, P. Saloheimo, J. Pyhtinen, and M. Hillbom, "Hypertension and diabetes as predictors of early death after spontaneous intracerebral hemorrhage: clinical article," *Journal of Neurosurgery*, vol. 110, no. 3, pp. 411–417, 2009.

[49] A. Hunter, L. Johnson, and A. Coustasse, "Reduction of intensive care unit length of stay: the case of early mobilization," *The Health Care Manager*, vol. 33, no. 2, pp. 128–135, 2014.

[50] S. A. Norton, L. A. Hogan, R. G. Holloway, H. Temkin-Greener, M. J. Buckley, and T. E. Quill, "Proactive palliative care in the medical intensive care unit: effects on length of stay for selected high-risk patients," *Critical Care Medicine*, vol. 35, no. 6, pp. 1530–1535, 2007.

[51] C. W. Hatler, C. Grove, S. Strickland, S. Barron, and B. D. White, "The effect of completing a surrogacy information and decision-making tool upon admission to an intensive care unit on length of stay and charges," *The Journal of Clinical Ethics*, vol. 23, no. 2, pp. 129–138, 2012.

[52] C. Chan, H. You, H. Huang, and H. Ting, "Using an integrated COC index and multilevel measurements to verify the care outcome of patients with multiple chronic conditions," *BMC Health Services Research*, vol. 12, no. 1, article 405, 2012.

[53] E. E. Schadt, "The changing privacy landscape in the era of big data," *Molecular Systems Biology*, vol. 8, article 612, 2012.

[54] A. McAfee and E. Brynjolfsson, "Big data: the management revolution," *Harvard Business Review*, vol. 90, no. 10, pp. 60–128, 2012.

[55] D. Howe, M. Costanzo, P. Fey et al., "Big data: the future of biocuration," *Nature*, vol. 455, no. 7209, pp. 47–50, 2008.

[56] H.-W. Ting, J.-T. Wu, C.-L. Chan, S.-L. Lin, and M.-S. Chen, "Decision model for acute appendicitis treatment with decision tree technology—a modification of the alvarado scoring system," *Journal of the Chinese Medical Association*, vol. 73, no. 8, pp. 401–406, 2010.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 291067, 10 pages
http://dx.doi.org/10.1155/2014/291067

Research Article

Gonadal Transcriptome Analysis of Male and Female Olive Flounder (*Paralichthys olivaceus*)

Zhaofei Fan,[1,2] Feng You,[1] Lijuan Wang,[1,2] Shenda Weng,[1] Zhihao Wu,[1] Jinwei Hu,[1,2] Yuxia Zou,[1] Xungang Tan,[1] and Peijun Zhang[1]

[1] *Key Laboratory of Experimental Marine Biology, Institute of Oceanology, Chinese Academy of Sciences, 7 Nanhai Road, Qingdao 266071, China*
[2] *University of the Chinese Academy of Sciences, Beijing 10049, China*

Correspondence should be addressed to Feng You; youfeng@qdio.ac.cn

Received 7 April 2014; Revised 12 June 2014; Accepted 15 June 2014; Published 6 July 2014

Academic Editor: Tatsuya Akutsu

Olive flounder (*Paralichthys olivaceus*) is an important commercially cultured marine flatfish in China, Korea, and Japan, of which female grows faster than male. In order to explore the molecular mechanism of flounder sex determination and development, we used RNA-seq technology to investigate transcriptomes of flounder gonads. This produced 22,253,217 and 19,777,841 qualified reads from ovary and testes, which were jointly assembled into 97,233 contigs. Among them, 23,223 contigs were mapped to known genes, of which 2,193 were predicted to be differentially expressed in ovary and 887 in testes. According to annotation information, several sex-related biological pathways including ovarian steroidogenesis and estrogen signaling pathways were firstly found in flounder. The dimorphic expression of overall sex-related genes provides further insights into sex determination and gonadal development. Our study also provides an archive for further studies of molecular mechanism of fish sex determination.

1. Introduction

According to Helfman et al., there are almost 30,000 species of fish distributed nearly in all the aquatic habitats around the world [1]. They are the most abundant vertebrates on Earth, showing a diversity of species unmatched by other classes. Not surprisingly given this extreme diversity, fish exhibit all known forms of vertebrate sex determination to adapt to the variable habitats [2]. In the meantime, economic values of growth rate, time and age of maturation, body shape, and carcass composition are related to their sexual development in some edible fish species [3]. Then, we are interested in fish sex determination mechanism. As reported, fish sex determination patterns can be classified as genetic sex determination (GSD) and environmental sex determination (ESD) forms. However, the feature of sex determination in fish is remarkably flexible; even individuals with GSD can be influenced by environmental factors like temperature, that is, GSD + EE (environmental effects) form [4]. Although fish have several different sex determination forms, it is hypothesized that genes involved in sex determination are probably conserved throughout evolution. Several genes have been confirmed as master genes of sex determination in some fish species. In the medaka (*Oryzias latipes*), a homologue of the *dmrt1* gene (called *dmy*) is located on the Y chromosome, and its expression is a necessary and sufficient condition for triggering testicular development in bipotential gonads [5, 6]. Recently, five novel sex determining genes (or candidates) have been reported in other fish: *amhy* in Patagonian pejerrey (*Odontesthes hatcheri*) [7], *irf9y* in rainbow trout (*Oncorhynchus mykiss*) [8], *gsdf* in *Oryzias luzonensis* (a relative of medaka) [9], *amhr2* in fugu (*Takifugu rubripes*) [10], and *dmrt1* in half-smooth tongue sole (*Cynoglossus semilaevis*) [11]. Besides the sex determining genes, some conserved genes shown to play important roles in mammal sex determination and differentiation were cloned and identified in fish. These include *cyp19*, *foxl2*, *sf1*, *dax1*, *wt1*, *mis*, *dmrt*, and *sox9* [12]. In mammals, these genes act together to constitute complicated network whereby sex phenotype is established [13]. However, studies on the function and connections of

the above genes in fish are limited. More or novel sex-related genes are also needed to be found out. And then, the complex mechanism of fish sex determination could be adequately explained.

Over the past decade, significant progress has been made in genome-wide gene expression profiling by the development and application of large scale sequencing technique, which can easily show more differential expressional genes in different traits, such as gender. Transcriptome profiling associated with sex determination and differentiation using RNA-seq of several fish, including platyfish (*Xiphophorus maculatus*) [14], rainbow trout (*O. mykiss*) [15], Nile tilapia (*Oreochromis niloticus*) [16], rockfish (*Sebastiscus marmoratus*) [17], catfish (*Ictalurus punctatus*) [18], and turbot (*Scophthalmus maximus*) [19], was shown. These data provided transcriptomic information expressed in gonads at particular condition and time and identified sex differentially expressed genes, while there is almost no report on biological pathways including gonadal steroidogenesis pathway in male and female fish.

As an important commercially marine flatfish, olive flounder (*Paralichthys olivaceus*) is mainly cultured in China, Japan, and Korea. The female flounder grows significantly faster and bigger than the male one [20, 21]. Breeding and culturing all-female population of flounder is a promising approach to boost production. Thus, the study on sex manipulation of flounder has been attracting researchers' interests. Olive flounder has XX (female)/XY (male) sex determination system indirectly inferred from the female-dominant phenotype among the gynogenesis offspring [21]. Its formation of sex phenotype is also influenced by environmental factors such as water temperature and external hormone [22]. Sex-related conserved genes have been studied to elucidate the sexual molecular mechanism in flounder. Kitano et al. firstly cloned flounder *cyp19a* gene and found female predominant expression pattern [23], whereafter cloning and expression profile analysis of *cyp19a* and its transcription factors such as *foxl2* [24] and *dmrt1* [25] and other sex-related genes including *cyp17* [26] and *dmrt4* [27] were conducted with male and female flounder. However, we have got only limited message about flounder sex determination, and more sex-related genes and their functions need to be studied. Further identification of the expression profile of genes involved in gonadal development using RNA-seq may help to illuminate the gene regulatory network controlling sex determination and subsequent maintenance of phenotypic sex. In this study, we sequenced flounder gonadal transcriptomes and identified the differences in gene expression profiles between ovary and testis and relevant biological pathways. These data would provide a useful genomic resource for future study on sex determination and for selection of candidate genes involved in these processes in flounder.

2. Materials and Methods

2.1. Fish. Adult and juvenile flounder (12~40 cm in total length, TL) used in the present study were collected from Shenghang fish farm (Weihai, China) or purchased from Nanshan market (Qingdao, China) and were temporarily cultured in a 3 m^3 aerated seawater tank at the institute aquarium and fed with commercial particle food twice a day. Gonads were retrieved from the abdomen of fish after anesthetization. Their genders were identified by morphological observation of gonads [28]. Each gonadal sample was divided into two halves. The first half was fixed in Davison's fixative solution for identification of gonadal developmental stage using a histology method as described below. The second half was stored immediately in liquid nitrogen for RNA isolation. Totally, the gonadal tissues at developmental stage I (3 testes and 3 ovaries), II (2 testes and 3 ovaries), and III (2 testes and 3 ovaries) were used in this project. Semen and eggs were gently squeezed out from 2 mature male flounder (40~50 cm in TL) and 3 mature female flounder (50~60 cm in TL), respectively, and the samples were immediately frozen and stored in liquid nitrogen for RNA isolation. All animal work has been conducted according to relevant national and international guidelines. Animal protocols were approved by the Institute of Oceanology, Chinese Academy of Science.

2.2. Gonadal Histology. For histological analysis, ovary or testes tissue from each adult or juvenile flounder was fixed in Davison fixative solution for 24 h and stored in 70% ethanol. Pieces of gonadal tissue were cut down, dehydrated using ethanol (gradient: 70%~100%), and finally embedded in paraffin. The sections were conducted at the thickness of 5~7 μm, dewaxed by ethanol (gradient: 100%~50%), washed by distilled water, stained by hematoxylin-eosin, and observed microscopically after air drying. Slices of each sample were mounted on glass slides, stained with hematoxylin, and counterstained with eosin (HE staining) to determine the developmental stage of gonads. Detailed methods can be found in Sun et al. [22] and Radonic and Macchi [29].

2.3. RNA Isolation and cDNA Library Construction. Total RNA was isolated from each sample using Trizol Reagent (Invitrogen, USA, http://www.lifetechnologies.com/cn/zh/home/brands/invitrogen) based on the protocol. The genomic DNA was eliminated by treatment with DNase1 (10 U/mL, Ambion, USA, http://www.lifetechnologies.com/cn/zh/home/brands/ambion) at 37°C for 1 h. The purified mRNA was enriched by Micropoly(A) Purist RNA purification kit (Amion, USA), and the concentration and integrity of mRNA were qualified using Agilent 2100 Bioanalyzer (Agilent Technologies, USA, http://www.home.agilent.com). The mRNAs from ovarian developmental stages I, II, and III and egg were mixed together to synthesize cDNA, and mRNAs from testicular developmental stages I, II, and III and sperm were mixed together correspondingly. These blended mRNAs served as templates to synthesize first-strand cDNA using GsuI-oligo dT primer; the reaction was performed with Superscript II reverse transcriptase (Invitrogen, USA) at 42°C for 1 h. Biotins were subsequently attached to the 5′cap of mRNA oxidized by $NaIO_4$ (Sigma, USA, http://www.sigmaaldrich.com), whereby the biotin-labeled mRNA /cDNA could be sublimed by Dynal M280 magnetic beads (Invitrogen, USA). Released from

the hybrid strands by alkaline lysis, the first-strand cDNA was attached with an adaptor at its 5′end. Second-strand cDNA was synthesized using Ex Taq polymerase (Takara, Japan, http://www.takara-bio.com/) based on the first-strand modified cDNA; then poly(A) and 5′adaptor were trimmed by GsuI enzyme. The synthesized cDNA was disrupted into short fragments (300–500 nt) by ultrasound instrument which were further enriched by Ampure beads (Agencourt, USA, https://www.beckmancoulter.com). These purified cDNA fragments were used to construct cDNA library with the method of TruSeqTM DNA sample Prep kit-set (Illumina, USA, http://www.illumina.com/). Finally, the two cDNA libraries were sequenced on Illumina Solexa using paired-end strategy in a single run.

2.4. Illumina Sequencing, Functional Annotation, and Bioinformatics Analysis. Total reads were produced through Illumina Solexa instrument (2 ∗ 100 bp pair-end sequencing) from Chinese National Human Genome Center in Shanghai. The clean reads were obtained from original data by filtering out reads inclusive of unknown nucleotides and low-quality reads in which Q5 percentage (Q5 percentage is proportion of nucleotides with quality value larger than 5) is less than 50%. All clean reads of the two libraries were jointly assembled into contigs performed by Trinity software. The assembled contigs were conducted to predict protein-coding region by GetORF module of EMBOSS package [30]. All the protein-coding sequences were submitted for blastp similarity searches against the NCBI nonredundant (NR) protein database and Eukaryotic Ortholog Groups (KOG) database with the *e*-value of top hit lower than 1 e^{-5}. Furthermore, GoPipe software was used to perform blastp (cut-off *e*-value of <1 e^{-5}) search against the Swiss-Prot database and TrEMBL database. With the result of blastp, gene ontology (GO) annotation associated with "biological process," "molecular function," and "cellular component" was obtained using the gene2go. Likewise, the predicted protein sequence was submitted for bidirectional blastp (cut-off *e*-value of <1 e^{-3}) similarity searches against Kyoto Encyclopedia of Genes and Genomes (KEGG) database to assign KEGG Orthology (KO) number. According to the KO assignment, metabolic pathways were generated with tools supplied by KEGG [31]. The mapped read count of given gene is affected by its length and sequencing depth; the reads per kb per million reads (RPKM) were calculated to standardize gene expression level [32]:

$$\mathrm{RPKM} = \frac{10^6 C}{NL/10^3}. \quad (1)$$

Here, C indicates the mapped read count of a given gene from a given library. L indicates the length of a given gene. N indicates total mapped read count of a given library.

2.5. Identification of Sex-Related Differentially Expressed Genes. RPKM was directly used to compare the difference of gene expression level between male and female. This process was completed by DEGseq (an *R* package) based on the MARS model (MA-plot-based with random sampling model)

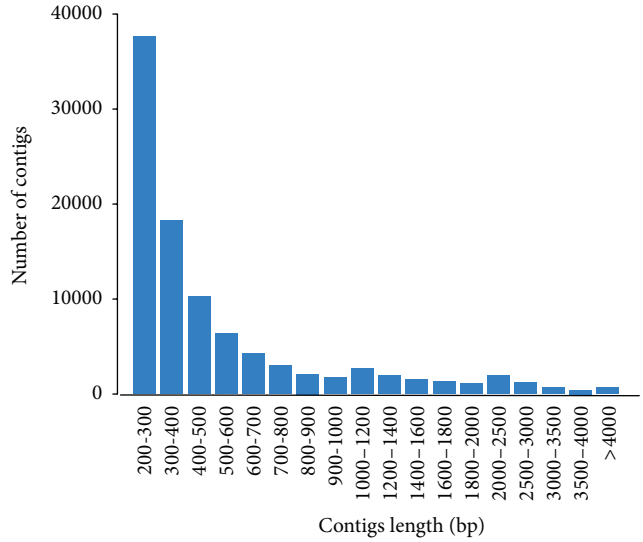


Figure 1: Length distribution of contigs.

[33]. We used Benjamini-Hochberg method to determine the threshold of the *q*-value in multiple testing. In our study, "$q < 10^{-3}$" and "$|\log_2(\mathrm{RPKM\_XX/RPKM\_XY})| \geq 2$" were chosen to identify sex-biased genes. Furthermore, we adopted four strategies to excavate sex-related genes from them: (i) complete cDNA sequences of well-known sex-related genes were downloaded from NCBI nucleotide database (http://www.ncbi.nlm.nih.gov/nucleotide/) and were conducted local blast (cut-off *e*-value of <1 e^{-10}) search against the local contigs database; (ii) sex-related keywords were retrieved in the annotation of sex-biased genes; (iii) among the sex-related KEGG pathways, the particular genes encoding pivotal enzyme were selected; (iv) genes reported to be relevant to sex differentiation were selected from the sex-biased genes.

3. Results

3.1. Assembly, Annotation, and Bioinformatical Analysis. Approximately 20.1 million and 22.4 million reads were obtained from male library and female library, respectively. After quality filtering, about 98.1% read of male and 99.0% read of female remain to be qualified for assembling. Sequencing saturation distribution (see Figure S1 in Supplementary Material available online at http://dx.doi.org/10.1155/2014/291067) and genes coverage statistic (Figure S2) analysis justified a deep sequencing coverage sufficient for the quantitative analysis of gene expression profiles. The clean reads were jointly assembled into 97,233 contigs with N50 of 809 bp and average length of 603 bp. The size distribution of assembled contigs was presented in Figure 1. Mapped to protein database, nearly 22.31% (21,697) of contigs were matched with known existing protein (Table S1). The function of the predicted proteins was classified with GO assignments statistically analyzed in Figure S3. In total, there were 20,582 proteins assigned with 132,920 GO terms and the three corresponding organizing principles.

Table 1: Overview of transcriptome analysis.

Total number of reads	42,640,333
Number of male reads	20,161,959
Number of female reads	22,478,374
Number of male reads in contigs	14,266,081
Number of female reads in contigs	20,939,047
Number of contigs from assembly	97,233
Number of annonated contigs	21,697
Number of contigs >400 bp	41,135
Number of male only contigs	25,226
Number of female only contigs	10,750
Number of mixed contigs	61,257
Sex-biased genes with q-value $\leq 1 \times 10^{-3}$ AND $\lvert\log_2 \text{ratio}\rvert \geq 2$	13,644
Male-biased contigs	10,348
Annonated male-biased genes	887
Female-biased contigs	3296
Annonated female-biased genes	2193

About 21 categories of biological process were assigned for 49,204 contigs, 11 categories of cellular component were assigned for 49,253 contigs, and 25 categories of molecular function were assigned for 41,788 contigs. Most of the cellular component genes were associated with cells and intracellular components, and most of the molecular function genes were associated with binding and catalytic activity. Mapped to reference canonical pathways in the KEGG database, 11,936 contigs were assigned with KO numbers. Through the KO number of the predicted protein, 328 metabolic pathways were constructed with various degrees (Table S1_KEGG pathway). Figure S4 showed the category distribution of biological pathways, in which the most enriched pathway is associated with signal transduction.

3.2. Sex-Biased Genes. Total 10348 male-biased and 3296 female-biased contigs were identified, respectively, (Table S2) showing significant expression difference between male and female. Among these sex-biased contigs, 887 (8.57%) male-biased contigs and 2193 (66.54%) female-biased contigs were annotated with known genes. In addition, 2111 contigs were identified to be expressed specifically in male and 75 contigs were identified to be expressed specifically in female. The statistical overview of transcriptome analysis was shown in Table 1. By using four strategies, sex-related well-documented genes were identified. In strategy i, we searched for well-known candidate genes already characterized to be sexually dimorphic in flounder. Among these genes, SRY-box containing protein 9 gene (*sox9*), *sox8a* mullerian inhibiting substance (*mis*), doublesex and mab-3 related transcription factor 1 gene (*dmrt1*), and so forth were predominantly expressed in male library, whereas P450 aromatase gene (*cyp19a)*, forkhead transcription factor L2 gene (*foxl2*), orphan nuclear receptor *dax1,* and so forth were obviously overexpressed in the female library. We also found out *sox6b* gene containing SRY-box, which is not reported in fish.In strategy ii, a set of keywords, including male, female, sex, sperm, egg, ovary, testis, estrogen, and androgen, were used to search sex-related genes based on annotation results. Amounts of genes consisting of the above keywords such as zona pellucida sperm-binding protein gene (*zp*), egg envelope glycoprotein-like precursor, and ovarian cancer-associated gene 2 (*ovca2*) exhibit sex-biased expression pattern. In strategy iii, among the steroidogenic enzyme genes, steroidogenic acute regulatory protein gene (*star*), 17-beta-hydroxysteroid dehydrogenase type 1 gene (*hsd17b1*), and estradiol 17-beta-dehydrogenase 12 gene (*hsd17b12*) present sex differential expression profile. In strategy iv, cathepsins (*ctss*), ropporin-1-like protein gene (*ropn1l*), ZPA domain containing protein precursor gene (*zpa*), *zpc5*, zygote arrest protein 1 gene (*zar1*), wee1-like protein kinase 2 gene (*wee2*), P43 5S RNA-binding protein gene (*42sp43*), histone H2Ax gene (h2ax), and so forth display differential expression profile between male and female. The sex-related genes identified were listed in Table 2.

3.3. Sexual Dimorphic Biological Pathway. To identify biological pathway that shows sexual dimorphism in flounder, the number of sex-biased genes was counted in different category of pathways. The result exhibits in Figure S5. It shows that the number of upregulated genes in most of metabolic pathways is much more in female than that in male. There are the most sex-biased genes in the signal transduction pathway compared with other pathways. And the significantly divergent pathways of male and female gene numbers mainly include lipid metabolism, signal transduction, translation, and cell growth death. Among them, several pathways associated with gonadal development and sex maintenance were found, such as ovarian steroidogenesis (shown as Figure 2 and Table 3), estrogen signaling pathway, progesterone-mediated oocyte maturation, prolactin signaling pathway, GnRH signaling pathway, oocyte meiosis, TGF-beta signaling pathway, steroid hormone biosynthesis, and Wnt signaling pathway.

Table 2: Representative sex-biased genes involved in sex determination and development.

Contigs	Length	Function	$\log_2$ (fold)	q-value	Sex	Gene
Comp134314_c0_seq1	2117	P450 aromatase	8.1274	$1.08E-71$	Female	*cyp19a*
Comp128200_c0_seq1	1721	Forkhead transcription factor L2	7.7734	$1.51E-29$	Female	*foxl2*
Comp129336_c0_seq1	697	Orphan nuclear receptor Dax1	2.3341	$2.96E-07$	Female	*dax1*
Comp135979_c0_seq1	3970	Vitellogenin receptor	2.4058	0	Female	*vtgr*
Comp131376_c0_seq1	1862	Zona pellucida sperm-binding protein 4	6.733436	0	Female	*zp4*
Comp129690_c1_seq1	1123	Zona pellucid sperm-binding protein 3	6.8619	0	Female	*zp3*
Comp126058_c0_seq1	1449	Zona pellucida sperm-binding protein	6.1157	0	Female	*zp*
Comp96429_c0_seq1	668	Histone H2A.x	6.3287	0	Female	*h2a.x*
Comp124668_c0_seq1	1372	Zygote arrest protein 1	6.556118	0	Female	*zar1*
Comp138502_c2_seq1	2515	ZPA domain containing protein precursor	7.7351	0	Female	*zpa*
Comp131365_c0_seq1	2442	ZPC5	6.5016	0	Female	*zpc5*
Comp130117_c0_seq1	1092	High choriolytic enzyme 2	7.5466	0	Female	*hce2*
Comp135460_c1_seq1	4433	Protein fem-1 homolog C	2.2905	$1.58E-173$	Female	*fem1c*
Comp130482_c0_seq1	1936	Frizzled-3	3.2588	$7.16E-20$	Female	*fzd3*
Comp126332_c0_seq1	1789	P43 5S RNA-binding protein	8.3272	0	Female	*42sp43*
Comp134412_c0_seq1	3026	Transcription factor IIIA	3.7385	0	Female	*gtf3a*
Comp129656_c0_seq1	2020	Wee1-like protein kinase 2	6.9006	0	Female	*wee2*
Comp126733_c0_seq1	1431	Cathepsin S precursor	5.1171	0	Female	*ctss*
Comp137073_c0_seq1	2247	G2/mitotic-specific cyclin-B1-like	6.7561	0	Female	*ccnb1*
Comp126534_c0_seq1	1128	Ovarian cancer-associated gene 2 protein	2.744	$4.03E-57$	Female	*ovca2*
Comp130477_c0_seq1	1312	17-beta-Hydroxysteroid dehydrogenase type 1	5.2759	$3.24E-34$	Female	*hsd17b1*
Comp128629_c0_seq1	1286	SRY-box containing protein 9	−4.0548	$3.25E-33$	Male	*sox9*
Comp138102_c0_seq1	1849	Mullerian inhibiting substance	−5.7273	$1.34E-283$	Male	*mis*
Comp135961_c0_seq1	2582	Sox8a	−4.3658	0	Male	*sox8a*
Comp125855_c0_seq1	1045	Steroidogenic acute regulatory protein	−2.1114	$6.75E-10$	Male	*star*
Comp70046_c1_seq1	523	SRY-box containing gene 6b	−2.9959	$2.38E-25$	Male	*sox6b*
Comp125855_c0_seq1	1045	Steroidogenic acute regulatory protein	−2.1114	$6.75E-10$	Male	*star*
Comp70129_c0_seq1	1376	Cytochrome c oxidase subunit I	−13.8122	0	Male	*cox I*
Comp127682_c0_seq1	960	ATP synthase F0 subunit 6	−12.5508	$4.52E-268$	Male	*atp5g*
Comp127263_c1_seq1	1000	Heat shock protein 90 alpha	−6.5865	0	Male	*hsp90α*
Comp132152_c1_seq1	3496	Sperm flagellar protein 2	−3.2844	$3.69E-302$	Male	*spef*
Comp138413_c1_seq1	2470	AMY-1-associating protein	−2.2291	$9.22E-53$	Male	*aat1*
Comp131350_c0_seq1	1398	Axonemal dynein light intermediate polypeptide 1	−9.422	0	Male	*dnali1*
Comp128419_c1_seq1	659	Ropporin-1-like protein	−8.2764	$2.46E-27$	Male	*ropn1l*

Note: Fold indicates RPKM (female)/RPKM (male); q-value indicates false discove rate (FDR).

4. Discussion

As a marine economic flatfish, flounder exhibit GSD + EE sex determination form, and GSD function is necessary for us to probe into. Now, little evidence can demonstrate the molecular mechanism of flounder sex determination; therefore more sex-related genes and explicit biological pathways are needed to imply the mechanism. In this study, we used RNA-seq technology to identify large quantities of sex-biased genes and illustrated its function from the perspective of biological pathways. Among these sex-biased genes, some are associated with the sex determination and gonadal development as reported. The other sex-biased genes need further study to investigate their connection with sex determination and differentiation. The sex-related genes identified in this study could provide an important clue for sex determination mechanism of flounder. In addition, the novel contigs may be from unknown gene sequence or alternatively spliced transcripts. Yano et al. [8] characterized a novel gene expressed only in testis through analyzing gonadal transcriptome of rainbow trout (*O. mykiss*), and further study revealed that this gene is Y chromosome sequence tightly linked with sex locus and necessary to trigger testicular differentiation. Therefore, our study also provides considerable novel sex-biased genes for further study on sex determination of flounder.

4.1. Sex Differences in Gene Expression Profiles of Gonads. We analyzed the overall gene expression profiles of gonads and identified numerous sex-related genes. Some of the sex-biased genes are known to show sexual dimorphism between ovary and testis testifying the reliability of the selection

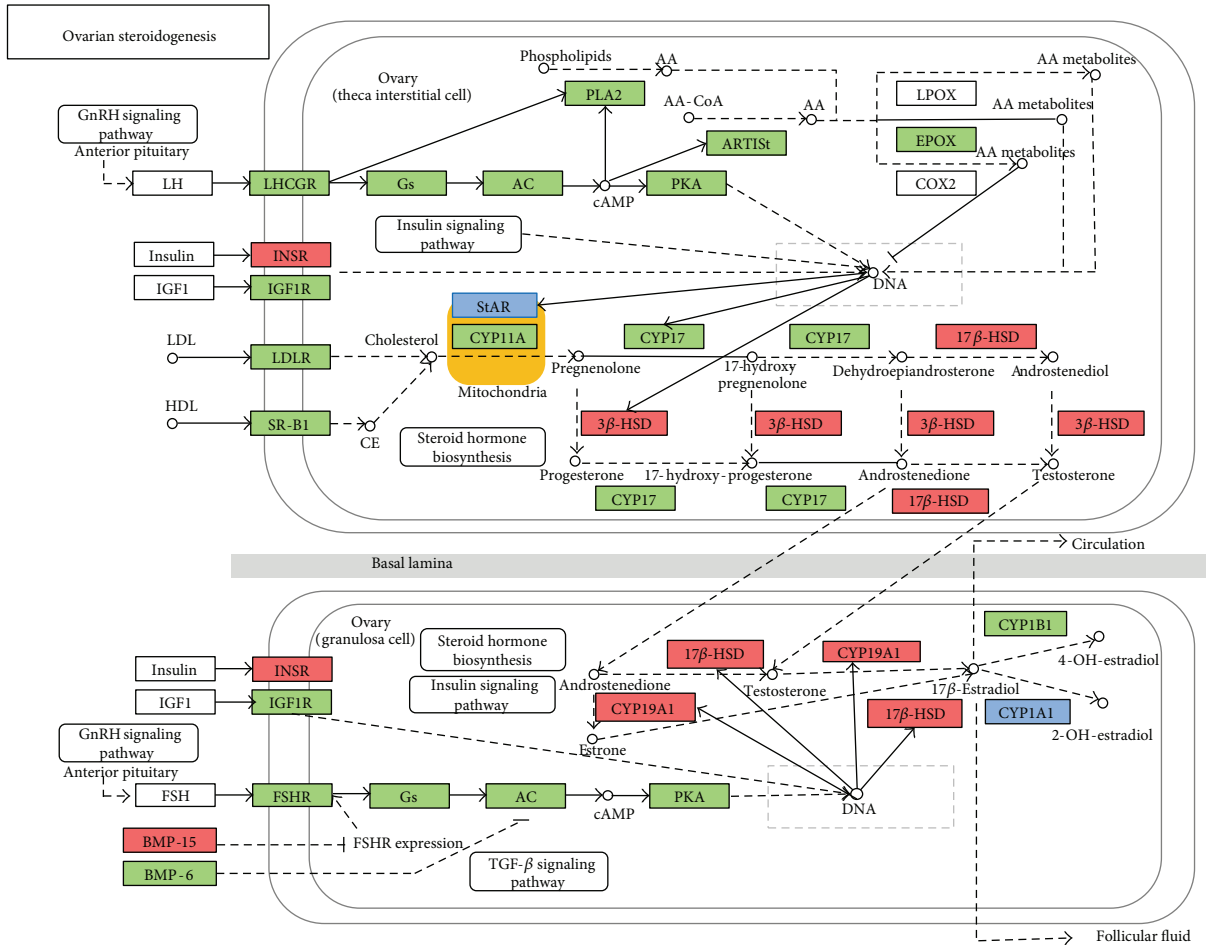


FIGURE 2: Ovarian steroidogenesis pathway. Green background indicates genes identified in flounder; red background indicates female-biased genes; blue background indicates male-biased genes.

criterion in the results. Our analyses also found considerable previously uncharacterized sex-biased genes. Further functional characterization of these genes using transgenic overexpression, knockout strategies, and knockdown strategies may help elucidate the molecular mechanisms controlling sex determination and gonadal development in teleost.

4.1.1. Male-Biased Genes. In the present study, more contigs were found to be male-biased genes than female-biased genes. For lack of genomic sequence of flounder, a large amount of male-biased contigs cannot be further assembled and annotated. Among the annotated genes, several well-documented and important male-enriched genes are listed in more detail. *Dmrt1* is expressed in the embryonic gonads of many vertebrates. It was thought to repress the female pathway through inhibition of *cyp19a1a* expression in Nile tilapia (*O. niloticus*) [34]. In medaka (*O. latipes*), *dmy*, a duplicated copy of the *dmrt1* on the Y chromosome, has been confirmed as sex determining gene [5, 6]. Our result shows the male-dominant expression of *dmrt1* in flounder as reported [25]. These reveal that *dmrt1* may be an important factor for flounder testicular differentiation. Three members of sox gene family including *sox9*, *sox8a*, and *sox6b* were identified to be male predominantly expressed genes. In teleost, *sox9* has been associated with testicular development, with *sox9*-expressing cells differentiating into Sertoli cells [35]. Thestudy on *sox8a* was only reported in orange-spotted grouper (*Epinephelus coioides*), and it was expressed in diverse tissues and increased during testicular developmental stages [36]. No document covered the function of *sox6b*. AMY-1-associating protein was found to be associated with protein kinase A anchor protein 84/149 in the mitochondria in human sperm, suggesting that it plays a role in spermatogenesis [37]. Heat shock protein 90 (HSP90) interacts with steroid hormone receptors, signaling kinases, and various transcription factors [38]. However, the mechanism by which HSP90 interacts with different proteins in various pathways remains unclear. Here in our study, we find *hsp 90α* genehighly expressed in both ovary and testis tissues, while *hsp 90β* gene exhibited upregulated expression pattern in testis. Several proteins associated with sperm mobility show absolute predominance in testis. In human, Ropporin-1-like protein is spermatogenic cell-specific protein that serves as an anchoring protein for the A-kinase anchoring protein, which may function as a regulator of both motility- and head-associated functions such as capacitation and the acrosome reaction [39]. Sperm

Table 3: Genes identified to take part in the ovarian steroidogenesis pathway.

Contigs	Gene function	Protein	*e*-value	log (fold)
Comp1153048_c0_seq1	Luteinizing hormone receptor	LHR	$3E-45$	NA/female
Comp67657_c0_seq1	Guanine nucleotide-binding protein G(s) subunit alpha	Gs	$1E-51$	0.299
Comp127442_c3_seq1	Insulin receptor	INSR	$1E-126$	2.911
Comp123852_c0_seq1	Low-density lipoprotein receptor-related protein 1	LDLR	0	1.607
Comp137524_c1_seq1	Follicle stimulating hormone receptor II	FSHR	0	−1.703
Comp108327_c0_seq1	cAMP-dependent protein kinase catalytic subunit PRKX	PKA	$8E-74$	1.911
Comp138426_c0_seq1	Acyl-protein thioesterase 2	ACOT2	$1E-104$	1.514
Comp1185767_c0_seq1	Cholesterol side chain cleavage cytochrome P450	CYP11A1	$8E-66$	NA/male
Comp945537_c0_seq1	17-alpha-Hydroxylase	CYP17	$9E-93$	1.446
Comp125855_c0_seq1	Steroidogenic acute regulatory protein, mitochondrial	StAR	$1E-65$	−2.111
Comp134314_c0_seq1	P450 aromatase	CYP19A1	0	8.127
Comp123514_c0_seq1	3-beta-Hydroxysteroid dehydrogenase type II	HSD3b	$2E-94$	2.288
Comp134625_c0_seq1	Estradiol 17-beta-dehydrogenase 12-A	HSD17b	$1E-168$	4.625
Comp131665_c2_seq1	Cytochrome P450 1B1	CYP1B1	0	−1.964
Comp123774_c0_seq1	Cytochrome P450 1A	CYP1A1	0	−2.691
Comp126639_c0_seq1	Catechol O-methyltransferase	COMT	$1E-134$	−1.249
Comp507499_c0_seq1	Cytosolic phospholipase A2	CPLA2	$3E-68$	−2.548
Comp80365_c0_seq1	Cytochrome P450 2J2	CYP2J	$1E-116$	0.451
Comp133304_c0_seq1	Scavenger receptor class B member	SCARB1	0	0.919
Comp136599_c2_seq1	Adenylate cyclase type 1	ADCY1	$3E-25$	4.797

Note: NA indicates solely identification in female or male; fold indicates RPKM (female)/RPKM (male).

flagellar protein 2 plays an important role in spermatogenesis and flagellar assembly. And a loss of function mutation in *spef2* of mice caused the big giant head phenotype [40]. Axonemal dynein light intermediate polypeptide may take part in the formation of sperm flagella and play a dynamic role in flagellar motility. As is well known, both cytochrome c oxidase subunit I and ATP synthase F0 subunit 6 are components of the respiratory chain supplying energy for sperm mobility.

4.1.2. Female-Biased Genes. Ovary is the female reproductive system and ovum-producing reproductive organ. It is responsible for the synthesis of estrogen and oogenesis. More and more female-biased genes were found to play important roles in the above processes. In this study, nearly 66.5% of female-biased genes were annotated, of which well-documented ovary markers such as *cyp19a*, *foxl2*, *zp*, cathepsins, and *42sp43* were identified as female-biased genes. *Cyp19a* is responsible for encoding P450 aromatase, critical enzyme catalyzing the process of transforming androgen into estrogen [41]. In teleost, *cyp19a* has been proved to play an important role in sex differentiation and ovarian development, and it is regarded as a reliable early marker of ovarian differentiation [42]. In addition, *foxl2*, encoding the activating transcription factor of *cyp19a*, has been identified to be uniquely expressed in female, which may be one reason for the higher expression level of *cyp19a* in female than in male justified by Yamaguchi et al. [24]. In mammal, *zar1* gene is oocyte-specific maternal-effect gene that functions at the oocyte-to-embryo transition [43]. Wee1-like protein kinase 2 is oocyte-specific protein tyrosine kinase that phosphorylates and inhibits cyclin-dependent kinase 1 (CDK1) and acts as a key regulator of meiosis in *Xenopus* [44]. Histone *H2Ax* is reported to play an important role in chromatin remodeling and associated silencing in male mouse meiosis [45]. However, we found *h2ax* highly expressed in ovary rather than testes in our result, which suggests that it may act on oogenesis of female flounder. The developing oocyte is surrounded by an acellular envelope that is composed of zona pellucida proteins. It is reported that *zpa*, *zpc*, and *zp3* genes and these proteins participate in taxon-specific sperm-egg binding during fertilization process and protect embryo at early developmental stage in mammals [46]. *Zps* were also identified as female-biased genes in this study. Vitellogenesis is the principal event responsible for the enormous growth of oocytes in many teleosts, during which most nutritive products are taken up and stored for developing embryo. Vitellogenin receptor is involved in uptake of vitellogenin by endocytosis. Enzymes such as cathepsins are responsible for the degradation of vitellogenin into yolk protein for storage in the oocyte [47]. P43 5S RNA-binding protein is combined with 5S rRNA to comprise 42S ribonucleoprotein storage particle. In addition, transcription factor IIIA acts as both a positive transcription factor for 5S RNA genes and a specific RNA-binding protein that complex with 5S RNA in oocytes to form the 7S ribonucleoprotein storage particle [48]. According to Diaz De Cerio et al., 5 S RNA and associate protein could constitute a sensitive and universal marker of oogenesis and oocyte differentiation in fish [49].

4.2. Sex Steroids and Biosynthetic Pathway. Gonad is an important organ that is responsible for producing gametes and sex hormones. Steroid hormones are small, hydrophobic hormones that can permeate membranes and therefore

can bind to specific nuclear receptors, including estrogen receptors (ER) and androgen [50]. It can form a complex with a hormone receptor and enter the nucleus where this complex can bind to DNA and result in the translation of specific mRNA and proteins. These can give rise to specific physiological responses including sex differentiation and germ-cell development [51]. In the present study, 20 genes were identified to be involved in flounder ovarian steroidogenesis pathway. In this pathway, all steroids are synthesized from cholesterol, whose transportation from intracellular source into the mitochondria is the rate-limiting step in steroidogenesis [52]. This transmembrane transport process is facilitated by the steroidogenic acute regulatory protein [53]. And the present study identified *star* as male-biased gene.

Male hormone testosterone is formed from pregnenolone by two pathways, delta5 pathway via dehydroepiandrosterone and delta4 pathway via androstenedione. The enzyme P450c17 is responsible for the 17,20-lyase and 17-alpha-hydroxylase activities in respective pathways. Two forms of P450c17 were identified in some teleost, which were P450c17-I and P450c17-II, encoded by *cyp17-I* and *cyp17-II* genes, respectively [54]. P450c17-I possesses both hydroxylase and lyase activities, while P450c17-II only has hydroxylase activity. In medaka, P450c17-I is essential for the production of estradiol-17β (E2) during oocyte growth, while P450c17-II plays a vital role in production of 17α, 20β-dihydroxy-4-pregnen-3-one (17α, 20β-DP) during oocyte maturation [55]. In our study, the transcripts of *cyp17-II* were only found. According to Ding et al., the variation trends of T and E2 level were consistent with the *cyp17-II* expression pattern in flounder ovary [26]. Nevertheless, *cyp19a* still plays a central role in the synthesis of E2. The suppression of *cyp19a* expression at male-inducing temperature leads to masculinization, which further supports the importance of this enzyme and its product (estrogen) in flounder ovarian differentiation [23]. Additionally, 17-beta-hydroxysteroid dehydrogenase (*hsd17b*) and 3-beta-hydroxysteroid dehydrogenase (*hsd3b*) were identified to be female predominantly expressed genes, which indicate their important role in synthesis of estrogen.

The steroidogenic enzymes expression profile may help us to demonstrate the difference of steroid level between male and female. Based on our previous study, both testosterone and estradiol-17β (E2) levels are on rise from ovarian developmental stage I to stage IV. The decrease of E2 level is detected in temperature-induced masculinized groups compared to control groups [56]. During testicular developmental stage, E2 in the serum stay in low level while T level varies significantly during different developmental stage. The variation of sex steroids level may be tightly linked with gonadal development and maturation of germ cells.

5. Conclusion

This is the first report of flounder gonadal transcriptome using RNA-seq technology. We generated a large number of ESTs collection and identified numerous differentially expressed genes between ovary and testis. According to annotation information, sex-related biological pathways including ovarian steroidogenesis were found. The dimorphic expression of overall sex-related genes provides further insights into sexual difference and gonadal development. Our result also provides an archive for further study on molecular mechanism underlying sex determination.

Conflict of Interests

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

Acknowledgments

This study was supported by the National High Technology Research and Development Program of China (863 Programme, no. 2012AA092203), the National Natural Science Foundation of China (no. 41276171), the National Key Basic Program of Science and Technology-Platforms of Aquaculture Stock Resources, and the National Flatfish Industry System Construction Programme (no. nycytx-50-G03).

References

[1] G. S. Helfman, B. B. Collette, D. E. Facey, and B. Bowen, *The Diversity of Fishes. Biology, Evolution, and Ecology*, Blackwell Science, Malden, Mass, USA, 2nd edition, 2009.

[2] J. A. Luckenbach, R. J. Borski, H. V. Daniels, and J. Godwin, "Sex determination in flatfishes: mechanisms and environmental influences," *Seminars in Cell & Developmental Biology*, vol. 20, no. 3, pp. 256–263, 2009.

[3] A. Cnaani and B. Levavi-Sivan, "Sexual development in fish, practical applications for aquaculture," *Sexual Development*, vol. 3, no. 2-3, pp. 164–175, 2009.

[4] N. Ospina-Álvarez and F. Piferrer, "Temperature-dependent sex determination in fish revisited: Prevalence, a single sex ratio response pattern, and possible effects of climate change," *PLoS ONE*, vol. 3, no. 7, Article ID e2837, 2008.

[5] I. Nanda, M. Kondo, U. Hornung et al., "A duplicated copy of DMRT1 in the sex-determining region of the Y chromosome of the medaka, *Oryzias latipes*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 18, pp. 11778–11783, 2002.

[6] M. Matsuda, Y. Nagahama, A. Shinomiya et al., "DMY is a Y-specific DM-domain gene required for male development in the medaka fish," *Nature*, vol. 417, no. 6888, pp. 559–563, 2002.

[7] R. S. Hattori, Y. Murai, M. Oura et al., "A Y-linked anti-Müllerian hormone duplication takes over a critical role in sex determination," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 8, pp. 2955–2959, 2012.

[8] A. Yano, R. Guyomard, B. Nicol et al., "An immune-related gene evolved into the master sex-determining gene in rainbow trout, *Oncorhynchus mykiss*," *Current Biology*, vol. 22, no. 15, pp. 1423–1428, 2012.

[9] T. Myosho, H. Otake, H. Masuyama et al., "Tracing the emergence of a novel sex-determining gene in medaka, *Oryzias luzonensis*," *Genetics*, vol. 191, no. 1, pp. 163–170, 2012.

[10] T. Kamiya, W. Kai, S. Tasumi et al., "A trans-species missense SNP in Amhr2 is associated with sex determination in the tiger Pufferfish, *Takifugu rubripes* (Fugu)," *PLoS Genetics*, vol. 8, no. 7, Article ID e1002798, 2012.

[11] S. Chen, G. Zhang, C. Shao, Q. Huang, and G. Liu, "Whole-genome sequence of a flatfish provides insights into ZW sex chromosome evolution and adaptation to a benthic lifestyle," *Nature Genetics*, vol. 46, pp. 253–260, 2014.

[12] G. Sandra and M. Norma, "Sexual determination and differentiation in teleost fish," *Reviews in Fish Biology and Fisheries*, vol. 20, no. 1, pp. 101–121, 2010.

[13] A. Cutting, J. Chue, and C. A. Smith, "Just how conserved is vertebrate sex determination?" *Developmental Dynamics*, vol. 242, no. 4, pp. 380–387, 2013.

[14] Z. Zhang, Y. Wang, S. Wang et al., "Transcriptome analysis of female and male *Xiphophorus maculatus Jp* 163 A," *PLoS ONE*, vol. 6, no. 4, Article ID e18379, 2011.

[15] M. Salem, C. E. Rexroad III, J. Wang, G. H. Thorgaard, and J. Yao, "Characterization of the rainbow trout transcriptome using Sanger and 454-pyrosequencing approaches," *BMC Genomics*, vol. 11, no. 1, article 564, 2010.

[16] W. Tao, J. Yuan, L. Zhou et al., "Characterization of gonadal transcriptomes from Nile Tilapia (*Oreochromis niloticus*) reveals differentially expressed genes," *PLoS ONE*, vol. 8, no. 5, Article ID e63604, 2013.

[17] L. Sun, C. Wang, L. Huang, M. Wu, and Z. Zuo, "Transcriptome analysis of male and female *Sebastiscus marmoratus*," *PLoS ONE*, vol. 7, no. 11, Article ID e50676, 2012.

[18] F. Sun, S. Liu, X. Gao et al., "Male-biased genes in catfish as revealed by RNA- seq analysis of the testis transriptome," *PLoS ONE*, vol. 8, no. 7, Article ID e68452, 2013.

[19] L. Ribas, B. G. Pardo, C. Fernández et al., "A combined strategy involving Sanger and 454 pyrosequencing increases genomic resources to aid in the management of reproduction, disease control and genetic selection in the turbot (*Scophthalmus maximus*)," *BMC Genomics*, vol. 14, no. 1, article 180, 2013.

[20] E. Yamamoto, "Studies on sex-manipulation and production of cloned populations in hirame, *Paralichthys olivaceus*," *Bulletin of the Tottori Prefectual Fisheries Experimental Station*, vol. 34, pp. 1–145, 1995.

[21] E. Yamamoto, "Studies on sex-manipulation and production of cloned populations in hirame, *Paralichthys olivaceus* (Temminck et Schlegel)," *Aquaculture*, vol. 173, no. 1–4, pp. 235–246, 1999.

[22] P. Sun, F. You, M. Liu et al., "Steroid sex hormone dynamics during estradiol-17β induced gonadal differentiation in *Paralichthys olivaceus* (Teleostei)," *Chinese Journal of Oceanology and Limnology*, vol. 28, no. 2, pp. 254–259, 2010.

[23] T. Kitano, K. Takamune, T. Kobayashi, Y. Nagahama, and S.-I. Abe, "Suppression of P450 aromatase gene expression in sex-reversed males produced by rearing genetically female larvae at a high water temperature during a period of sex differentiation in the Japanese flounder (*Paralichthys olivaceus*)," *Journal of Molecular Endocrinology*, vol. 23, no. 2, pp. 167–176, 1999.

[24] T. Yamaguchi, S. Yamaguchi, T. Hirai, and T. Kitano, "Follicle-stimulating hormone signaling and Foxl2 are involved in transcriptional regulation of aromatase gene during gonadal sex differentiation in Japanese flounder, *Paralichthys olivaceus*," *Biochemical and Biophysical Research Communications*, vol. 359, no. 4, pp. 935–940, 2007.

[25] A. Y. Wen, F. You, and P. Sun, "CpG methylation of dmrt1 and cyp19a promoters in relation to their sexual dimorphic expression in the Japanese flounder *Paralichthys olivaceus*," *Journal of Fish Biology*, vol. 84, no. 1, pp. 193–205, 2014.

[26] Y. Ding, F. He, H. Wen et al., "DNA methylation status of cyp17-II gene correlated with its expression pattern and reproductive endocrinology during ovarian development stages of Japanese flounder (*Paralichthys olivaceus*)," *Gene*, vol. 527, no. 1, pp. 82–88, 2013.

[27] A. Wen, F. You, X. Tan et al., "Expression pattern of dmrt4 from olive flounder (*Paralichthys olivaceus*) in adult gonads and during embryogenesis," *Fish Physiology and Biochemistry*, vol. 35, no. 3, pp. 421–433, 2009.

[28] P. Sun, F. You, L. Zhang et al., "Histological evaluation of gonadal differentiation in olive flouder (*Paralichthys olivaceus*)," *Marine Sciences*, vol. 33, pp. 53–58, 2009.

[29] M. Radonic and G. J. Macchi, "Gonadal sex differentiation in cultured juvenile flounder, *Paralichthys orbignyanus* (Valenciennes, 1839)," *Journal of the World Aquaculture Society*, vol. 40, no. 1, pp. 129–133, 2009.

[30] Z. Z. Chen, C. H. Xue, S. Zhu et al., "GoPipe: streamlined gene ontology annotation for batch anonymous sequences with statistics," *Progress in Biochemistry and Biophysics*, vol. 32, no. 2, pp. 187–191, 2005.

[31] M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, and M. Hirakawa, "KEGG for representation and analysis of molecular networks involving diseases and drugs," *Nucleic Acids Research*, vol. 38, no. 1, pp. D355–D360, 2009.

[32] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-Seq," *Nature Methods*, vol. 5, no. 7, pp. 621–628, 2008.

[33] L. Wang, Z. Feng, X. Wang, and X. Zhang, "DEGseq: an R package for identifying differentially expressed genes from RNA-seq data," *Bioinformatics*, vol. 26, no. 1, pp. 136–138, 2010.

[34] D. S. Wang, L. Y. Zhou, T. Kobayashi et al., "Doublesex- and Mab-3-related transcription factor-1 repression of aromatase transcription, a possible mechanism favoring the male pathway in tilapia," *Endocrinology*, vol. 151, no. 3, pp. 1331–1340, 2010.

[35] J. Kent, S. C. Wheatley, J. E. Andrews, A. H. Sinclair, and P. Koopman, "A male-specific role for SOX9 in vertebrate sex determination," *Development*, vol. 122, no. 9, pp. 2813–2822, 1996.

[36] Q. Liu, H. Lu, L. Zhang, J. Xie, W. Shen, and W. Zhang, "Homologues of sox8 and sox10 in the orange-spotted grouper *Epinephelus coioides*: sequences, expression patterns, and their effects on cyp19a1a promoter activities in vitro," *Comparative Biochemistry and Physiology B*, vol. 163, no. 1, pp. 86–95, 2012.

[37] H. Yukitake, M. Furusawa, T. Taira, S. M. M. Iguchi-Ariga, and H. Ariga, "AAT-1, a novel testis-specific AMY-1-binding protein, forms a quaternary complex with AMY-1, a-kinase anchor protein 84, and a regulatory subunit of cAMP-dependent protein kinase and is phosphorylated by its kinase," *Journal of Biological Chemistry*, vol. 277, no. 47, pp. 45480–45492, 2002.

[38] W. Liu, F. X. Zhang, M. J. Cai et al., "The hormone-dependent function of Hsp90 in the crosstalk between 20-hydroxyecdysone and juvenile hormone signaling pathways in insects is determined by differential phosphorylation and protein interactions," *Biochimica et Biophysica Acta*, vol. 1830, no. 11, pp. 5184–5192, 2013.

[39] D. W. Carr, A. Fujita, C. L. Stentz, G. A. Liberty, G. E. Olson, and S. Narumiya, "Identification of sperm-specific proteins that interact with A-kinase anchoring proteins in a manner similar to the type II regulatory subunit of PKA," *Journal of Biological Chemistry*, vol. 276, no. 20, pp. 17332–17338, 2001.

[40] A. Sironen, J. Hansen, B. Thomsen et al., "Expression of SPEF2 during mouse spermatogenesis and identification of IFT20 as an interacting protein," *Biology of Reproduction*, vol. 82, no. 3, pp. 580–590, 2010.

[41] A. Conley and M. Hinshelwood, "Mammalian aromatases," *Reproduction*, vol. 121, no. 5, pp. 685–695, 2001.

[42] Y. Guiguen, A. Fostier, F. Piferrer, and C.-F. Chang, "Ovarian aromatase and estrogens: a pivotal role for gonadal sex differentiation and sex change in fish," *General and Comparative Endocrinology*, vol. 165, no. 3, pp. 352–366, 2010.

[43] X. Wu, M. M. Viveiros, J. J. Eppig, Y. Bai, S. L. Fitzpatrick, and M. M. Matzuk, "Zygote arrest 1 (Zar1) is a novel maternal-effect gene critical for the oocyte-to-embryo transition," *Nature Genetics*, vol. 33, no. 2, pp. 187–191, 2003.

[44] K. Okamoto, N. Nakajo, and N. Sagata, "The existence of two distinct Wee1 isoforms in Xenopus: implications for the developmental regulation of the cell cycle," *EMBO Journal*, vol. 21, no. 10, pp. 2472–2484, 2002.

[45] O. Fernandez-Capetillo, S. K. Mahadevaiah, A. Celeste et al., "H2AX is required for chromatin remodeling and inactivation of sex chromosomes in male mouse meiosis," *Developmental Cell*, vol. 4, no. 4, pp. 497–508, 2003.

[46] C. Modig, T. Modesto, A. Canario, J. Cerdà, J. von Hofsten, and P. Olsson, "Molecular characterization and expression pattern of zona pellucida proteins in gilthead seabream (*Sparus aurata*)," *Biology of Reproduction*, vol. 75, no. 5, pp. 717–725, 2006.

[47] S. Brooks, C. R. Tyler, and J. P. Sumpter, "Egg quality in fish: what makes a good egg?" *Reviews in Fish Biology and Fisheries*, vol. 7, no. 4, pp. 387–416, 1997.

[48] H. Schneider, M. Dabauvalle, N. Wilken, and U. Scheer, "Visualizing protein interactions involved in the formation of the 42S RNP storage particle of *Xenopus oocytes*," *Biology of the Cell*, vol. 102, no. 8, pp. 469–478, 2010.

[49] O. Diaz De Cerio, I. Rojo-Bartolomé, C. Bizarro, M. Ortiz-Zarragoitia, and I. Cancio, "5S rRNA and accompanying proteins in gonads: powerful markers to identify sex and reproductive endocrine disruption in fish," *Environmental Science and Technology*, vol. 46, no. 14, pp. 7763–7771, 2012.

[50] G. E. Sandra and M. M. Norma, "Sexual determination and differentiation in teleost fish," *Reviews in Fish Biology and Fisheries*, vol. 20, no. 1, pp. 101–121, 2010.

[51] R. H. Devlin and Y. Nagahama, "Sex determination and sex differentiation in fish: an overview of genetic, physiological, and environmental influences," *Aquaculture*, vol. 208, no. 3-4, pp. 191–364, 2002.

[52] M. B. Rone, J. Fan, and V. Papadopoulos, "Cholesterol transport in steroid biosynthesis: Role of protein-protein interactions and implications in disease states," *Biochimica et Biophysica Acta*, vol. 1791, no. 7, pp. 646–658, 2009.

[53] P. R. Manna, M. T. Dyson, and D. M. Stocco, "Regulation of the steroidogenic acute regulatory protein gene expression: present and future perspectives," *Molecular Human Reproduction*, vol. 15, no. 6, pp. 321–333, 2009.

[54] L.-Y. Zhou, D.-S. Wang, T. Kobayashi et al., "A novel type of P450c17 lacking the lyase activity is responsible for C21-steroid biosynthesis in the fish ovary and head kidney," *Endocrinology*, vol. 148, no. 9, pp. 4282–4291, 2007.

[55] L. Y. Zhou, D. S. Wang, Y. Shibata, B. Paul-Prasanth, A. Suzuki, and Y. Nagahama, "Characterization, expression and transcriptional regulation of P450c17-I and -II in the medaka, *Oryzias latipes*," *Biochemical and Biophysical Research Communications*, vol. 362, no. 3, pp. 619–625, 2007.

[56] P. Sun, F. You, D. Ma, J. Li, and P. Zhang, "Sex steroid changes during temperature-induced gonadal differentiation in *Paralichthys olivaceus* (Temminck & Schegel, 1846)," *Journal of Applied Ichthyology*, vol. 29, no. 4, pp. 886–890, 2013.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 346074, 11 pages
http://dx.doi.org/10.1155/2014/346074

Research Article

MAVTgsa: An R Package for Gene Set (Enrichment) Analysis

Chih-Yi Chien,[1] Ching-Wei Chang,[2] Chen-An Tsai,[3] and James J. Chen[2,4]

[1] *Community Medicine Research Center, Keelung Chang Gung Memorial Hospital, No. 200, Lane 208, Jijinyi Road, Anle District, Keelung 204, Taiwan*
[2] *Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, FDA, 3900 NCTR Road, HFT-20, Jefferson, AR 72079, USA*
[3] *Department of Agronomy, National Taiwan University, No. 1, Section 4, Roosevelt Road, Taipei 106, Taiwan*
[4] *Graduate Institute of Biostatistics and Biostatistics Center, China Medical University, Taichung, Taiwan*

Correspondence should be addressed to Chen-An Tsai; catsai@ntu.edu.tw and James J. Chen; jamesj.chen@fda.hhs.gov

Received 2 April 2014; Revised 27 May 2014; Accepted 4 June 2014; Published 3 July 2014

Academic Editor: Tzu-Ya Weng

Gene set analysis methods aim to determine whether an a priori defined set of genes shows statistically significant difference in expression on either categorical or continuous outcomes. Although many methods for gene set analysis have been proposed, a systematic analysis tool for identification of different types of gene set significance modules has not been developed previously. This work presents an R package, called MAVTgsa, which includes three different methods for integrated gene set enrichment analysis. (1) The one-sided OLS (ordinary least squares) test detects coordinated changes of genes in gene set in one direction, either up- or downregulation. (2) The two-sided MANOVA (multivariate analysis variance) detects changes both up- and downregulation for studying two or more experimental conditions. (3) A random forests-based procedure is to identify gene sets that can accurately predict samples from different experimental conditions or are associated with the continuous phenotypes. MAVTgsa computes the P values and FDR (false discovery rate) q-value for all gene sets in the study. Furthermore, MAVTgsa provides several visualization outputs to support and interpret the enrichment results. This package is available online.

1. Introduction

DNA microarray technology enables simultaneous monitoring of the expression level of a large number of genes for a given experimental study. Much initial research on methods for data analysis has focused on the techniques to identify a list of differentially expressed genes. After selection of a list of differentially expressed gene, the list is then examined with biologically predefined gene sets to determine whether any sets are overrepresented in the list compared with the whole list ([1–3]). Mootha et al. [4] proposed gene set enrichment analysis (GSEA), which considers the entire distribution of a predefined gene set rather than a subset from the differential expression list. GSEA provides a direct approach to the analysis of gene sets of interest and the results are relatively easy to interpret. Furthermore, microarray experiments inherit various sources of biological and technical variability, and analysis of a gene set is expected to be more reproducible than an individual gene analysis.

GSEA is a statistical approach to determine whether a functionally related set of genes expresses differently (enrichment and/or deletion) under different experimental conditions. The GSEA approach has inspired the development of various statistical tests for identifying differentially expressed gene sets [5–16]. There are two fundamental hypotheses for GSEA: competitive hypothesis and self-contained hypothesis [8]. The competitive hypothesis tests if the association of a gene set with the phenotype is equal to those of the other gene sets. The self-contained hypothesis tests if the expression of a gene set differs by the experimental condition. In either test, resampling methods are typically used to generate the null distribution of test statistics. The null distributions of statistic under the competitive hypothesis are generated by gene sampling; the null distributions under the self-contained hypothesis are generated by subject-sampling. Various GSEA

statistics have been proposed for testing either the competitive hypothesis or self-contained hypothesis. There are also hybrid-type methods that utilize gene and sample sampling (e.g., [5, 9]). The differences of the two hypotheses were evaluated and summarized in Nam and Kim [14] and Dinu et al. [15]. Gene sampling under the competitive hypothesis simply reassigns the genes into different gene sets. As a result, the sampled distributions are not independent. On the other hand, subject sampling is consistent with the null hypothesis, which the null distributions of statistical tests are identically and independently distributed. In addition, the self-contained hypothesis is a generalization from identification of differentially expressed genes to identification of differentially expressed gene sets. Therefore, we consider self-contained hypothesis and corresponding permutation P values in this paper.

In this paper, a statistical test to determine whether some functionally predefined classes of genes express differently under different experimental conditions is referred to as gene set analysis (GSA). When there are two experimental conditions, a GSA can be a one-sided or two-sided test. A one-sided test is to detect the changes of gene expressions in the gene set in one direction, either up- or downregulation. The two-sided test is to detect the changes in both up- and downregulation [6, 15]. When the goal is to detect coordinated changes in one direction, the one-sided hypothesis is appropriate. However, in an exploratory context, it is impossible to prespecify how individual genes in a gene set will respond in different conditions. Hence, two-sided hypothesis is generally suggested.

The original GSEA statistic was developed for a one-sided test to identify downregulated genes in type 2 diabetes mellitus subjects [4]. Other one-sided tests included T-like statistic [8]; MaxMean statistic [9]; standardized weighted sum statistic [10]; and OLS statistics [6]. Chen et al. [6] showed that OLS statistic performed well for GSA statistics in one-sided test. Two-sided tests were considered in most GSA statistics (e.g., [7, 11–17]). Tsai and Chen [16] proposed a multivariate analysis of variance (MANOVA) GSA test for two or more conditions. The MANOVA approach was compared with principal component analysis [13], SAM-GS [15], GSEA [5], MaxMean [9], analysis of covariance [7], and Goeman's global test [11]. They found that MANOVA test performed the best in terms of control of type I error and power.

The GSEA software only performs one-sided test in two conditions at a time [4]. Another software provides either one-sided or two-sided test; for example, BRB-Array tools [18] provided LS (KS) statistics, Goeman's global test, and MaxMean test. When the null hypothesis is that no genes in the gene set are differentially expressed, the two-sided is more appropriate than other methods.

When the purpose of study is to build up prediction rule based on gene expression profile, utilizing the existing biological knowledge, such as biological pathway or cellular function information, has been showed to improve the classification accuracy (e.g., [19–21]). The selection of predictors can be either preselected by testing or chosen based on classification algorithm, and the final identified predictors are also considered as differential expressed. Hsueh et al. [22] and Pang et al. [23] proposed a random forests-based differential analysis of gene set data in terms of predictive performance of gene sets. The analysis contained not only a classifier but also the feature importance of the input gene sets.

The gene set analysis has been considered to accommodate continuous phenotype. Linear combination test [17] has been extended from binary to continuous phenotype (LCT) by utilizing linear regression function [24]. Significance analysis of microarrays for gene sets (SAM-GS) [15] and global test [11] have also been extended in a generalized linear model (GLM) framework. Dinu et al. [24] compared the type I error rates and powers for the three methods and concluded that LCT approach is powerful and computationally attractive. The random forests-based approach could be also applied to continuous phenotype by modifying random forest classification to random forest regression. A small simulation experiment to compare the random forests-based and LCT is reported in this paper.

The main purpose of this paper is to present the MAVTgsa R package tool for GSA for study with categorical phenotypes or continuous phenotypes. The OLS one-sided test, MANOVA test, and random forest-based analysis are implemented in MAVTgsa. When categorical phenotypes involve more than two classes, the three multiple comparison procedures are implemented: Dunnett, Tukey, and sequential pairwise comparison. The program provides two visualization plots: GSA plot, a P value plot of GSA for all gene sets in the study, and GST, a plot of the empirical distribution function of the ranked test statistics of a selected gene set. The researcher is able to summarize and visualize the gene expression data in gene set analysis. More importantly, the adjusted P value of family-wise error rate (FWER) [25] and FDR step-up procedure [26] are computed in this package. When the outcome of interest is a continuous phenotype, the random-forests based analysis is applied in the regression context. Figure 1 describes the procedure of how to implement the MAVTgsa package for gene set enrichment analysis.

Section 2 describes the methods implemented in the MAVTgsa package. In Section 3, we present a description of the MAVTgsa package including data input, optional parameters, output, and result visualization. In Section 4, we apply to two real data sets, P53, and breast cancer, for illustration. This paper is concluded by a summary.

2. Methods

2.1. The OLS Test. The OLS test was developed to detect changes in one direction, either up- or downregulation, for a study with two experimental conditions. Consider a gene set consisting of $\mathbf{m}$ genes with two conditions of sample size $\mathbf{n_1}$ and $\mathbf{n_2}$. Let $y_{ij} = (y_{ij1}, \ldots, y_{ijm})$ be the m-vector of intensities for simple j ($j = 1, \ldots, n_i$) in ith condition ($i = 1, 2$). Denote the standardized variable $y^*_{ijk} = (y_{ijk} - \overline{y}_k)/s_k$, where $\overline{y}_k$ is the overall sample mean for the kth gene and s_k is the pooled standard deviation. Let $z_{ik} = \Sigma_j y^*_{ijk}/n_i$ be the mean of the standardized variable for the kth in the ith condition. Denote the $\mathbf{z}_i = (z_{i1}, \ldots, z_{im})$ as the m-vector of the standardized

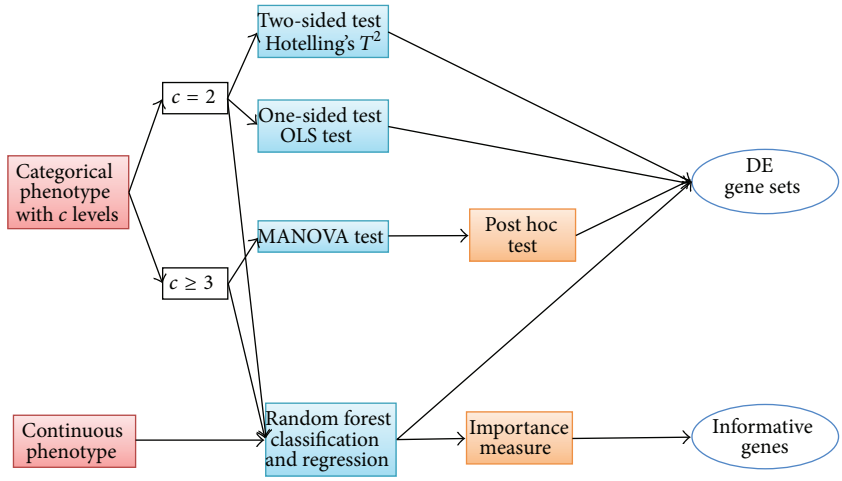


Figure 1: A schematic flowchart of GSEA using the MAVTgsa package.

mean variable z_{ij}'s for the ith condition ($i = 1, 2$). Let $\mathbf{d}_i$ be an m-dimensional vector for the mean difference between two conditions $\mathbf{d}_i = (\mathbf{z}_1 - \mathbf{z}_2)$. The O'Brien's OLS statistic [27] is

$$T_{\text{ols}} = \frac{\mathbf{1}'\mathbf{d}_i}{(\mathbf{1}'\mathbf{V}_i\mathbf{1})^{1/2}}, \tag{1}$$

where $\mathbf{1}$ is a $m_i \times 1$ vector of $\mathbf{1}$'s and $\mathbf{V}_i$ is the pooled sample covariance matrix of $\mathbf{d}_i$. If $\mathbf{d}_i$ is a multivariate normal, then the OLS statistic T_{ols} has an approximately t distribution with $n_1 + n_2 - 2$ degrees of freedom.

The one-sided OLS statistic is the most widely used global test for the analysis of multiple clinical endpoints [27]. This test is very powerful when the changes in the gene expression are in the same direction. The direction of changes of significant gene set can be checked from the OLS statistic The OLS statistics account for gene set size and correlation structure [6]. In addition, the permutation P values for the OLS test are suggested to compute as the level of significance.

2.2. The MANOVA Test. Consider a gene set consisting of $\mathbf{m}$ genes with $\mathbf{c}$ conditions of sample size $\mathbf{n_1}, \ldots, \mathbf{n_c}$. Let $y_{ij} = (y_{ij1}, \ldots y_{ijm})$ be the m-vector of intensities for simple j ($j = 1, \ldots, n_i$) in ith condition ($i = 1, \ldots, c$). The MANOVA model [16] can be expressed as $y_{ij} = \mu_i + e_{ij}$, where e_{ij} is m-vector of residuals with $\text{Var}(e_{ij}) = \Sigma$ and μ_i is the m-vector of means for the ith condition. MAVTgsa occupied the Wilks' Λ as the statistic for MANOVA test. The formula of Wilks' Λ is given as

$$\Lambda = \Pi \frac{1}{(1 + \lambda_k)}, \tag{2}$$

where λ_k's are the eigenvalues of the matrix S ($= E^{-1}H$), and E is within sum of squares matrix (sample covariance matrix) and H is between sum of squares matrix. The number of eigenvalues k is equal to the minimum of the number of genes ($\mathbf{m}$) and the number of conditions minus 1 ($\mathbf{c} - 1$). When the number of genes in the gene set is greater than the number of samples, the matrix E is singular and ill-conditioned. The shrinkage covariance matrix estimator (s^*_{ij}) proposed by Schafer and Strimmer [28] is applied to estimate the sample covariance matrix and given as

$$s^*_{ij} = \begin{cases} s_{ii} & \text{if } i = j \\ r^*_{ij}\sqrt{s_{ii}s_{jj}} & \text{if } i \neq j \end{cases} \tag{3}$$

and $r^*_{ij} = r_{ij} \min\{1, \max(0, 1 - \widehat{\lambda}^*)\}$, where s_{ii} and r_{ij}, respectively, denote the empirical sample variance and sample correlation, and the optimal shrinkage intensity $\widehat{\lambda}^*$ is estimated by

$$\widehat{\lambda}^* = \frac{\sum_{i \neq j} \widehat{\text{Var}}\left(r_{ij}\right)}{\sum_{i \neq j} r^2_{ij}}. \tag{4}$$

The distribution of Wilks' Λ under the null hypothesis of no difference in responses among the conditions was estimated by the permutation method. The Wilks' Λ test is equivalent to Hotelling's T^2 test when there are only two conditions. The Hotelling's T^2 statistic is

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}\left(\overline{x}_1 - \overline{x}_2\right)^t S_p^{-1}\left(\overline{x}_1 - \overline{x}_2\right), \tag{5}$$

where $\overline{x}_i$ and S_i denote the sample mean vector and sample covariance matrix of the ith group ($i = 1, 2$), respectively, and $S_p = ((n_1 - 1)S_1 + (n_2 - 1)S_2)/(n_1 + n_2 - 2)$ denotes the pooled covariance matrix. Using the shrinkage covariance matrix and permutation test, the P values of the Hotelling's T^2 test are computed. The P values of the MANOVA test are computed using the permutation method.

The MANOVA test is a multivariate generalization of the univariate analysis of variance (ANOVA) as Hotelling's T^2 test is the generalization of the univariate t-test. The parametric MANOVA and T^2 test statistics are commonly used for analyzing multivariate data. The MANOVA test uses a shrinkage covariance matrix estimator [28] to incorporate

the correlations structure among the genes in the test statistic and account for the singularity and ill-condition of the sample covariance matrix. The MANOVA test uses the permutation method to compute the P values. Tsai and Chen [16] have compared the MANOVA test with several GSEA tests including the two one-sided tests: GSEA [5] and MaxMean [9], and four two-sided tests: ANCOVA [7], Global [11], PCA [13], and SAM-GS [15]. Their simulation showed that MANOVA test performs well in terms of controlling the type I error and power as compared to other tests, except that the ANCOVA [7] was more powerful when the variances were equal across all genes in the gene set.

2.3. Random-Forests Based Analysis. The random forests (RF) [29] is a popular classification and regression algorithm based on an ensemble of many classification/regression trees combined with a bootstrap sample out of the original dataset. The error rate of a classification tree or the mean of squared error (MSE) of a regression tree is calculated by the observations outside the bootstrap sample called the out-of-bag (OOB) data. Lower error rate and MSE indicates that the corresponding gene set is with higher prediction accuracy for a categorical phenotype and higher percent variance explained for a continuous phenotype, respectively. The variable selection for each split at the nodes of trees is conducted only from a small random subset of the predictors, so that there is no need to deal with the "small n large p" problem. A permutation based P value can be obtained as

$$P\text{-value} = \frac{\sum_{k=1}^{N} I\left\{R^{(k)} \leq R_0\right\}}{N}, \tag{6}$$

where R_0 is the observed score from the random forests analysis, $R^{(k)}$ is the score in the kth permutation, and N is the total number of permutations. The score is calculated as the OOB error rate and MSE for the categorical and continuous phenotypes, respectively. The gene set is considered as significantly expressed if the correspondent P value is less than or equal to the significance level α. To measure the importance of each predictor variable, the random forests algorithm assesses the importance of a variable by looking at how much prediction error increases or MSE decreases when that variable of the (OOB) data is permuted, while remaining variables are left unchanged. Here, the mean decrease in the Gini index (MDG) and the mean decrease in MSE are employed as the importance measures of genes for categorical and continuous phenotypes, respectively.

2.4. Multiple Comparisons Methods. In MANOVA test, the null hypothesis is rejected if one or more of the mean differences or some combination of mean differences among the genes in gene set differs from zero. If the null hypothesis is rejected, three multiple comparison procedures are implemented to identify which two conditions differ in expression between gene sets for studies with more than two conditions. Three multiple comparisons methods are Dunnett, Tukey, and sequential pairwise comparison. Dunnett test is specifically designed for situations where all groups are to be compared against the "reference" group. Tukey test is for all pairwise comparisons. When the conditions are in order, sequential pairwise comparison is appropriate to be occupied for comparing with previous condition.

3. MAVTGSA Package Description

MAVTgsa is a software tool to evaluate the expressions of a priori defined gene sets under different experimental conditions. The package implements essentially the method described in the previous section and the main functions are *MAVTn()* and *MAVTp()*.

3.1. MAVTn. For experiments with two or more than two conditions, the input parameters to perform the testing are described as follows.

MAVTn(DATA, GS, Alpha, nbPerm, MCP)

(i) *DATA* is a gene expression data matrix with samples in columns. The first row contains the information of the experimental condition of each sample. The remaining rows contain gene expression.

(ii) *GS* is a binary matrix coded 0 or 1 with genes in rows. Each column represents a predefined gene set, with row equal to 1 indicating that the gene is in the gene set, and 0 otherwise.

(iii) *Alpha* is the significance level.

(iv) *nbPerm* specifies the number of permutations, and at least 5,000 is recommended.

(v) *MCP* specifies one of three multiple comparison methods, Dunnett = 1, Tukey = 2, and sequential pairwise comparison = 3. MCP can be ignored when the number of experimental conditions is 2.

The output of the *MAVTn()* function is a list of objects which contains the following.

(i) *P value:* a list of the gene sizes and the P values and adjusted P values of the statistical tests for GSA. The adjusted P values of the family-wise error rate (FWE) [25] and the FDR step-up procedure [26] are computed using permutation method. In case of two experimental conditions, the P values and adjusted P values for OLS and Hotelling's T^2 test are listed. In case of more than two conditions, the P values for MANOVA and *post hoc* tests are listed.

(ii) *Significant gene set:* a list of the P values of individual genes for those significant gene sets in statistical test of GSA.

In addition, a plot of P values for all gene sets is drawn to examine the adequacy of the assumptions on which the distributions of the test statistics are based.

Example of applying MAVTn to a simulated 3 conditions data is as follows:

```
R> mu <- c(rep(1,5),rep(5,5))
R> Sigma <- diag(1,10)
```

```
R> set.seed (10) # fix random seed
R> GE <− rmvnorm (15,mu,Sigma) # generate 15 samples with 10 gene expression
R > set.seed(10) # fix random seed
R > GS <− matrix (1∗(runif (40,0,1)<=0.7), ncol=4,nrow=10) # simulate gene set matrix
R> cl <− c(rep(1,5),rep(2,5),rep(3,5)) #clinical treatment
R> data <− rbind(cl,t(GE))
R> MAVTn(data,GS,MCP = 1).
```

3.2. MAVTp. For experiments with categorical or continuous phenotypes, the input parameters to perform the random forests analysis are described as follows.

MAVTp(DATA, GS, nbPerm, Numoftree, Type, Impt)

(i) *DATA* is a gene expression data matrix with samples in columns. The first row contains the information of the experimental condition of each sample. The remaining rows contain gene expression. If the first row is a factor, RF classification is assumed, otherwise RF regression is assumed.

(ii) *GS* could be a binary matrix coded 0 or 1 with genes in rows. Each column represents a predefined gene set, with row equal to 1 indicating that the gene is in the gene set, and 0 otherwise.

(iii) *nbPerm* specifies the number of permutations.

(iv) *Numoftree* specifies the number of trees to grow.

(v) *Type* indicates categorical or continuous phenotype.

(vi) *Impt* is an option for outputting the important measure.

The output of the *MAVTp()* function is a list of objects which contains the following.

(i) *P value:* a list of the *P* values of the random forests for GSA.

(ii) *Important gene set:* a list of the importance measurement of individual genes for those significant gene sets.

Example of applying MAVTp to a simulated continuous phenotype data is as follows:

```
R> mu <− c(rep(1,5),rep(5,5))
R> Sigma <− diag(1,10)
R> set.seed(15) # fix random seed
R> GE <− rmvnorm(15,mu,Sigma) # generate 15 samples with 10 gene expression
R> y = mvrnorm (1,rep(0,10),diag(rep(1,10))) # generate continuous phenotype
R> data <− rbind(y,GE)
R> GS <− matrix (1∗(runif (30,0,1)<=0.7),ncol = 3,nrow=10) # simulate gene set matrix
R> test_rf_con <− MAVTp(data,GS,nbPerm=1000, numoftree=500,type="cont",impt= TRUE).
```

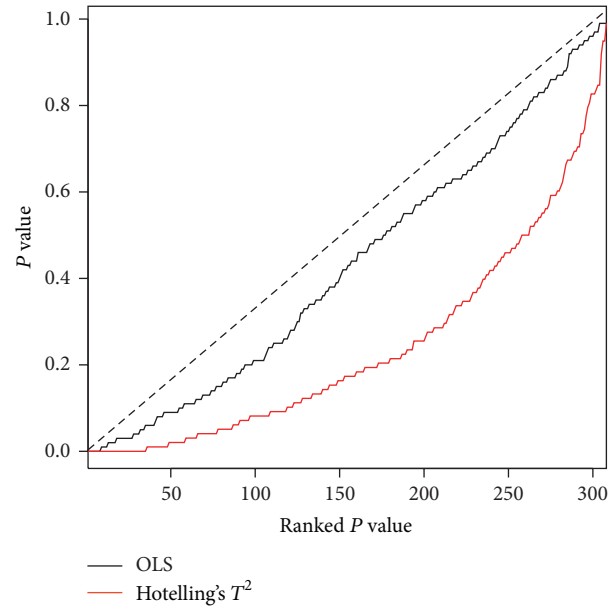


FIGURE 2: The GSA-plot for OLS test and Hotelling's T^2 test in P53 study. Both *P* values in OLS test (black line) and T^2 test (red line) are not close to the diagonal dash line. That means both tests could identify that several gene sets showed truly significant of the testing hypotheses.

3.3. Visualization Plots. Two visualization plots are implemented, *GSA* and *GST*. The *GSA* plot (Figure 2) provides a *P* value plot of all gene sets considered in the study. The *P* value plot is the plot of ordered *P* value versus its rank. The *P* value plot can provide an overall assessment of differences in expression among conditions for all gene sets considered in the study. Under the null hypothesis of no differences, the *P* values should be uniformly distributed on the interval (0, 1); the *P* value plot should be a straight line. If a null hypothesis is not true, then its *P* value will tend to be small. The *GST* plot displays the relative direction (in two conditions) and statistics ranking for genes in a gene set. The *GST* plot is derived from the SAFE plot which provides the empirical distribution function for the ranked statistics of a given gene category [30]. The *GST* plot displays the ranked test statistics (red line) and empirical cumulative distribution function of these test statistics for expressed genes in a gene set (solid line). Tick marks above the plot display the location of genes with gene names. The shaded regions are set to represent the statistics that the *P* values below the given alpha value (Figures 3–5). The input parameters to perform the analysis are described as follows. Two examples are presented in next section to show the output of *GST* plot.

4. Results

4.1. P53 Study. The MAVTgsa was applied to a P53 dataset. The P53 dataset is from a study to identify targets of the transcription factor P53 from 10,100 gene expression profiles in the NCI-60 collection of cancer cell lines. There are 308

Table 1: The original P values and adjusted P values (FDR and FWE) of the OLS and Hotelling's T^2 test of P53 data.

GS name	GS size	OLS			T^2		
		P value	FDR	FWE	P value	FDR	FWE
hsp27Pathway	15	0*	0	0	0.0001*	0.0015	0.0003
p53hypoxiaPathway	20	0*	0	0	0*	0	0
p53Pathway	16	0*	0	0	0*	0	0
P53_UP	40	0*	0	0	0*	0	0
Radiation_sensitivity	26	0.0001*	0.0062	0.0741	0*	0	0
rasPathway	22	0.0015*	0.0770	0.3618	0.0989	0.2621	0.0576
HTERT_DOWN	64	0.0029*	0.1276	0.5490	0.0383	0.1636	0.1239
ST_Interleukin_4_Pathway	24	0.0079*	0.3042	0.2743	0.0039*	0.0375	0.1799
ck1Pathway	15	0.0089*	0.3046	0.3107	0.0068*	0.0582	0.1486
ngfPathway	19	0.0105	0.3102	0.2071	0.1313	0.2952	0.0862
inflamPathway	28	0.0133	0.3102	0.0737	0.0308	0.1492	0.0674
lairPathway	15	0.0134	0.3102	0.1119	0.0393	0.1636	0.2736
no2il12Pathway	17	0.0136	0.3102	0.2248	0.0245	0.1417	0.0549
cytokinePathway	21	0.0141	0.3102	0.2706	0.0290	0.1492	0.0217
badPathway	21	0.0157	0.3224	0.4702	0.0001*	0.0015	0

*Denote P values < 0.01.

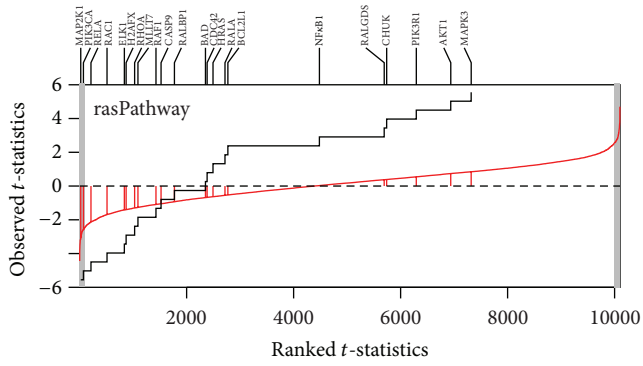


Figure 3: GST-plot for the gene set rasPathway in P53 dataset. The solid line is the empirical cumulative distribution function of the rank t-statistics for 10,100 genes in the array. The two-tailed shaded regions represent the t-statistics that had the P value less than 0.01. There are 22 tick marks above the plot which display the location of the P value of the genes from the gene set. The gene set shows underexpressed.

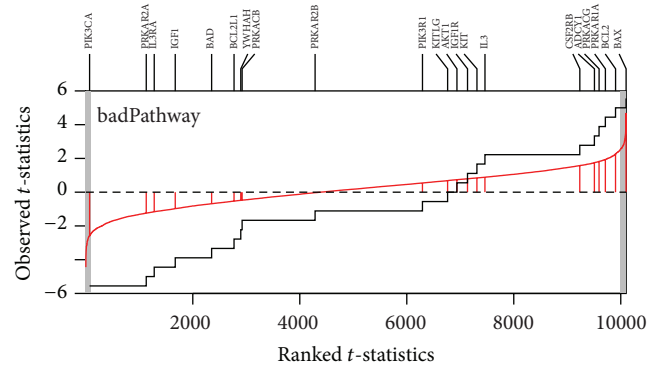


Figure 4: GST-plot for the gene set badPathway in P53 dataset. The solid line is the empirical cumulative distribution function of the rank t-statistics for 10,100 genes in the array. The two-tailed shaded regions represent the t-statistics that had the P value less than 0.01. There are 21 tick marks above the plot which display the location of the P value of the genes from the gene set. The gene set shows both under- and overexpressed.

gene sets in the P53 study. The mutation status of the P53 gene has been reported for 50 cell lines included 17 wild-type and 33 mutation samples. The dataset is publicly available at the GSEA website (http://www.broad.mit.edu/gsea/). Table 1 shows the result of fifteen gene sets in which the P values for OLS test are the top fifteen smallest P values under 10,000 permutations ($nbPerm = 10,000$). The significant level is set as 0.01 ($alpha = 0.01$). The P values for OLS test and MANOVA (Hotelling's T^2) were calculated and listed the P values of individual genes for those significant gene sets. The P values from the OLS test are different from the P values from two-sided T^2. The gene set rasPathway, HTERT_DOWN, and ngfPathway are highly significant ($P < 0.01$) by OLS test, but they are not significant in Hotelling's T^2 test. In contrary, the gene set badPathway is highly significant in Hotelling's T^2 test but not in OLS test. Figure 2 is the GSA-plot for two groups. Figures 3 and 4 are the GST plots for the gene set rasPathway and badPathway, respectively. In Figure 3, two of 22 genes are underexpressed with the P value less than 0.01. On the other hand, one of the 21 genes in badPathway shows underexpressed and one shows overexpressed in Figure 4. The results indicate that the power of methods to detect differential expressed gene set depends on the global pattern of genes within gene set. Combining the information from the two tests and GSA-plot will be useful to get biological meaningful interpretation.

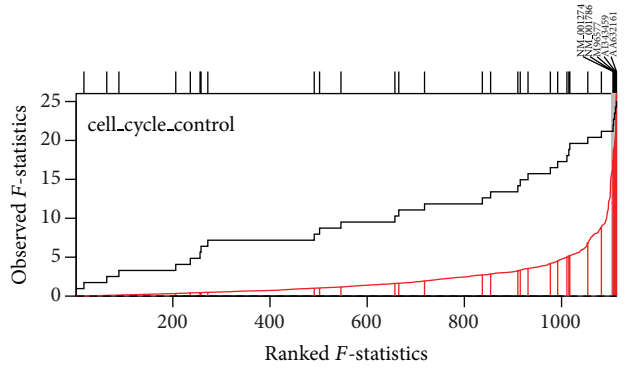


Figure 5: GST-plot for the gene set cell_cycle_control in breast cancer dataset. The solid line is the empirical cumulative distribution function of the rank F-statistics for 1,113 genes in the array. The shaded regions represent the F-statistics that had the P value less than 0.01. There are 5 tick marks above the plot which display the location of the P value of the genes from the gene set.

4.2. Breast Cancer Dataset. We applied the MAVTgsa to a breast cancer dataset [31] and illustrate the RF, MANOVA, and a multiple comparison analysis. The dataset consisted of three conditions with 1,113 genes and 96 samples. Three conditions were tumor grades 1, 2, and 3 with the sample sizes of 11, 25, and 60, respectively. There were nine cancer related pathways for gene set analysis. Table 2 shows the P values of the MANOVA and the post hoc Dunnett's analysis for each of the nine pathways. Figure 5 is the GST plots for the 31 genes in the gene set cell_cycle_control. Five of the 31 genes are differential expressed with P value less than 0.01 in the breast cancer samples. In this analysis the total computation time used to perform the analysis was approximately four hours and 10 minutes with nbPerm = 10,000. A classification rule to classify the three tumor grades is also constructed. The error rates of 10-fold cross-validation are given in Table 2.

4.3. Simulation Study. In order to understand how well the random forests-based analysis performs for continuous phenotypes, we conducted a number of simulations to evaluate the performance in terms of type I error and power and compare to the LCT method [24]. The simulation design was similar to that considered by Dinu et al. [24]. For each gene set of size p 24, we generated a p by n gene expression matrix $X_{n\times p}$ with sample size of n and a linear model was used to simulate the phenotype data associated with the gene set X. The gene expressions matrix X was generated from a multivariate normal distribution with mean from a uniform (0, 10) distribution, variance from a uniform (1, 5) distribution, and a mixed intragene set correlation structure as follows:

$$\rho_{ij} = \begin{cases} \rho, & 1 \le i \ne j \le p_1; \\ \rho^{|i-j|}, & p_1 + 1 \le i \ne j \le 2p_1; \\ 0, & \text{otherwise}, \end{cases} \quad (7)$$

where the correlation ρ was set at 0, 0.3, 0.5, and 0.9 in this study. For each gene set, a continuous phenotype vector $Y_{n\times 1}$ is generated from a multivariate normal distribution MVN($X\mu$, $\mathbf{I}$), where μ is the effect size vector with length p and $\mathbf{I}$ the identity covariance matrix. In the null model, with no association of gene expressions on the phenotype, we set μ to be 0 to investigate the type I error and the simulation scenarios varied according to sample sizes (n = 10, 20, or 50), gene set sizes (p = 20, 100, or 200), and the levels of correlation among genes (ρ = 0, 0.3, 0.5, or 0.9) within a gene set at different number of genes (p_1 = 5, 20, or 40). In the alternative model, the sample size and gene set size were, respectively, fixed at 20 and 100, as well as only 10 of genes were simulated to be associated with the phenotype to investigate power of two gene set analysis methods. First, we randomly generated 5 of the first 20 components of μ from normal distribution $N(\nu, |\nu|)$ and another five of the next 20 components of μ from normal distribution $N(-\nu, |\nu|)$ as up- and downregulated genes, respectively. The rest of components of μ were set to be 0. The effect sizes (ν) for the associations were set at 0.2, 0.6, 1.0, 1.4, and 1.8. For each scenario, the simulation data were replicated 1000 times to estimate type I error rate or power. The P values were based on 1000 permutations. Power was then estimated as the proportion of significance using the nominal level of 0.05. Table 3 showed the empirical type I errors using the nominal level of 0.05 for each scenario. The type I errors from the random forests method were reasonably close to or below the nominal level, while the LCT method appeared to have an inflated type I error rate in most cases. It indicates that the RF method gives a conservative conclusion. Such conservativeness may lead to power loss in detecting a difference. Figure 6 illustrated the empirical powers using the nominal level of 0.05 for ρ = 0.0, 0.3, 0.6, and 0.9. As expected, the RF method was slightly inferior to the LCT method, while LCT was unable to adequately control the type I error rate. However, with increasing of correlations among genes, both methods appeared to be equivalent. The powers of both methods increased gradually with increasing correlations.

In addition, to explore the robustness of the RF method with regard to nonlinear association data, the continuous phenotype vector $Y_{n\times 1}$ was generated from a multivariate normal distribution MVN(exp($X\mu$), $\mathbf{I}$). Figure 7 showed the average power over 1000 simulations for each method using the nominal level of 0.05. As a result, the RF provided a more powerful test than the LCT method to detect the non-linear association between gene sets and continuous phenotypes.

5. Conclusion

The MAVTgsa package performs a systematic gene set analysis for identification of different types of gene set significance modules. The user can select the most appropriate analysis or combine them to provide insight into gene sets that respond in a similar manner to varying phenotypes and that might therefore be coregulated. For studies with more than two conditions, MAVTgsa not only provides the MANOVA test to identify gene sets consisting of differentially expressed genes but also implements three multiple comparison methods for post hoc analysis. The method implemented in MAVTgsa

Table 2: Results of the MANOVA, 10-fold CV error, and post hoc test (Dunnett's method) of breast cancer data with three groups for nine pathways.

GS name	GS size	MANOVA P-value	10-fold CV Error rate	adjusted P value FDR	adjusted P value FWE	Multiple comparisons $T_2 - T_1$	Multiple comparisons $T_3 - T_1$
androgen_receptor_signaling	72	0.0000*	0.3438	0	0	0.2088	0.0027
apoptosis	187	0.0031*	0.3542	0.0047	0.1751	0.5836	0.0517
cell_cycle_control	31	0.0000*	0.3333	0	0	0.3823	0.0002
notch_delta_signalling	34	0.0040*	0.3750	0.0051	0.2179	0.4558	0.0468
P53_signalling	33	0.0001*	0.2917	0.0002	0.0078	0.4891	0.0029
ras_signalling	266	0.0049*	0.3542	0.0055	0.5569	0.9132	0.0239
tgf_beta_signaling	82	0.0564	0.3438	0.0564	0.4885	0.6840	0.0664
tight_junction_signaling	326	0.0001*	0.3854	0.0002	0.0294	0.2937	0.0058
wnt_signaling	176	0.0004*	0.3542	0.0007	0.3299	0.6117	0.0037

* Denote P values < 0.01 in Wilks' Λ test.

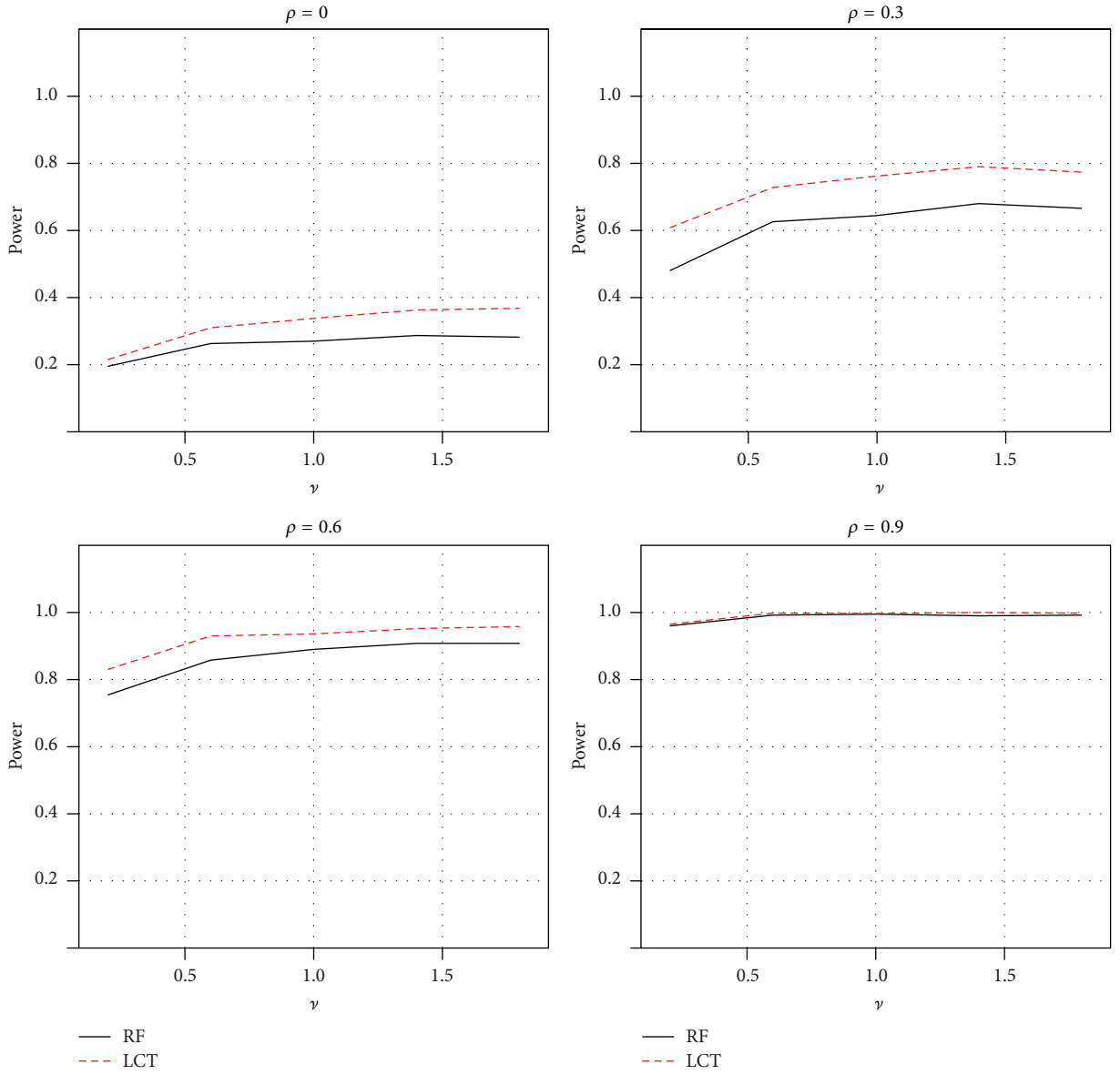


Figure 6: Power comparisons of two GSA methods for a linear association between gene sets and continuous phenotypes: random forests and LCT.

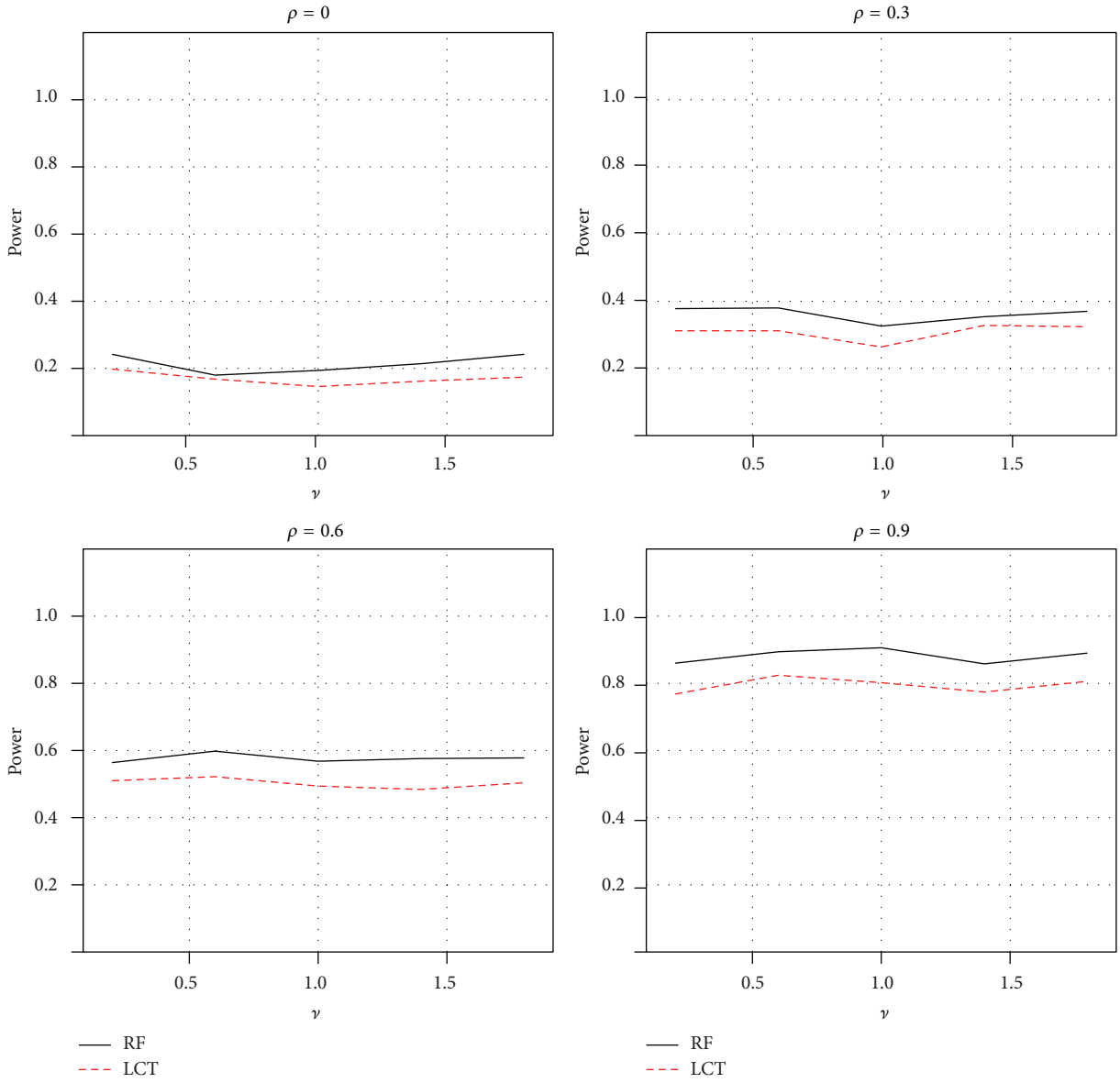


Figure 7: Power comparisons of two GSA methods for a nonlinear association between gene sets and continuous phenotypes: random forests and LCT.

package has the advantage of real application. First, the MAVTgsa package provides the adjusted P values using both family-wise error rate (FWER) [25] and the FDR step-up procedure [26] using permutation method. Second, MAVTgsa is a permutation-based method to compute P values and adjusted P values. Third, MAVTgsa displays the results for individual genes test of significant gene sets. Finally, MAVTgsa draws the GST plot to display the empirical cumulative function for the ranked test statistics of a given gene set with the gene location and gene name above the plot. These allow the user to analyze the interesting gene set of the data easily. In addition, MAVTgsa provides a random forests-based procedure to identify gene sets in terms of predictive performance or in association with the continuous phenotypes. Random forests method has been proved to perform well in comparison with the other classification methods and successfully applied to various problems. Most importantly, it can accommodate multiclass and continuous phenotypes for the GSA, even if the associations between gene sets and phenotypes are nonlinear and involve complex high-order interaction effects.

6. Hard Ware and Software Specifications

The implementation and examples run of this package were conducted on a laptop computer with 2.8 GHz CPU and 3.0 GB RAM under the Microsoft Windows XP Professional SP3 using the R software version 2.14.1.

Table 3: Type I error rate comparisons of two GSA methods for continuous phenotype, RF, and LCT, at a significance level of 0.05.

Method	$\rho = 0.0$	$\rho = 0.3$	$\rho = 0.6$	$\rho = 0.9$
$n = 10$, $p = 20$, $p_1 = 5$				
RF	0.050	0.049	0.046	0.036
LCT	0.052	0.060	0.044	0.052
$n = 20$, $p = 100$, $p_1 = 20$				
RF	0.051	0.042	0.049	0.052
LCT	0.060	0.051	0.050	0.064
$n = 50$, $p = 200$, $p_1 = 40$				
RF	0.052	0.040	0.040	0.052
LCT	0.060	0.044	0.048	0.066

7. Availability

The *MAVTgsa* package is available from http://cran.r-project.org/web/packages/MAVTgsa/.

Disclaimer

The views presented in this paper are those of the authors and do not necessarily represent those of the U.S. Food and Drug Administration.

Conflict of Interests

The authors declare that they have not competing interests.

Authors' Contribution

Chih-Yi Chien and Ching-Wei Chang contributed equally to this work.

Acknowledgments

Chih-Yi's research was supported by the Post-doctoral Fellowship Program at the NCTR administered by the Oak Ridge Institute for Science and Education through an interagency agreement between USDOE and USFDA.

References

[1] S. Drăghici, P. Khatri, R. P. Martins, G. C. Ostermeier, and S. A. Krawetz, "Global functional profiling of gene expression," *Genomics*, vol. 81, no. 2, pp. 98–104, 2003.

[2] P. Khatri and S. Drăghici, "Ontological analysis of gene expression data: current tools, limitations, and open problems," *Bioinformatics*, vol. 21, no. 18, pp. 3587–3595, 2005.

[3] I. Rivals, L. Personnaz, L. Taing, and M. C. Potier, "Enrichment or depletion of a GO category within a class of genes: which test?" *Bioinformatics*, vol. 23, no. 4, pp. 401–407, 2007.

[4] V. K. Mootha, C. M. Lindgren, K. Eriksson et al., "PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes," *Nature Genetics*, vol. 34, no. 3, pp. 267–273, 2003.

[5] A. Subramanian, P. Tamayo, V. K. Mootha et al., "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 43, pp. 15545–15550, 2005.

[6] J. J. Chen, T. Lee, R. R. Delongchamp, T. Chen, and C.-A. Tsai, "Significance analysis of groups of genes in expression profiling studies," *Bioinformatics*, vol. 23, no. 16, pp. 2104–2112, 2007.

[7] U. Mansmann and R. Meister, "Testing differential gene expression in functional groups: goeman's global test versus an ANCOVA approach," *Methods of Information in Medicine*, vol. 44, no. 3, pp. 449–453, 2005.

[8] L. Tian, S. A. Greenberg, S. W. Kong, J. Altschuler, I. S. Kohane, and P. J. Park, "Discovering statistically significant pathways in expression profiling studies," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 38, pp. 13544–13549, 2005.

[9] B. Efron and R. Tibshirani, "On testing the significance of sets of genes," *The Annals of Applied Statistics*, vol. 1, no. 1, pp. 107–129, 2007.

[10] J. Lauter, "Exact t and F tests for analyzing studies with multiple endpoints," *Biometrics*, vol. 52, no. 3, pp. 964–970, 1996.

[11] J. J. Goeman, S. A. van de Geer, F. de Kort, and H. C. van Houwellingen, "A global test for groups fo genes: testing association with a clinical outcome," *Bioinformatics*, vol. 20, no. 1, pp. 93–99, 2004.

[12] J. Tomfohr, J. Lu, and T. B. Kepler, "Pathway level analysis of gene expression using singular value decomposition," *BMC Bioinformatics*, vol. 6, article 225, 2005.

[13] S. W. Kong, W. T. Pu, and P. J. Park, "A multivariate approach for integrating genome-wide expression data and biological knowledge," *Bioinformatics*, vol. 22, no. 19, pp. 2373–2380, 2006.

[14] D. Nam and S. Y. Kim, "Gene-set approach for expression pattern analysis," *Briefings in Bioinformatics*, vol. 9, no. 3, pp. 189–197, 2008.

[15] I. Dinu, J. D. Potter, T. Mueller et al., "Improving gene set analysis of microarray data by SAM-GS," *BMC Bioinformatics*, vol. 8, article 242, 2007.

[16] C. A. Tsai and J. J. Chen, "Multivariate analysis of variance test for gene set analysis," *Bioinformatics*, vol. 25, no. 7, pp. 897–903, 2009.

[17] X. Wang, I. Dinu, W. Liu, and Y. Yasui, "Linear combination test for hierarchical gene set analysis," *Statistical Applications in Genetics and Molecular Biology*, vol. 10, article 13, 2011.

[18] R. Simon, "BRB-Array tools version 4.3 user's manual," Biometrics Research Branch, National Cancer Institute, 2012, http://linus.nci.nih.gov/BRB-ArrayTools.html.

[19] J. Kang, A. D. D'Andrea, and D. Kozono, "A DNA repair pathway-focused score for prediction of outcomes in ovarian cancer treated with platinum-based chemotherapy," *Journal of the National Cancer Institute*, vol. 104, no. 9, pp. 670–681, 2012.

[20] N. Li, H. H. Nguyen, M. Byrom, and A. D. Ellington, "Inhibition of cell proliferation by an anti-EGFR aptamer," *PLoS ONE*, vol. 6, no. 6, Article ID e20299, 2011.

[21] E. Bicaku, Y. Xiong, D. C. Marchion et al., "In vitro analysis of ovarian cancer response to cisplatin, carboplatin, and paclitaxel identifies common pathways that are also associated with overall patient survival," *British Journal of Cancer*, vol. 106, no. 12, pp. 1967–1975, 2012.

[22] H.-M. Hsueh, D.-W. Zhou, and C.-A. Tsai, "Random forests-based differential analysis of gene sets for gene expression data," *Gene*, vol. 518, no. 1, pp. 179–186, 2013.

[23] H. Pang, A. Lin, M. Holford et al., "Pathway analysis using random forests classification and regression," *Bioinformatics*, vol. 22, no. 16, pp. 2028–2036, 2006.

[24] I. Dinu, X. Wang, L. E. Kelemen, S. Vatanpour, and S. Pyne, "Linear combination test for gene set analysis of a continuous phenotype," *BMC Bioinformatics*, vol. 14, no. 1, article 212, 2013.

[25] P. H. Westfall and S. S. Young, *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*, John Wiley & Sons, New York, NY, USA, 1993.

[26] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society B: Methodological*, vol. 57, no. 1, pp. 289–300, 1995.

[27] P. C. O'Brien, "Procedures for comparing samples with multiple endpoints," *Biometrics*, vol. 40, no. 4, pp. 1079–1087, 1984.

[28] J. Schäfer and K. Strimmer, "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics," *Statistical Applications in Genetics and Molecular Biology*, vol. 4, no. 1, pp. 1544–6115, 2005.

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

[30] W. T. Barry, A. B. Nobel, and F. A. Wright, "Significance analysis of functional categories in gene expression studies: a structured permutation approach," *Bioinformatics*, vol. 21, no. 9, pp. 1943–1949, 2005.

[31] L. J. van't Veer, H. Dai, M. J. van de Vijver et al., "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, no. 6871, pp. 530–536, 2002.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 538549, 9 pages
http://dx.doi.org/10.1155/2014/538549

Research Article

Target Capture and Massive Sequencing of Genes Transcribed in *Mytilus galloprovincialis*

Umberto Rosani,[1] Stefania Domeneghetti,[1] Alberto Pallavicini,[2] and Paola Venier[1]

[1] *Department of Biology, University of Padua, Via U. Bassi 58/b, 32121 Padua, Italy*
[2] *Department of Life Sciences, University of Trieste, Via L. Giorgeri, No. 5, 34121 Trieste, Italy*

Correspondence should be addressed to Paola Venier; paola.venier@unipd.it

Received 20 February 2014; Revised 29 May 2014; Accepted 7 June 2014; Published 30 June 2014

Academic Editor: Tzu-Ya Weng

Next generation sequencing (NGS) allows fast and massive production of both genome and transcriptome sequence datasets. As the genome of the Mediterranean mussel *Mytilus galloprovincialis* is not available at present, we have explored the possibility of reducing the whole genome sequencing efforts by using capture probes coupled with PCR amplification and high-throughput 454-sequencing to enrich selected genomic regions. The enrichment of DNA target sequences was validated by real-time PCR, whereas the efficacy of the applied strategy was evaluated by mapping the 454-output reads against reference transcript data already available for *M. galloprovincialis* and by measuring coverage, SNPs, number of *de novo* sequenced introns, and complete gene sequences. Focusing on a target size of nearly 1.5 Mbp, we obtained a target coverage which allowed the identification of more than 250 complete introns, 10,741 SNPs, and also complete gene sequences. This study confirms the transcriptome-based enrichment of gDNA regions as a good strategy to expand knowledge on specific subsets of genes also in nonmodel organisms.

1. Introduction

Genome enrichment methods are efficient ways to reduce sequencing efforts and costs by examining only selected target regions of a given genome [1, 2]. Indeed, the target capture and sequencing approach increases the level of coverage and adds multiplexing options. Various strategies of target enrichment have been employed in many areas of genetic research, from whole exome sequencing [3], sequencing of genes causally linked to diseases [4], and extensive exome resequencing [5] to microbial metagenomics [6] and investigations on ancient DNA [7]. Recently, target enrichment approaches have been applied also to organisms without sequenced genome in order to target the exome for SNP identification or determine gene copy number [8–10]. Before the advent of NGS sequencing, four different methodologies were already available to enrich DNA targets of interest: PCR-based enrichment, microarray- or liquid-based hybridization, restriction enzyme-based enrichment, and physical isolation of mRNA [11]. Among these techniques, in-solution hybrid capture coupled with PCR amplification is one of the most efficient methods for sequencing small- and medium-size targets and it represents a cost-effective procedure in case of low DNA amounts [12, 13].

Mytilus spp. are widespread aquaculture bivalves also used as biosensors for coastal water pollution [14]. *M. galloprovincialis* has a diploid complement of 28 chromosomes and the haploid DNA content is estimated in 1.38–1.88 Gbp [15]. Its genome is not available and gene sequences are still limited. Nevertheless, mussel gene transcript data are increasing in public databases (19,617 nucleotide sequences, 2,292 proteins, and 319 GEO datasets available at NCBI, May 2014) and support the development of gene-centered studies [16–22]. In order to analyze the molecular variability of transcripts coding for antimicrobial mussel peptides (AMPs) in *M. galloprovincialis*, we have previously made available high-throughput amplicon sequence data [23].

This work aims to increase the current knowledge on the *M. galloprovincialis* genome through the analysis of selectively enriched DNA regions and to assess the feasibility of a small target capture approach in a nonsequenced organism. Toward this end, we have designed a high number of probes

on 1,518 mussel transcripts originated by Sanger sequencing [24] and used them to target the related DNA regions by massive 454-sequencing.

2. Materials and Methods

2.1. DNA and Library Preparation. Genomic DNA (gDNA) was extracted from the foot of one adult mussel using a standard phenol/chloroform method [25]. We assessed the DNA quality on 2% agarose gel with SYBR Safe staining (Invitrogen, Carlsbad, Germany) and DNA concentration with a NanoDrop spectrophotometer (NanoDrop Technologies, Wilmington, USA). The sampled animal was confirmed to be *M. galloprovincialis* by Sanger sequencing of a variable interspecific region of the gene MSLAP [26].

To prepare a single-stranded DNA library, the purified gDNA was fragmented by nebulization and two adapters for GS-FLX sequencing (A 5′-CCATCTCATCCCTGCGTGTCTCCGAC-3′ and B 5′-CCTATCCCCTGTGTGCCTTGGCAGTC-3′) were ligated following the manufacturer's protocol (Roche Life Sciences). Library amount and size were assessed with the NanoDrop 3300 fluorimeter (NanoDrop Technologies) and Bioanalyzer DNA7500 lab-chip (Agilent Technologies, Santa Clara, USA), respectively. The DNA library was then PCR-amplified by using sequencing adapters as forward and reverse primers in order to obtain the quantity necessary for the enrichment procedure.

2.2. Probe Design and DNA Enrichment and Sequencing. The *M. galloprovincialis* EST contigs previously catalogued in *Mytibase* were ordered by length. Those longer than 750 bp were selected, so as to design oligonucleotide probes and target the related genomic regions. Shorter sequences representing mussel AMPs were also considered in the probe design, for a total of 1,518 selected contigs covering altogether a genomic target region of 1.35 Mb. RNA probes (120-mers) were designed on the contig sequences, tiled every 60 bp, in order to obtain a final 2× coverage. Biotinylated probes were synthesized by Mycroarray.com (Ann Arbor, USA) with a technology which allows the conversion of a DNA oligonucleotide library into biotinylated RNA baits by *in vitro* transcription. The RNA probe size was accurately checked on RNA6000 lab-chip (Agilent Technologies).

We started the capture-enrichment step from 500 ng of gDNA following manufacturer's instructions [27]. The number of cycles in the postcapture PCR amplification step was set at 15 using a Herculase High Fidelity Taq (Invitrogen) in a Mastercycler Gradient Thermal Cycler (Eppendorf, Hamburg, Germany) as follows: 95°C for 1′, 15 cycles at 95°C for 30″, 60°C for 30″, 68°C for 1′, and a final extension step at 68°C for 5′. After purification, the PCR product was subjected to emulsion PCR and 454-sequencing according to standard protocols. Two independent half runs of a PicoTitre plate were performed (BMR Genomics, Padua, Italy).

2.3. Enrichment Controls. Four transcripts (3 on-target sequences: MGC05878, MGC00300, and MGC04518; 1 off-target sequence) were evaluated before and after enrichment of the genomic library by RT-PCR analysis. Raw and enriched samples were diluted 10, 50, 100, and 500 times and used as template for RT-PCR, each one with 3 replicates per dilution. PCR was performed in a CFX96 real-time PCR detection system (Bio-Rad) using iQ SYBR Green Supermix (Bio-Rad) with the following thermal profile: 5′ at 95°C, 40 cycles of amplification of 30″ at 95°C, 20″ at 56°C, and 30″ at 72°C. Enrichment ratios were computed with the Ct values resulting from the raw DNA library and the enriched one.

2.4. Sequencing Data Analysis. Raw sequencing data are deposited at the NCBI SRA archive under accession number PRJNA246144. The output reads having a minimal Phred quality score (>Q20) [28] were retained for data analysis. Such reads were trimmed for length (>50 nt), quality (<2 Ns), and adaptor sequences. Moreover, duplicated reads and reads mapping on publicly available *Mytilus* microsatellite sequences were excluded. In order to estimate the target coverage, we mapped the trimmed reads on the initially selected EST contigs with the CLC Genomic Workbench, version 5.1 (CLC Bio, Katrinebjerg, Denmark).

All high quality reads, as well as the subsets of on-target and off-target reads, were separately subjected to *de novo* assembling with standard parameters to improve the data analysis (mismatch, insertion, and deletion costs were set at 2\3\3, resp.; length fraction and similarity were set at 0.5\0.8). The contigs resulting from the whole assembly were subjected to SNP discovery and intron-exon analysis. SNPs were prudently considered genuine when coverage was at least 10× and the sequence variation was present at least in 35% of the locally aligned reads. Putative gene structures inferred from the *de novo* contigs were validated using a *M. galloprovincialis* Illumina RNA-seq dataset (SRA ID: PRJNA88481). Briefly, raw Illumina reads were quality trimmed and good quality reads were back mapped on the previously obtained genomic contigs. Evaluation of coverage depth and percentage of covered hits was performed on contigs composed by introns, exons, and the two of them. The latter were used to identify the completely sequenced introns.

2.5. Sequence Validation. Six contigs resulting from the genomic assembly (3 intronic sequences, 2 mytilin C regions, and 1 MSLAP gene) were validated by Sanger sequencing. After primer design, PCR was carried out in 50 μL final volume using a Mastercycler Gradient Thermal Cycler set as follows: 95°C for 1′, 35 cycles at 95°C for 30″, 55°C for 20″, 68°C for 1′, and a final extension step at 68°C for 5′. Purified PCR products (PureLink PCR purification kit, Invitrogen) were checked by electrophoresis in 2% agarose gel with SYBR Safe staining (Invitrogen) and sequenced in forward and reverse direction (BMR Genomics, Padua, Italy). Primer sequences and other details are available in Supplementary Materials (SM1) (see SM1 in the Supplementary Material available online at http://dx.doi.org/10.1155/2014/538549).

3. Results

Using a liquid capture strategy and designing RNA capture probes on the *Mytibase* EST collection [24], we have selected

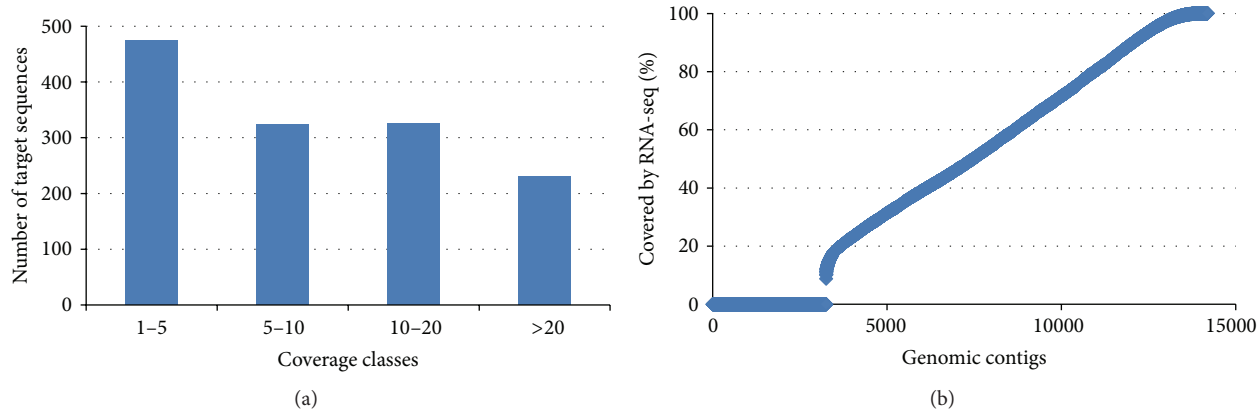


Figure 1: (a) Target coverage distribution. Number of targets per coverage class. (b) Length percentage of the genomic contigs covered by RNA-seq reads.

and sequenced a small genome portion of the marine bivalve *M. galloprovincialis*. The target enrichment performed with the *Mybaits* in-solution hybrid capture system (MYcroarray.com) spanned over 1.35 M bases of sequences expressed in the Mediterranean mussel (1,518 contigs targeted by 12,039 RNA probes of 120-mers). Theoretically, our target regions represent 0.1% of the mussel genome, currently estimated at 1.38–1.88 Gbp [15]. For the purpose of comparison, Table 1 reports that the estimated genome size of mollusk species is already sequenced.

We used a medium-size genomic library (average dimension of 680 bp) suitable to the sequencing capacity of the 454 Titanium platform to produce two 454 datasets in parallel (RUN_1 and RUN_2; details are in Table 2 and SM2). Reads of low quality, duplicates, and reads mapping on microsatellite repeats were excluded from the analysis. Clonality of reads, evaluated by counting the identical reads, was around 36%. Therefore, we obtained a total of 626,769 high quality reads (average length: 235 bp) and 354,633 of them could be mapped on the reference set of *Mytibase* transcript sequences, with 1,355 out of 1,518 contigs having at least 1× coverage. Accordingly, the capture efficiency in RUN_1 and RUN_2 was 62% and 52%, respectively.

Most of the 1,355 covered targets had a sequence coverage ranging from 1× to 10× (median value 8×), with about 600 targets displaying more than 5× coverage (Figure 1(a)). We did not observe any positive correlation between target length—directly related to the number of designed probes—and target coverage (SM3).

The sequence coverage observed in the selected cases (3 on-target and 1 off-target sequences) was consistent with the enrichment level measured before sequencing by RT-PCR analysis (Table 3).

De novo assembling of the whole sequence read dataset yielded 14,339 contigs, with average length of 511 bp (200–4,217 bp), as summarized in Table 2 and SM4. A first-hit BLAST annotation is reported for the contigs covered with more than 100 reads (SM5).

The reliability of the probe enrichment approach is evident when taking into account the C1q gene family: a total of 1,753 reads were mapped on the 99 gene sequences of the C1q multigene family [34]. In this study, the majority of the mapped reads (96%) matched exclusively with the 28 C1q genes targeted by the designed RNA probes.

Table 1: Genome size of sequenced mollusk species.

Species name	*C*-value (pg)	Length (Gbp)
Crassostrea gigas (see Zhang et al. [29])	0.91	1.00
Lottia gigantean (see Simakov et al. [30])	0.43	0.35
Pinctada fucata (see Takeuchi et al. [31])	/	1.20
Aplysia californica (see Broad Institute (US) [32])	1.8/2.0	0.74
Pecten maximus (see Biscotti et al. [33])	1.42	1.40

C-values (pg) are summarized according to http://www.genomesize.com.

In order to identify gene structures, we mapped our *M. galloprovincialis* read collection (SRA ID: PRJNA88481) on the genomic contigs resulting from this study. More than 29 M reads gave a positive match (Figure 1(b)) and allowed us to predict 4,587 gene regions and identify introns and exons. About 20% of the genomic contigs did not match any RNA-seq read and, either complete or partial, such lack of sequence coverage suggests the presence of introns. Mytilin C (Figure 2) and GADD45 (SM6) exemplify gene structures identified in the assembled genomic contigs. In order to confirm some of the novel introns, we designed primers flanking the intron region and, after amplification, we subjected the PCR products to Sanger sequencing. The resulted sequences, amplified from a different individual mussel, were consistent with the genomic contigs previously assembled (SM1).

With a view to recognizing and counting the fully sequenced introns, *de novo* contigs showing at least 250 bp divergence between the covered length and the total length were blasted against the initial *Mytibase* targets (allowing only one high-scoring segment pair). As a result of retrieving the genomic contigs with more than one hit, we were able to

Table 2: Sequencing output data and summary of *de novo* assembling results.

Sequencing output	RUN_1		RUN_2		RUN_1 + 2	
Total reads	472,122		473,409		**945,531**	
Total high quality reads	287,362		339,407		**626,769**	
Average length (bp)	380		114		**235**	
On-target reads (number and %)	179,201	62%	175,432	52%	**354,633**	**57%**
Covered targets (number and %)	1,262	83%	1,032	68%	**1,355**	**89%**
De novo assembly	On-target reads		Off-target reads		All dataset	
Total contigs	5,547		12,423		**14,339**	
Total assembled reads (number and %)	279,922	79%	347,439	45%	**444,145**	**71%**
Average contig length (bp)	490		476		**511**	
N50 (bp)	557		523		**552**	
N75 (bp)	388		402		**405**	
Longest contig (bp)	2,234		3,538		**4,217**	
Contigs with blast annotation	28%		21%		**44%**	

Total raw and HQ reads, average length (bp), number of mapped reads (on-target), and covered contigs are reported for the subsets (RUN_1, RUN_2) and total sequenced data (RUN_1 + 2).
De novo assembly of the reads on-target and off-target and of the whole dataset. Number of resulting contigs and related reads, contig length, quality parameters, longest contig, and percentage of annotated contigs are reported.

Table 3: Enrichment fold and coverage of selected transcripts.

ID	Status	Enrichment fold (RT_PCR)	Sequenced reads (NGS)
MGC04518	On-target	32	22
MGC00300	On-target	60	45
MCG05878	On-target	2	7
Target 4	Off-target	−867	/

Enrichment real-time analysis was performed on 3 targets and on 1 not selected transcript (target 4).
Enrichment fold was measured in qRT-PCR by comparing the DNA library before and after enrichment and, subsequently, by reporting the number of reads that mapped uniquely on the targets.

identify 263 fully sequenced introns on a total of 474 contigs (Table 4).

Among other findings, several genomic contigs related to the NF-kB pathway transcripts were identified. For instance, we found the complete gene sequence of MgIRAK-4 composed by one 1,774 bp long exon (GenBank ID KC994891, 533aa; contigs 1408 and 2351) and one MgIkB-a (KF015301, 392aa; contig 5756). Both of these genes include only one exon, a gene structure conserved also in the *C. gigas*, *P. fucata,* and *L. gigantea* genomes. We found the complete MyD88 sequence spread for 2.4 kbp over several genomic contigs and composed by three exons, with the typical DEATH domain completely localized in the first exon and the TIR domain along the remaining ones. A similar gene organization is present in the *C. gigas* (CGI_1002602 and CGI_10013672) and *L. gigantea* (sca_120023) genomes.

With regard to the lipopolysaccharide-induced TNF factor-like sequences (LITAF), seven different transcripts are present in *Mytilus* spp. (GenBank IDs: KF051277 and KF110675-82); this indicates the presence of a multigene LITAF family like in the oyster genome. In this study, the genomic contigs matching the LITAF 4, 5, and 6 transcripts show the same gene organization at the 3′ region. In detail, a common intron/exon junction (Figure 3), with the last 38 highly conserved amino acids entirely localized in one exon,

Table 4: Fully sequenced introns.

Total contigs with introns	204
Total introns	263
Total intron length (bp)	110,643
Average intron length (bp)	434
Maximum intron length (bp)	1,008
Minimum intron length (bp)	100

suggests evolutionary diversification by alternative splicing coupled with gene duplication.

The *Mytibase* transcript sequences for the mussel AMPs were included in the list of targets to be enriched, even if their length was less than 750 bp (see M&M). This can explain the low coverage. Regarding mytilin B, the complete 3,125 bp long gene sequence was already available (NCBI ID: AF177540) [16]. Although the presence of at least 3 mytilin transcripts (B, C, and D) has been reported, their gene structures are not yet disclosed. Manual extraction and assembling of the 374 genomic reads aligned on the mytilin targets allowed us to partially reconstruct the gene structure. Figure 2(b) exemplifies the case of mytilin C.

We used both the genomic and transcriptomic reads of *M. galloprovincialis* to evaluate the nucleotide variability per

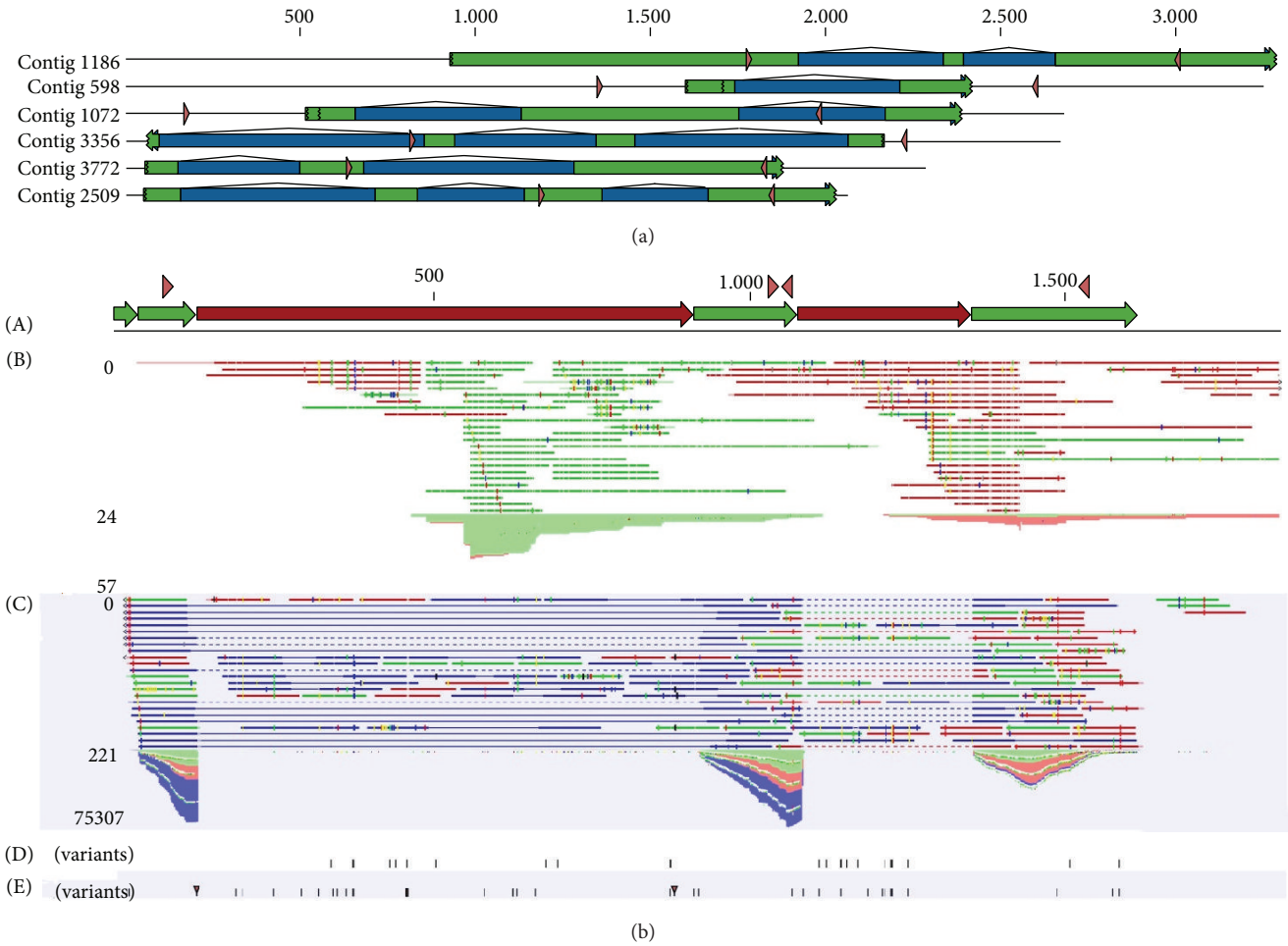


Figure 2: (a) Genomic contigs resulting from *de novo* assembling and evaluated for confirmatory Sanger sequencing. Exon and intron regions are depicted in green and blue, respectively. Pink arrows represent the primer positions. (b) Predicted mytilin C gene: (A) gene structure with exons (green), introns (red), and primer positions (pink); (B) mapping of genomic reads with related coverage graph; (C) mapping of transcriptomic reads with related coverage graph; (D) variant positions detected on genomic sequences; (E) variant positions detected on transcriptomic sequences.

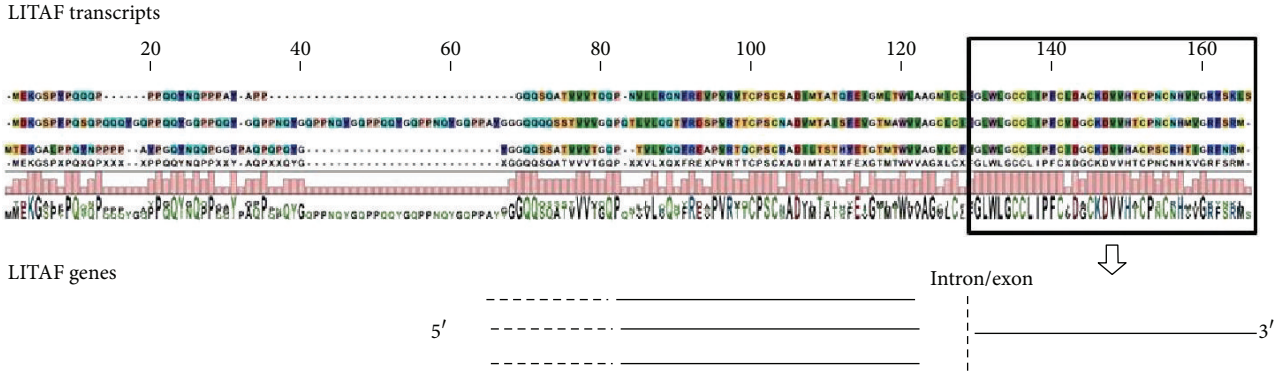


Figure 3: Partial LITAF gene structure. Alignment of the three targeted LITAF transcripts, translated into amino acids. Conservation graph shows the high conservation of the 3′ region, located entirely in one exon. The 5′ region displays a lower conservation level.

Table 5: SNP identification in genomic and transcriptomic data of *M. galloprovincialis*.

	Total SNPs	Total contigs with SNPs	SNP frequency (%)	SNPs in exons (%)
Genome	10,741	2,326	0.71	0.58
Transcriptome	13,821	2,057	0.96	0.87
Common	1,135	447	0.31	/

Table 6: Overview on *Mytilus* AMPs data.

AMP name	ID NCBI	Reads	Sequence length (NCBI) (bp)	Sequence extension (bp)	SNPs (genomic)	SNPs (transcriptomic)	Common SNPs
Mytilin B	AF177540	271	3,125	0	17	9	3
Mytilin C	/	130	/	1,834	7	23	4
Mytilin D	EU810204	53	/	1,165	8	5	3
Myticin A	/	95	/	1,650	87	54	17
Myticin B	EU088427	72	2,775	0	20	30	14
Myticin C	EU927419	163	1,409	466	61	78	26

Selection of mussel AMPs listed by name, NCBI ID (if present), number of aligned reads, length of public available sequences (bp), sequence elongation (bp), and number of genomic, transcriptomic, and common SNCs.

contig. Regarding the genomic reads, the SNP count highlighted 10,741 variable positions in 2,327 out of 14,339 contigs, with an average variation frequency of 0.71% (SNP/nt) and 60% of the SNPs localized on mRNA regions. A higher variation rate was evident by mapping the transcriptomic reads on the same genomic contigs (Table 5). Fewer SNPs are common in the genome and transcriptome datasets: 1,135 SNPs appear in 447 contigs, indicative of 551 transitions and 485 transversions (Figure 4). Among the most variable contigs, only 30% of them display at least a partial functional annotation, suggesting that most of the observed variability is located in intronic sequences.

Despite the low coverage obtained for the AMP-related targets, we could extend the available genomic sequences for some of them (Table 6). As expected, the AMP-related genomic sequences displayed high molecular variability, further confirming the different level of variability previously reported by Mytilin and Myticin gene families [23]. In detail, the intronic nucleotide variants reported by Pallavicini et al. [35] are all confirmed in a single mussel, laying the foundations for investigating the presence of genes coding for nontranslated RNAs or other possible causes of intron conservation in the antimicrobial precursor sequences.

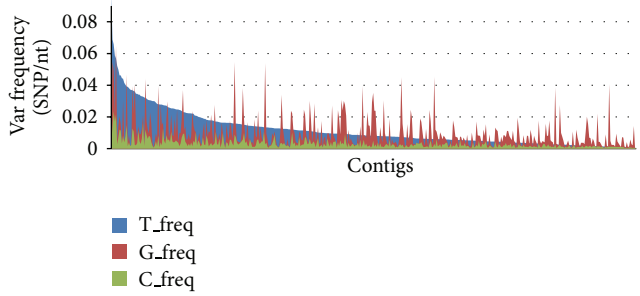


Figure 4: SNP frequency in mussel exons. Predicted frequency of SNPs in exons using transcriptomic data (T, in blue), genomic data (G, in red), and the common SNP dataset (C, in green).

4. Discussion

Sequence enrichment strategies take advantage of the NGS high sequencing power and, at the same time, reduce the analytical complexity by focusing on limited portions of a given genome. Nowadays, sequence enrichment protocols can be applied to an increased number of organisms since many genomes of interest have been completely sequenced. Although the genome of *M. galloprovincialis* is not yet sequenced and gene-specific data are scarce, we designed 12 k RNA probes on the Mytibase transcript collection to capture a small portion of the mussel genome by PCR amplification and subsequent 454-sequencing. The *M. galloprovincialis* genome size is remarkable (estimated length: 1.38–1.88 Gbp) and not easily affordable with a random-shotgun approach, due to the number of sequencing runs necessary to obtain an adequate genome coverage and the redundancy expected from repetitive sequences. Focusing on genome regions related to the mussel exome, we defined a small target of study (theoretically, about 0.1% of the nuclear DNA) which could be regarded as preliminary to full genome analysis.

The design of 120-mer probes tiled every 60 nucleotides aimed to ensure, as much as possible, the linear coverage of all targets as reported by Tewhey et al. [12]. Long probes were also chosen to assure a high specificity of the capture reaction, though the presence of introns was expected to interrupt the probe-target matching. Nevertheless, the analysis carried out on the C1q multigene family returned an encouraging result, since only the genes targeted by RNA probes showed a positive read coverage.

We based our analysis on two highly comparable read datasets (RUN1 and RUN2 differ only in their read length) and found a positive correlation between read length and positive matching on target (Table 2, SM2). Overall, 57% of sequenced reads were positively assigned and 89% of

targets were covered, consistently with other target capture experiments [3]. On the whole, we observed an average target enrichment of 50×, even though with uneven coverage distribution among targets (SM3). Such uneven distribution could have been influenced by the gene copy number and, more probably, by target redundancy, which might have increased due to the presence of unknown ESTs. It bears remarking that the testing of the four enrichment controls indicated the effectiveness of the enrichment strategy, with comparable results between library enrichment and sequencing coverage.

Data analysis performed on the on-target reads highlighted some procedural constraints, mainly due to the essential differences between the sequenced reads (genomic) and probes designed on transcripts. Moreover, the absence of a mussel genome scaffold makes it difficult to discriminate between completely off-target reads and reads located at 5′ or 3′ of the targets. The *de novo* assembly performed on the whole read dataset was the most effective way to overcome these constraints. Such assembly produced more than 14 k contigs with an average length of 511 bp (range 200–4, 217 bp). In addition, the contig annotation mainly showed sequence similarities to oyster genes, with the percentage of unknown contigs similar to that of other nonsequenced organisms [36, 37].

We estimated the number of intron contigs by counting the contigs without RNA-seq coverage (2.5 k in total, Figure 1(b)). Since 51% of the total contig length was covered by the RNA-seq reads, the remaining 49% (about 3.5 Mbp) may refer to introns or gene transcripts not represented in our RNA-seq data. The presence of other nuclear RNA types has not been analyzed in our work, due to the lack of a complete transcriptome dataset for *M. galloprovincialis*. We then evaluated the number of completely sequenced introns by blasting the genomic contigs against the selected *Mytibase* targets (Table 4). Complete introns were only 3% of the total intron length (i.e., the majority of these introns have still to be completed). Due to the initial choice of a medium-size DNA library which could be completely covered by our 454-sequencing effort, in this study we were not able to fully sequence large introns or to recover regulatory gene elements.

With reference to the resulting data, we found contigs identifying IRAK, IkB, and MyD88 genes, that is, elements belonging to the NFkB pathway recently described in mussels [22, 38] and other important genes (e.g., GADD45, mytilin C, and LITAF). We used these contigs to reconstruct, as much as possible, the gene sequences and compare the gene structure among related species. In the cases of IRAK, GADD45, mytilin C, and IkB, we reconstructed the whole gene. Some of these genes are composed by a single exon, whereas other genes (MyD88, GADD45, and mytilin C) include many exons. In the case of LITAF, we could recover only a part of the gene, and the intercomparison of the several LITAF transcripts in mussels suggests mechanisms for the evolutionary diversification of this multigene family.

We used both the genomic and transcriptomic mussel datasets for SNC analysis, also ranking the contigs on the basis of their sequence diversity. In order to minimize the number of false positives, we applied stringent parameters for SNP calling and removed the reads with identical mapping location. As a matter of fact, variability could be introduced by 454-sequencing [39] and PCR amplification; in particular, errors produced in the early PCR cycles could then be spread in multiple reads with a low possibility to distinguish them from real SNCs. About 60% of the observed variability was located in exome regions and 20% of the SNCs were confirmed by RNA-seq data (Table 5). Furthermore, we analyzed in greater detail the SNPs found in myticin and mytilin genes, known as antimicrobial mussel peptides. Our genomic data support the different sequence diversity previously observed at transcript level [23] and confirm the myticin C as one of the most variable antimicrobials of *M. galloprovincialis,* also at gene level. Moreover, only about 33% of the SNPs present in the analyzed AMP transcripts were detected in the related genomic contigs, probably because the latter refer to the analysis of one individual mussel whereas the RNA-seq reads were produced from pooled mussel samples. Therefore, both transcriptional and posttranscriptional mechanisms are expected to play a role in the production of the observed transcript diversity of mussel AMPs [40, 41].

5. Conclusions

This work has exploited the feasibility of a genome-targeted sequencing based only on transcriptomic data of *M. galloprovincialis*. Relying on a continuous and redundant probe design on the expressed mussel sequences, the selected strategy allowed us to improve knowledge on the targeted gene regions, representing a first overview of the genome of the Mediterranean mussel. This work could be further developed by implementing the RNA probe design, library dimension, and total target size. At the same time, whole genome sequencing of *Mytilus spp.* is still underway.

Abbreviations

AMP: Antimicrobial peptide
bp: Base pair
EST: Expressed sequence tag
gDNA: Genomic DNA
NGS: Next generation sequencing
nt: Nucleotide
PCR: Polymerase chain reaction
RT-PCR: Real-time PCR
SNP: Single nucleotide polymorphism.

Conflict of Interests

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

Acknowledgments

The work has been developed within the Cross-border Cooperation Operational Programme Italia-Slovenija 2007–2013 (Innovaqua) and PRAT 2012 CPDA128951 (University of Padua). The authors are grateful to Sara Boscariol who kindly did the final linguistic revision.

References

[1] D. Summerer, "Enabling technologies of genomic-scale sequence enrichment for targeted high-throughput sequencing," *Genomics*, vol. 94, no. 6, pp. 363–368, 2009.

[2] M. Harakalova, M. Mokry, B. Hrdlickova et al., "Multiplexed array-based and in-solution genomic enrichment for flexible and cost-effective targeted next-generation sequencing," *Nature Protocols*, vol. 6, no. 12, pp. 1870–1886, 2011.

[3] S. B. Ng, K. J. Buckingham, C. Lee et al., "Exome sequencing identifies the cause of a mendelian disorder," *Nature Genetics*, vol. 42, no. 1, pp. 30–35, 2010.

[4] A. Hoischen, C. Gilissen, P. Arts et al., "Massively parallel sequencing of ataxia genes after array-based enrichment," *Human Mutation*, vol. 31, no. 4, pp. 492–499, 2010.

[5] Y. Li, N. Vinckenbosch, G. Tian et al., "Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants," *Nature Genetics*, vol. 42, no. 11, pp. 969–972, 2010.

[6] J. Denonfoux, N. Parisot, E. Dugat-Bony et al., "Gene capture coupled to high-throughput sequencing as a strategy for targeted metagenome exploration," *DNA Research*, vol. 20, no. 2, pp. 185–196, 2013.

[7] M. C. Ávila-Arcos, E. Cappellini, J. A. Romero-Navarro et al., "Application and comparison of large-scale solution-based DNA capture-enrichment methods on ancient DNA," *Scientific Reports*, vol. 1, 2011.

[8] S. J. Jakhesara, V. B. Ahir, K. B. Padiya, P. G. Koringa, D. N. Rank, and C. G. Joshi, "Tissue-specific temporal exome capture revealed muscle-specific genes and SNPs in Indian buffalo (Bubalus bubalis)," *Genomics, Proteomics and Bioinformatics*, vol. 10, no. 2, pp. 107–113, 2012.

[9] J. Li, R. Lupat, K. C. Amarasinghe et al., "CONTRA: copy number analysis for targeted resequencing," *Bioinformatics*, vol. 28, no. 10, pp. 1307–1313, 2012.

[10] L. Zhou and J. A. Holliday, "Targeted enrichment of the black cottonwood (*Populus trichocarpa*) gene space using sequence capture," *BMC Genomics*, vol. 13, no. 1, article 703, 2012.

[11] R. Cronn, B. J. Knaus, A. Liston et al., "Targeted enrichment strategies for next-generation plant biology," *The American Journal of Botany*, vol. 99, no. 2, pp. 291–311, 2012.

[12] R. Tewhey, M. Nakano, X. Wang et al., "Enrichment of sequencing targets from the human genome by solution hybridization," *Genome Biology*, vol. 10, no. 10, article R116, 2009.

[13] F. Mertes, A. ElSharawy, S. Sauer et al., "Targeted enrichment of genomic DNA regions for next-generation sequencing," *Briefings in Functional Genomics*, vol. 10, no. 6, pp. 374–386, 2011.

[14] E. D. Goldberg and K. K. Bertine, "Beyond the mussel watch—new directions for monitoring marine pollution," *Science of the Total Environment*, vol. 247, no. 2-3, pp. 165–174, 2000.

[15] T. R. Gregory, J. A. Nicol, H. Tamm et al., "Eukaryotic genome size databases," *Nucleic Acids Research*, vol. 35, no. 1, pp. D332–D338, 2007.

[16] G. Mitta, F. Hubert, E. A. Dyrynda, P. Boudry, and P. Roch, "Mytilin B and MGD2, two antimicrobial peptides of marine mussels: gene structure and expression analysis," *Developmental and Comparative Immunology*, vol. 24, no. 4, pp. 381–393, 2000.

[17] C. N. Pantzartzi, A. Kourtidis, E. Drosopoulou, M. Yiangou, and Z. G. Scouras, "Isolation and characterization of two cytoplasmic hsp90s from *Mytilus* galloprovincialis (Mollusca: Bivalvia) that contain a complex promoter with a p53 binding site," *Gene*, vol. 431, no. 1-2, pp. 47–54, 2009.

[18] M. Vera, P. Martínez, L. Poisa-Beiro, A. Figueras, and B. Novoa, "Genomic organization, molecular diversification, and evolution of antimicrobial peptide myticin-C genes in the mussel (*Mytilus galloprovincialis*)," *PLoS ONE*, vol. 6, no. 8, Article ID e24041, 2011.

[19] A. Romero, S. Dios, L. Poisa-Beiro et al., "Individual sequence variability and functional activities of fibrinogen-related proteins (FREPs) in the Mediterranean mussel (*Mytilus galloprovincialis*) suggest ancient and complex immune recognition models in invertebrates," *Developmental and Comparative Immunology*, vol. 35, no. 3, pp. 334–344, 2011.

[20] S. Aceto, G. Formisano, F. Carella, G. de Vico, and L. Gaudio, "The metallothionein genes of *Mytilus galloprovincialis*: genomic organization, tissue expression and evolution," *Marine Genomics*, vol. 4, no. 1, pp. 61–68, 2011.

[21] M. Sonthi, F. Cantet, M. Toubiana et al., "Gene expression specificity of the mussel antifungal mytimycin (MytM)," *Fish and Shellfish Immunology*, vol. 32, no. 1, pp. 45–50, 2012.

[22] M. Toubiana, M. Gerdol, U. Rosani, A. Pallavicini, P. Venier, and P. Roch, "Toll-like receptors and MyD88 adaptors in *Mytilus*: complete cds and gene expression levels," *Developmental and Comparative Immunology*, vol. 40, no. 2, pp. 158–166, 2013.

[23] U. Rosani, L. Varotto, A. Rossi et al., "Massively parallel amplicon sequencing reveals isotype-specific variability of antimicrobial peptide transcripts in *Mytilus galloprovincialis*," *PLoS ONE*, vol. 6, no. 11, Article ID e26680, 2011.

[24] P. Venier, C. de Pittà, F. Bernante et al., "MytiBase: a knowledgebase of mussel (*M. galloprovincialis*) transcribed sequences," *BMC Genomics*, vol. 10, article 72, 2009.

[25] S. W. M. John, G. Weitzner, R. Rozen, and C. R. Scriver, "A rapid procedure for extracting genomic DNA from leukocytes," *Nucleic Acids Research*, vol. 19, no. 2, p. 408, 1991.

[26] K. Inoue, J. H. Waite, M. Matsuoka, S. Odo, and S. Harayama, "Interspecific variations in adhesive protein sequences of *Mytilus edulis, M. galloprovincialis*, and *M. trossulus*," *The Biological Bulletin*, vol. 189, no. 3, pp. 370–375, 1995.

[27] A. Gnirke, A. Melnikov, J. Maguire et al., "Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing," *Nature Biotechnology*, vol. 27, no. 2, pp. 182–189, 2009.

[28] B. Ewing and P. Green, "Base-calling of automated sequencer traces using phred. II. Error probabilities," *Genome Research*, vol. 8, no. 3, pp. 186–194, 1998.

[29] G. Zhang, X. Fang, X. Guo et al., "The oyster genome reveals stress adaptation and complexity of shell formation," *Nature*, vol. 490, no. 7418, pp. 49–54, 2012.

[30] O. Simakov, F. Marletaz, S. Cho et al., "Insights into bilaterian evolution from three spiralian genomes," *Nature*, vol. 493, no. 7433, pp. 526–531, 2013.

[31] T. Takeuchi, T. Kawashima, R. Koyanagi et al., "Draft genome of the pearl oyster *Pinctada fucata*: a platform for understanding bivalve biology," *DNA Research*, vol. 19, no. 2, pp. 117–130, 2012.

[32] Broad Insitute (USA).

[33] M. A. Biscotti, A. Canapa, E. Olmo et al., "Repetitive DNA, molecular cytogenetics and genome organization in the King scallop (*Pecten maximus*)," *Gene*, vol. 406, no. 1-2, pp. 91–98, 2007.

[34] M. Gerdol, C. Manfrin, G. de Moro et al., "The C1q domain containing proteins of the Mediterranean mussel *Mytilus galloprovincialis*: a widespread and diverse family of immune-related

molecules," *Developmental and Comparative Immunology*, vol. 35, no. 6, pp. 635–643, 2011.

[35] A. Pallavicini, M. del Mar Costa, C. Gestal et al., "High sequence variability of myticin transcripts in hemocytes of immune-stimulated mussels suggests ancient host-pathogen interactions," *Developmental and Comparative Immunology*, vol. 32, no. 3, pp. 213–226, 2008.

[36] M. Milan, A. Coppe, R. Reinhardt et al., "Transcriptome sequencing and microarray development for the Manila clam, *Ruditapes philippinarum*: genomic tools for environmental monitoring," *BMC Genomics*, vol. 12, article 234, 2011.

[37] E. E. R. Philipp, L. Kraemer, F. Melzner et al., "Massively parallel RNA sequencing identifies a complex immune gene repertoire in the lophotrochozoan *Mytilus edulis*," *PLoS ONE*, vol. 7, no. 3, Article ID e33091, 2012.

[38] M. Toubiana, U. Rosani, S. Giambelluca et al., "Toll signal transduction pathway in bivalves: complete cds of intermediate elements and related gene transcription levels in hemocytes of immune stimulated *Mytilus galloprovincialis*," *Developmental & Comparative Immunology*, vol. 45, no. 2, pp. 300–312, 2014.

[39] A. Gilles, E. Meglécz, N. Pech, S. Ferreira, T. Malausa, and J.-F. Martin, "Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing," *BMC Genomics*, vol. 12, article 245, 2011.

[40] W. M. Gommans, S. P. Mullen, and S. Maas, "RNA editing: a driving force for adaptive evolution?" *BioEssays*, vol. 31, no. 10, pp. 1137–1145, 2009.

[41] V. Pelechano, W. Wei, and L. M. Steinmetz, "Extensive transcriptional heterogeneity revealed by isoform profiling," *Nature*, vol. 497, no. 7447, pp. 127–131, 2013.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 675158, 9 pages
http://dx.doi.org/10.1155/2014/675158

Research Article

Defining Loci in Restriction-Based Reduced Representation Genomic Data from Nonmodel Species: Sources of Bias and Diagnostics for Optimal Clustering

Daniel C. Ilut,[1] Marie L. Nydam,[2] and Matthew P. Hare[3]

[1] *Department of Plant Breeding and Genetics, Cornell University, Ithaca, NY 14850, USA*
[2] *Division of Science and Mathematics, Centre College, Danville, KY 40422, USA*
[3] *Department of Natural Resources, Cornell University, Ithaca, NY 14850, USA*

Correspondence should be addressed to Daniel C. Ilut; dci1@cornell.edu

Received 28 February 2014; Revised 2 June 2014; Accepted 4 June 2014; Published 25 June 2014

Academic Editor: Tzu-Ya Weng

Next generation sequencing holds great promise for applications of phylogeography, landscape genetics, and population genomics in wild populations of nonmodel species, but the robustness of inferences hinges on careful experimental design and effective bioinformatic removal of predictable artifacts. Addressing this issue, we use published genomes from a tunicate, stickleback, and soybean to illustrate the potential for bioinformatic artifacts and introduce a protocol to minimize two sources of error expected from similarity-based de-novo clustering of stacked reads: the splitting of alleles into different clusters, which creates false homozygosity, and the grouping of paralogs into the same cluster, which creates false heterozygosity. We present an empirical application focused on *Ciona savignyi*, a tunicate with very high SNP heterozygosity (~0.05), because high diversity challenges the computational efficiency of most existing nonmodel pipelines while also potentially exacerbating paralog artifacts. The simulated and empirical data illustrate the advantages of using higher sequence difference clustering thresholds than is typical and demonstrate the utility of our protocol for efficiently identifying an optimum threshold from data without prior knowledge of heterozygosity. The empirical *Ciona savignyi* data also highlight null alleles as a potentially large source of false homozygosity in restriction-based reduced representation genomic data.

1. Introduction

As population genomic applications of next generation sequencing continue to expand beyond model organisms to nonmodel species and wild populations, reduced representation libraries (RRL) have been used to subsample genomes in a repeatable manner. Here, "nonmodel" refers to species with few genomic resources and in particular the absence of a reference genome. Early efforts, especially within plants, focused on RRL methods that could avoid highly repetitive genomic fractions. Two generalizable approaches to RRL construction have been Cot analysis to isolate the low-copy fraction of a genome [1] and screening of restriction enzymes for those that yield the desired fragment size range, in some cases taking advantage of enzyme methyl sensitivity to target certain genomic regions [2, 3]. Compared with Cot analysis, restriction digestion is a rapid and easy method of generating an RRL whether size selection is gel-based (e.g., RAD-tags, [4]) or a consequence of PCR and sequencing biases (genotyping-by-sequencing or GBS [5]). When a related genome sequence is available, in silico digests can help determine the optimum restriction enzyme with respect to fragment size distribution and avoid repetitive elements [6]. The simplicity of this restriction-based approach has made it the method of choice for genomic sampling apart from transcriptomes and hybridization-based sequence capture [7], but with the expectation that some portion of the data will be from multicopy DNA sequences and will possibly bias estimates of genetic diversity.

With nonmodel taxa, analysis of reduced representational data must contend with a potentially large fraction of repetitive DNA without the quality control benefits of mapping

to a reference genome. Also, the clustering or binning of homologous sequences into allelic loci without the benefit of a reference genome has the potential to create multiple artifacts with different effects on inferred heterozygosity. When the goal is to simply find a set of single nucleotide polymorphisms to analyze spatially among population samples, stringent quality filtering can yield valuable data at the cost of enormous data loss. In contrast, when the goal is to identify individual candidate loci using genome scans or association mapping, or to make inferences requiring unbiased estimates of heterozygosity, results will be sensitive to biases and artifacts introduced by library construction, quality filtering, or allelic clustering methods. These latter applications are where the full and exciting potential of next generation sequence data hold the greatest promise to transform the questions that can be empirically addressed in natural populations, by making it possible to find and analyze both neutral and potentially nonneutral genomic variation. Therefore, our goal was to investigate the potential for large biases from repetitive DNA and copy number variation and explore novel de novo clustering methods that can minimize heterozygosity biases.

In what follows, we focus on reduced representation methods of genomic sampling such as RADseq and GBS. These methods focus sequencing effort on loci (sometimes referred to as "tags") that are no larger than the sequence read length from high-throughput platforms (currently 100–150 bp read lengths). Usually in these protocols there is no "assembly" of partially overlapping sequences into longer contigs (e.g., [9]) but instead there is simple "stacking" or clustering of similar-sequence reads [10]. Our focus here is also on data from single individuals because many studies are barcoding individuals within multiplexed population samples, allowing for bioinformatic quality control and analysis at this informative organismal level when there is at least moderate (5–10 reads per locus) average coverage.

When clustering reads without a reference genome, a major biological factor affecting the potential for artifacts and biased heterozygosity is sequence diversity among paralogs versus alleles. First, the amount of repetitive DNA or large scale genomic duplication in the genome being sequenced is a potential concern, with high levels of recent duplication increasing the chances of mistakenly clustering paralogous sequences with alleles at a locus. However, typical measures of genome complexity and repetitive DNA content, when known, are of uncertain relevance for predicting the frequency of confounding paralogs at the specific scale of sequence read comparisons (~100 bp). Using high read counts to filter out problematic clusters is an intuitive approach to remove highly repetitive loci, but high stochastic variation in read counts among single copy loci makes it practical to remove only the most egregious and obvious artifacts with this type of filter. More recently a "ploidy informed" filter has been proposed wherein clustering involves all high-quality unique sequences but the resulting clusters with >2 distinct sequences per individual are removed from consideration [11, 12]. Because false heterozygosity can still result from clustering of two homozygous paralogous loci, particularly in more homozygous populations, one of our goals was to evaluate the effectiveness of the ploidy-based paralog filter.

Second, many natural populations have a wide range of sequence differences among selectively neutral alleles, and an even wider range must be considered to detect different forms of selection. In many cases, especially in populations with large effective population size and high heterozygosity or recent genome duplication events, the distribution of allelic sequence differences is likely to overlap with the distribution of paralogous sequence differences (Figure 1(a)), increasing the odds that paralogs will create false heterozygosity under a wide range of de novo clustering parameters. It is not possible to distinguish similarly divergent sequences from this distribution overlap without a reference genome, so either to be conservative or to be because of computational constraints, many studies and analysis pipelines consider only minimal allelic differences (1-2 bp per read; [10, 13]). Low allowable sequence differences will tend to split real alleles into separate clusters, or putative loci, creating false homozygosity. Other studies have used a subset of the data for computationally intensive de novo clustering allowing higher sequence difference thresholds, and after discarding clusters deemed to be artifact-prone used the concatenated contigs as a reference for mapping all sequence reads [11]. This latter procedure would benefit from an empirical method for determining, from the data, a minimum sequence difference threshold for clustering that eliminates false homozygosity. The optimum clustering threshold will strike a balance by minimizing false homozygosity, but avoiding higher thresholds because they increase computational demands and have greater potential for false heterozygous calls due to clustering of paralogs. Therefore, our second goal was to develop and assess a protocol for evaluating clustering thresholds using short read data in single individuals in order to avoid arbitrary thresholds and reduce both the amount of discarded data and the homozygosity bias of resulting data.

2. Methods and Materials

2.1. Simulations

2.1.1. Genome Data. Genome assemblies were obtained for three species with published genomes: (1) *Gasterosteus aculeatus* (stickleback) genome assembly gasAcu1.0 ([14]; Broad Institute, http://www.broadinstitute.org/models/stickleback), (2) *Glycine max* (soybean) genome assembly Glyma1.1 ([8]; Joint Genome Institute, http://www.phytozome.net/dataUsagePolicy.php?org=Org_Gmax_v1.1), and (3) *Ciona savignyi* genome assembly version 2.1 ([15]; Broad Institute, http://mendel.stanford.edu/sidowlab/ciona.html).

2.1.2. Genome Repetitiveness Evaluation. For each of the three genome sequences we defined a confounding divergence parameter (CDP) as the amount of divergence between paralogous sequences that would mirror average allelic divergence within the organism. For CDP in each species we used integer values approximating their genomic SNP heterozygosity: 1%, 2%, and 5% for stickleback, soybean, and *Ciona*,

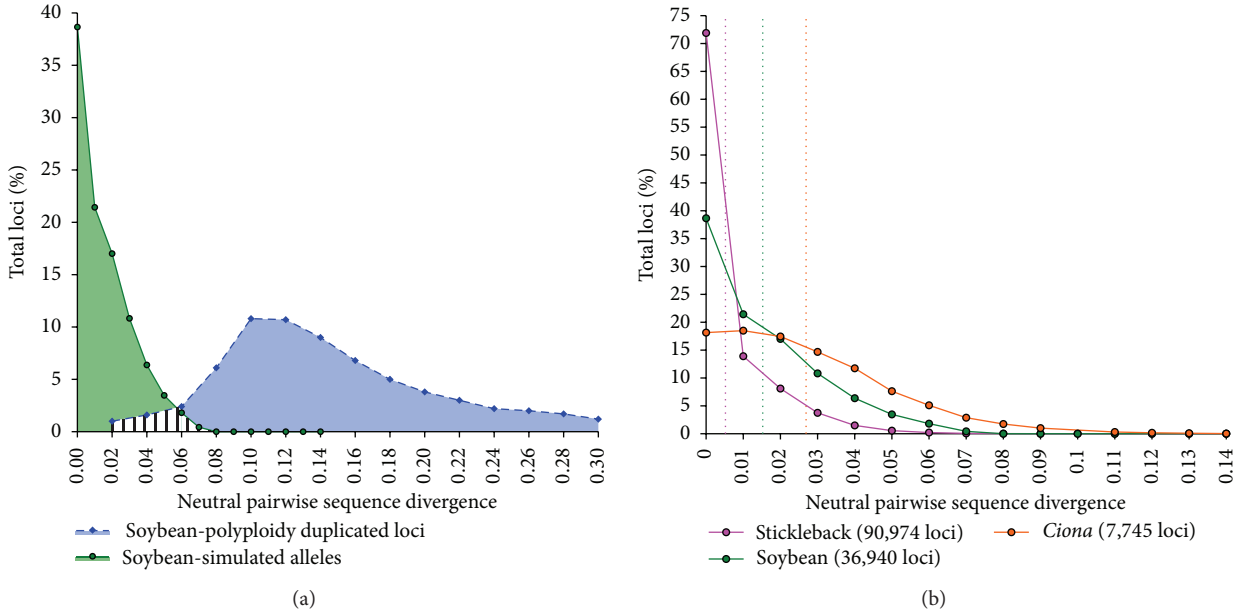


FIGURE 1: Pairwise divergence distribution for simulated allelic loci (green) and polyploidy duplicated loci (blue) in soybean (a), and pairwise divergence distribution of stickleback (pink), soybean (green), and *C. savignyi* (orange) simulated alleles (b). The shaded region in (a) between 0.02 and 0.08 pairwise divergence represents the "confounding duplication" region in which alleles and paralogs are indistinguishable during de novo assembly. The divergence of polyploidy duplicated paralogs in (a) is derived from [8]. Dashed vertical lines in (b) indicate distribution means. Different ranges are used for the *X*- and *Y*-*axis* values in (a) and (b).

respectively [8, 15, 16]. Genome repetitiveness was estimated for each of the three genomes with the corresponding CDP value as follows: using the gem_mappability component of the GEM software library [17], each possible 100 bp fragment in the genome (sliding window, single bp increments) was assigned a value corresponding to the number of times that fragment had a match in the same genome with a divergence less than or equal to the CDP value for that genome. Single copy loci were identified as the class of 100 bp fragments with only one match.

2.1.3. Read Simulation. For the three representative taxa we simulated the equivalent of paired-end Illumina GBS reads from single diploid genomes. We selected the six-base pair sequence GCAGCA as a digest pattern and performed in silico digest of the three genome assemblies using custom scripts. This GC-rich pattern was selected in order to preferentially select gene-rich regions [18, 19] that are typically targeted by the GBS protocol [5]. The sequence was selected to represent a general sampling of these regions, and does not correspond to any specific restriction enzyme. For *Ciona* the version 2.1 assembly used was a merger of the best segments from two haplome assemblies [15]. We removed any in silico digested fragments with a length greater than 800 bp or less than 200 bp, and read the first and last 100 bp of sequence from the remaining fragments.

From this pool of haploid sequences, we simulated an alternative allele for each locus in each species in a coalescent framework. First, we sampled a coalescence time for the two alleles from an exponential distribution with a rate of 1, followed by sampling the number of differences between the simulated alleles from a Poisson distribution with a rate equal to the coalescence time scaled by the average allelic divergence observed in the organism and the length of the sequence. Second, we calculated a per-base pair probability of mutation by dividing the number of differences sampled from the Poisson distribution by the length of the sequence. Finally, the 100 bp sequence (corresponding to a simulated error-free Illumina read) was duplicated and each nucleotide of the duplicate read was changed with probability equal to the per-base pair probability of mutation, replacing it with a nucleotide chosen with equal probability among the other three nucleotides. As a result of this process, for each genome we simulated allele pairs at each locus (referred to below as simulated reads) with a divergence between alleles sampled from a coalescence-adjusted Poisson distribution (Figure 1(b)). Indel polymorphisms and sequencing error were not simulated. A FASTQ [20] format file was generated for these reads, with the basecall quality value set to 40 for each nucleotide.

2.1.4. Simulated Read Clustering and Clustering Threshold Optimization. The simulated reads for each genome were assembled de novo without the use of the reference sequence. First, the pool of reads was reduced to a nonredundant set using the tool SEED [21] allowing for no mismatches or gaps. Second, the nonredundant set was searched against itself for matching reads with a divergence of 15% or less using the tool SlideSort [22]. In order to illustrate the effects of splitting alleles into distinct clusters and combining paralogs

into the same clusters, the 15% pairwise divergence threshold value for the high end of the clustering series was selected such that it was much larger than the largest average allelic pairwise divergence in our dataset (~5% in *Ciona*; [15]) and large enough to include polyploidy duplicated paralogs in soybean (Figure 1; [8]). Each read was paired to every other matching read sequentially, each pair of sequences and their divergence were recorded, and transitive clusters were built for all pairs whose divergence was below a given divergence threshold. A transitive cluster is a grouping of sequences such that, given three sequences A, B, and C, if the divergence between sequence A and B is below the chosen threshold, and the divergence between B and C is also below the threshold, then A, B, and C are grouped together regardless of the divergence between A and C. This clustering process was repeated for each integer threshold between 0 and 15%, and an "optimal threshold" was selected for each organism that maximized the number of clusters with two haplotypes and simultaneously minimized the number of clusters with one haplotype (see discussion). We selected clusters at the optimal mismatch threshold for further evaluation. Simulated data had no redundancy (multiple reads per allele), therefore read count thresholds were not applied and sequence sampling error was not evaluated.

2.1.5. Simulated Clustering Evaluation. For the simulated data each haplotype had a known chromosomal origin, and each locus contained at most two haplotypes. Therefore, the heterozygosity of the simulated data can be computed directly, and the difference between the exact simulated heterozygosity and the inferred heterozygosity from de novo clustering represents the expected inaccuracy of assemblies for the given organism that results solely from the interaction between simulated allelic divergence, the amount of genome duplication, and the clustering threshold employed.

2.2. *Ciona savignyi* Sequencing and Clustering

2.2.1. Source Material, Library Preparation, and Sequencing. Genomic DNA from a wild-collected single diploid *Ciona savignyi* individual, previously used for genomic sequencing [15], was kindly provided by A. Sidow. The DNA was prepared for genotyping-by-sequencing as described in Elshire et al. [5] and barcoded separately from other samples. Paired-end (2×88 nucleotides) data was collected on an Illumina Genome Analyzer (Illumina Inc., San Diego, California, USA) in the Cornell Laboratory Core.

2.2.2. Sequence Cleanup and Clustering. Using custom scripts, raw sequence files in the QSEQ format [23] for an individual were scanned for any barcode tag, sequencing adaptor, and enzyme cut site sequence and these were trimmed from the sequence ends. After trimming we removed from analysis any sequences shorter than 75 bp, containing internal enzyme cut sites, or containing at least one nucleotide quality value less than 20. The remaining sequences were assembled into unique tag clusters using the same pipeline as the simulated reads (see above), with the following modifications: (1) any tag that did not have at least four confirming reads was discarded from further analysis prior to the nonredundant reduction step using SEED, and (2) in the SlideSort pairwise matching step, internal indel mismatches were counted as one mismatch regardless of indel size, and edge indels were not counted towards the total number of mismatches between two reads. Finally, to reduce the probability of false homozygosity due to sequence sampling error below 1%, given a minimum of 4 reads supporting each allele (calculations in Supplementary Material S1; see S1 in Supplementary Material available online at http://dx.doi.org/10.1155/2014/675158), we only analyzed clusters (putative loci) that had 11 or more reads assigned to them within the single individual.

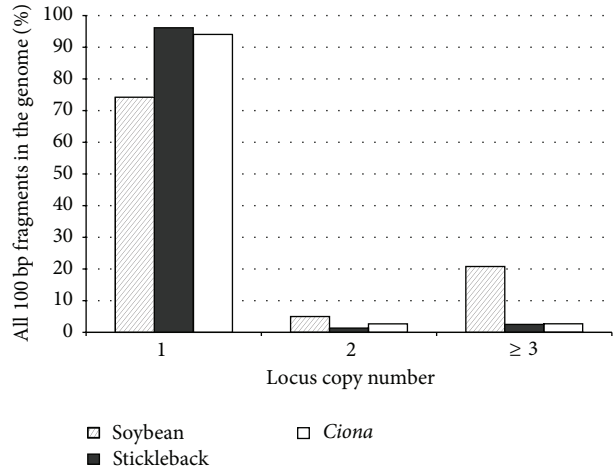


Figure 2: Potentially confounding duplication across all possible 100 bp fragments in three genomes. At this scale of comparison most loci are single-copy in each genome (copy class 1) and do not present a paralog clustering problem when trying to de novo assemble alleles. Paralogous loci from copy class 2 have the potential to be clustered together and if each paralog is homozygous, the cluster will be incorrectly interpreted as a diploid heterozygous locus even after a ploidy-informed filter. In principle, a ploidy-informed filter should be able to identify and remove most clusters derived from loci in the 3+ copy class.

3. Results

3.1. Genome Characteristics. The three genomes selected for study vary greatly in size, GC content, and amount of duplication within the genome. For stickleback, the reference genome sequence is 401 Mbp in size with a GC content of 43.49%. The soybean reference genome is 974 Mbp in size with a GC content of 34.10%, and the *C. savignyi* reference genome is 177 Mbp in size with a GC content of 35.89%. The amount of genome repetitiveness is presented in Figure 2, with the 1-copy class representing the proportion of 100 bp loci that had a single copy in the genome, the 2-copy class representing the proportion of loci that had exactly one other (paralogous) match in the genome, and the 3+ copy class representing the proportion of loci with multiple matches across the genome. As expected given its history of whole

genome duplication [24], soybean had the largest percentage of loci duplicated across the genome. When sequencing reduced representation libraries, sequences sampled from the 2 and 3+ copy classes for a given organism have the potential to confound de novo clustering of reads as their paralogous divergence is within the expected divergence range for alleles within the organism (Figure 1(a)).

3.2. Simulated and Sequenced Reads. Summary of statistics for the simulated GBS data are presented in Table 1. For the experimental data from a *C. savignyi* individual, we obtained 12,188,018 high quality reads of 75 bp each. These reads were collapsed into 3,033,429 nonredundant sequences (representing both allelic variants and different loci). Of these, 427,919 (14.11%) had 4 or more reads supporting the tag sequence, and these were used for further analysis.

3.3. Clustering Threshold Series. For each of the three simulated data sets and the experimental *Ciona* data, we generated clustering threshold series in order to evaluate the impact of the maximum allowable mismatch threshold on the proportion of clusters free from artifacts. The simulated data for all three species showed a trend with increasing mismatch thresholds from 2% to 14% whereby single haplotype clusters (putative homozygous loci) decreased and two-haplotype clusters (putative heterozygous loci) increased (Figures 3(a)–3(c)). The greatest degree of disparity between simulated read cluster heterozygosity and true simulated heterozygosity occurred at low mismatch thresholds. For simulated stickleback reads, clusters with only one haplotype were minimized at mismatch thresholds of 4% and greater; at lower mismatch values oversplit alleles inflated the total cluster number and apparent homozygosity (Figure 3(a)). The number of clusters with 3 or more distinct sequences was negligible at all thresholds, and therefore the number of both homozygous and heterozygous clusters converged to the true value once the clustering threshold was raised to ≥4%. For simulated soybean (Figure 3(b)) the symmetry of changes in cluster proportions is more apparent as increasingly larger mismatch thresholds convert oversplit 1-haplotype clusters into heterozygous clusters. The magnitude of the artifact at low mismatch thresholds scales with heterozygosity. As mismatch thresholds increase, the inferred number of heterozygous clusters increases dramatically in both soybean (37% to 55%, Figure 3(b)) and *Ciona* (~35% to 77%, Figure 3(c)). The proportion of paralogous clusters slowly increases with increasing mismatch threshold but is typically small, reaching a maximum among these species of 11% in soybean at a 14% mismatch threshold. Because artifactual heterozygosity in the 2-haplotype class is expected to be minor, these results suggest that an optimum mismatch threshold would be the point at which there exists 1-haplotype and 2-haplotype cluster proportions asymptote or 4%, 8%, and 12% maximum sequence difference for stickleback, soybean, and *Ciona*, respectively (Figure 3, vertical dashed line). At higher mismatch thresholds there may be diminishing returns, with paralogs contributing to more new 2-haplotype clusters than collapsing oversplit clusters. The experimental *C. savignyi* data (Figure 3(d), sequence differences included variant nucleotides and indels) showed the same tendency to asymptote at higher sequence difference thresholds, but a large bias toward homozygosity remained at the asymptote.

3.4. Computational Streamlining of Threshold Determination. For experimental data sets from high heterozygosity species it can be computationally intensive to determine the cluster category distribution at large sequence difference thresholds. For the *C. savignyi* experimental data the calculation across the entire clustering threshold range took approximately 4 days on a modern quad-core Intel processor machine with 16 GB RAM, with the time limiting step being the SlideSort pairwise alignment. Two procedures were implemented to streamline these computations. First, clustering with a full all by all alignment was restricted to the maximum sequence difference threshold, and cluster results at lesser thresholds were rapidly estimated by pruning according to pairwise sequence difference. Specifically, we recorded the pairwise distance between all reads in the set used for maximum difference threshold clustering. Then to estimate cluster results for each lower threshold, read pairs with a divergence above this new threshold were separated as unpaired reads, and transitive clusters were built anew to account for multiread clusters being separated into new 1-read and 2-read clusters. Second, in order to evaluate whether subsampled data could generate robust clustering threshold series to determine an asymptote, we generated the maximum divergence clustering series (14 bp difference) using subsets of the *C. savignyi* experimental data uniformly subsampled without replacement to represent a range between 10% and 100% of the total reads. It should be noted that subsampling can inflate specific estimates of homozygosity due to unsampled alleles, and estimates of these specific values should be done with the full data set. However, as shown in Figure S2 the shape of the clustering series curve generally holds at all levels of subsampling, and it quickly converges to the shape and values of the full data set once 20–30% or more of the data are used.

4. Discussion

In order to explore genomic repetitiveness and allelic divergence distributions with simulated data, we selected three reference genomes from organisms with very different genomic parameters. The two genomes with low amounts of duplication (stickleback and *Ciona*, Figure 2) have very low and very high heterozygosity, with 0.5% and ~5% average divergence between alleles in stickleback and *Ciona*, respectively. The soybean genome, by contrast, has a history of polyploidy leading to large parts of the genome being maintained in multiple copies (Figure 1) but its heterozygosity falls in between the two extremes above. Simulated data from these genomes is therefore expected to illustrate the effects of duplication and heterozygosity on clustering quality under different read clustering regimes. However, since additional factors contribute to clustering errors in an experimental dataset, we also compared the results of our simulations with de novo assemblies of reads from genotyping-by-sequencing

Table 1: Simulated restriction-based genome sampling for the three species.

Species	Total fragments	200–800 bp fragments	Total 100 bp reads	Homozygous Tags percentage	Heterozygous Tags percentage	Mean allele divergence
Stickleback	345,704	90,974 (26.32%)	181,948	72%	28%	0.5%
Soybean	269,756	36,940 (13.69%)	73,880	39%	61%	1.5%
Ciona	61,102	7,746 (12.68%)	15,492	18%	82%	2.7%

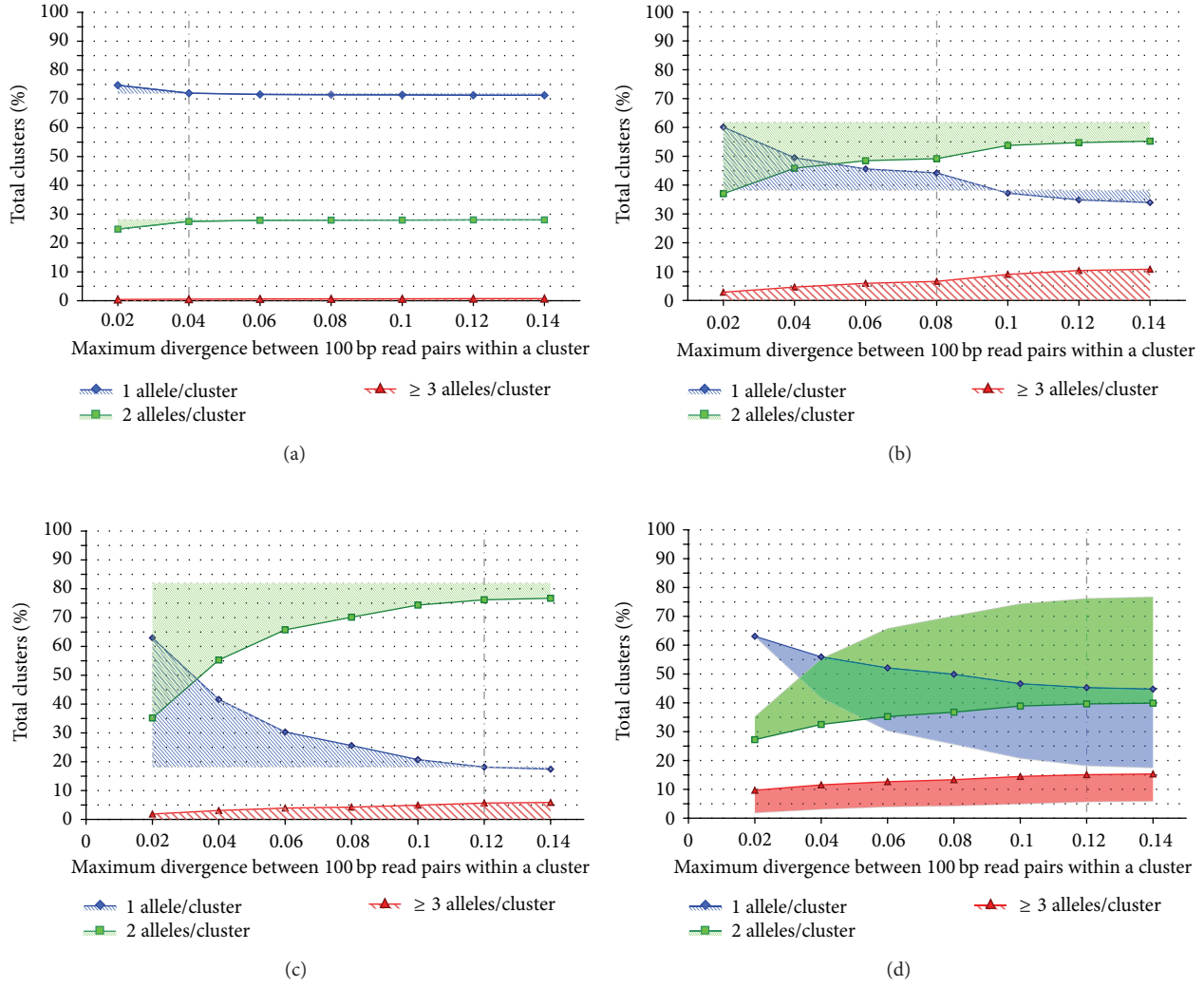


Figure 3: Clustering mismatch threshold series for simulated stickleback reads (a), simulated soybean reads (b), simulated *C. savignyi* reads (c), and experimental *C. savignyi* data (d). The *Y-axis* represents the percentage of total clusters for a given organism at a given mismatch value and the *X-axis* represents the maximum proportion of differences (mismatches) allowed between reads within a cluster. Single haplotype clusters (putative homozygous loci) are represented by a solid blue line and diamonds, two-haplotype clusters (putative heterozygous loci) are represented by a solid green line and squares, and three or more haplotype clusters (combined alleles from 2 or more paralogous loci) are represented by a solid red line and triangles. Striped shaded areas for simulated data represent deviation from the true values due to assembly artifacts such as splitting alleles into different clusters when the threshold is low or combining paralogs into a cluster when the threshold is high. Solid shaded areas for experimental data represent deviation from the expected simulated values due to assembly artifacts and null alleles. The uptick in heterozygosity observed between 0.8 and 0.10 in (b) is likely a result of the surge in 2+ paralog clustering over the same interval, perhaps due to clustering of duplicated loci from the soybean polyploidy event. The mean values of total cluster counts across assembly thresholds for plots (a)–(d) are 88300 (SD = 1987), 29201 (SD = 4798), 7700 (SD = 1244), and 167824 (SD = 10863).

results on the genome of the same individual *C. savignyi* specimen that was used to generate the publicly available reference genome sequence.

4.1. The Paralogs Problem. Regarding the issue of false heterozygous clusters due to the clustering of paralogs, our simulated data shows that even in the case of a highly duplicated genome their impact is relatively low (Figure 3(b)). Partly, this is because only duplications with divergences within the same range as allelic divergence present a problem for clustering. With the exception of extremely recent polyploidy events, the majority of maintained duplicated loci are expected to be much older (have much higher divergence) than the alleles in a population, resulting in only a small number of duplicated loci exhibiting "confounding divergence." For example, duplicates from the most recent polyploidy event in soybean (~14 million years ago (mya); [25]) have a mean pairwise divergence of ~10–20% [8, 24, 25]. The age of the soybean polyploidy event is typical of most recent polyploidy events in other crop species such as maize (~12 mya; [26]) and cotton (~13–20 mya; [27]). Therefore, a ploidy-informed cut-off on the number of haplotypes per cluster per individual appears to be sufficient to address this issue in most cases. This result suggests that when devising clustering methods that minimize paralog clusters while also minimizing the splitting of divergent alleles into artifactual 1-haplotype clusters, the former problem is largely ameliorated by filtering out 3+ haplotype clusters. This provides flexibility to apply generous sequence difference thresholds to maximize allelic clusters without large loss of data from the elimination of putative paralog clusters. In our simulated soybean data, only ~7% of putative loci were eliminated by this filter when selecting the maximum 8 differences threshold (Figure 3(b)).

4.2. Insights from Simulated Data. Given the small amount of paralog-induced erroneous clustering observed, the clustering quality is mainly reflected by the change in relative proportions of 1-haplotype and 2-haplotype clusters over different sequence difference thresholds. As the clustering threshold increases up to the maximum divergence between alleles, oversplit allelic clusters are collapsed, decreasing the proportion of 1-haplotype clusters, with a concomitant increase in 2-haplotype clusters. Clustering thresholds higher than allelic divergence are not expected to alter observed counts of 1-haplotype and 2-haplotype clusters except for new clustering errors from paralogous loci. Indeed, as the simulated data in Figure 3 shows, cluster counts plateau once the threshold reaches a certain value: 4 bp difference for stickleback, 8 bp for soybean, and 12 bp for *Ciona*. Only soybean exhibits a secondary abrupt increase in heterozygous clusters at high clustering thresholds due to its polyploid nature. This suggests that as paralog clusters accumulate at higher mismatch thresholds in species that have not recently undergone whole genome duplication, they mostly fall into the 3+ read class and are easily filtered out. Very few 2-haplotype clusters are accounted for by clustering two homozygous paralogs. These simulated data comparisons suggest that the clustering threshold series asymptote marks an optimal clustering threshold regardless of genomic heterozygosity because at that point collapsing of alleles is fully realized but hidden paralog artifacts are typically minimal. Even for model species the distribution of allelic differences in a particular population will be unknown, so it is this empirical basis for identifying the optimum clustering threshold that makes this protocol valuable.

4.3. Evaluating Clustering Thresholds with Real Data. From a computational perspective, the main challenge an investigator faces when attempting de novo clustering in a high heterozygosity organism is the practical computational constraints on clustering with large sequence divergences allowed. Many tools such as Stacks [10] use a k-mer based approach to clustering, and its memory requirements increase exponentially with the average number of allowed mismatches between reads. Therefore, it is often impractical to attempt clustering with more than 2 or 3 mismatches allowed between reads, and in fact some tools limit this parameter to one mismatch (e.g. UNEAK [13]). As our simulated data show (Figure 3), low threshold clustering can easily misidentify heterozygous loci as homozygous by splitting alleles into different locus clusters even in species with very low heterozygosity. For example, a 2.5% difference in estimated heterozygosity is observed for stickleback when moving from 2% to 4% sequence difference thresholds (Figure 3(a)). However, while extremely useful in assembling partially overlapping reads at a locus, k-mer based approaches are not necessary (or ideal) for analyzing restriction based RAD-tag or GBS type data that produce sequence reads always starting at the same position for a given locus. Such data are better suited for a classic all-versus-all read comparison, which does not have as severe computational divergence limitation. We chose SlideSort [22] for our analysis, but other similar tools have been available for over a decade and have been used by other groups for similar purposes (e.g., CAP3, BLAST).

From a biological perspective, ideal clustering thresholds are highly species dependent and therefore a priori thresholds that worked well in one species might not work in another. Our clustering series approach allows an investigator to determine an optimal threshold for a given sample directly from the data. To our knowledge, this is the first time such an approach has been presented, and the very different optimal thresholds obtained for the three species simulated here suggest that such an analysis is worthwhile in order to improve the quality of de novo assemblies of restriction-based genome samples.

4.4. Null Alleles. An important question is whether the approach presented here will work as well to identify clustering mismatch thresholds in experimental data as it does with simulated data. At first glance the large difference between homozygosity asymptotes in simulated versus experimental *Ciona* data suggests caution in generalizing from the former. However, the observed discrepancy can easily be explained as an expected result of restriction-based sampling of a high heterozygosity species. Heterozygous restriction

sites will generate sequence data for only one of the two homologous chromosomes. The frequency of loci with these null alleles is directly proportional to genomic heterozygosity. Impacts of this biased sampling include inflated estimates of homozygosity, increased coalescent variance across loci, and increased heterogeneity of locus sampling among individuals (for polymorphic restriction sites, the locus will not get sampled in homozygous null individuals). In population samples the bias can be effectively reduced by narrowing analyses to the set of loci represented in all individuals [28], but only at the cost of large data loss, so model-based filters are needed. Methods for estimating the proportion of unsampled loci due to heterozygous restriction sites have been previously developed [29], and based on these calculations, we expect at least 24% of the loci in *C. savignyi* to be falsely identified as homozygous due to variation at restriction sites resulting in null alleles. Homozygosity in *C. savignyi* simulated data, at the optimal clustering threshold and with no null alleles, is 18%. Therefore, the expected homozygosity in the experimental data when accounting for the expected null alleles is at least 42%, in agreement with the observed 45% from the experimental data.

5. Conclusions

In general, the clustering of paralogous sequences is a minor source of error in restriction-based reduced representation data for many species and can be effectively addressed by ploidy-based filtering. False homozygous loci, however, can produce a potentially strong bias, and both mechanisms potentially generating this bias, oversplit alleles and null alleles, have effects that scale with heterozygosity. We have provided formulae to calculate the read count filter necessary to achieve a given expected amount of false homozygous loci due to lack of second allele sampling during sequencing (Eq. iii, Supplementary 1). More importantly, we demonstrated an empirical procedure to determine a clustering mismatch threshold that minimizes the splitting of alleles into artifactual 1-haplotype clusters. The clustering threshold series approach can help select the appropriate cut-off for clustering the full data set without requiring any a-priori knowledge about heterozygosity in the organism under study, and this diagnostic step can be done with a subsample of the data in order to accelerate analysis.

Conflict of Interests

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

References

[1] D. G. Peterson, S. R. Schulze, E. B. Sciara et al., "Integration of cot analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery," *Genome Research*, vol. 12, no. 5, pp. 795–807, 2002.

[2] S. R. McCouch, G. Kochert, Z. H. Yu et al., "Molecular mapping of rice chromosomes," *Theoretical and Applied Genetics*, vol. 76, no. 6, pp. 815–829, 1988.

[3] P. D. Rabinowicz, K. Schutz, N. Dedhia et al., "Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome," *Nature Genetics*, vol. 23, no. 3, pp. 305–308, 1999.

[4] N. A. Baird, P. D. Etter, T. S. Atwood et al., "Rapid SNP discovery and genetic mapping using sequenced RAD markers," *PLoS ONE*, vol. 3, no. 10, Article ID e3376, 2008.

[5] R. J. Elshire, J. C. Glaubitz, Q. Sun et al., "A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species," *PLoS ONE*, vol. 6, no. 5, Article ID e19379, 2011.

[6] C. P. van Tassell, T. P. L. Smith, L. K. Matukumalli et al., "SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries," *Nature Methods*, vol. 5, no. 3, pp. 247–252, 2008.

[7] J. W. Davey, P. A. Hohenlohe, P. D. Etter, J. Q. Boone, J. M. Catchen, and M. L. Blaxter, "Genome-wide genetic marker discovery and genotyping using next-generation sequencing," *Nature Reviews Genetics*, vol. 12, no. 7, pp. 499–510, 2011.

[8] J. Schmutz, S. B. Cannon, J. Schlueter et al., "Genome sequence of the palaeopolyploid soybean," *Nature*, vol. 463, no. 7278, pp. 178–183, 2010.

[9] P. D. Etter, J. L. Preston, S. Bassham, W. A. Cresko, and E. A. Johnson, "Local de novo assembly of rad paired-end contigs using short sequencing reads," *PLoS ONE*, vol. 6, no. 4, Article ID e18561, 2011.

[10] J. M. Catchen, A. Amores, P. Hohenlohe, W. Cresko, J. H. Postlethwait, and D.-J. de Koning, "Stacks: building and genotyping loci de novo from short-read sequences," *G3, Genes, Genomes, Genetics*, vol. 1, no. 3, pp. 171–182, 2011.

[11] T. L. Parchman, Z. Gompert, J. Mudge, F. D. Schilkey, C. W. Benkman, and C. A. Buerkle, "Genome-wide association genetics of an adaptive trait in lodgepole pine," *Molecular Ecology*, vol. 21, no. 12, pp. 2991–3005, 2012.

[12] B. K. Peterson, J. N. Weber, E. H. Kay, H. S. Fisher, and H. E. Hoekstra, "Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species," *PLoS ONE*, vol. 7, no. 5, Article ID e37135, 2012.

[13] F. Lu, A. E. Lipka, J. Glaubitz et al., "Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol," *PLoS Genetics*, vol. 9, no. 1, Article ID e1003215, 2013.

[14] F. C. Jones, M. G. Grabherr, Y. F. Chan et al., "The genomic basis of adaptive evolution in threespine sticklebacks," *Nature*, vol. 484, no. 7392, pp. 55–61, 2012.

[15] K. S. Small, M. Brudno, M. M. Hill, and A. Sidow, "A haplome alignment and reference sequence of the highly polymorphic Ciona savignyi genome," *Genome Biology*, vol. 8, no. 3, article R41, 2007.

[16] P. A. Hohenlohe, S. J. Amish, J. M. Catchen, F. W. Allendorf, and G. Luikart, "Next-generation RAD sequencing identifies thousands of SNPs for assessing hybridization between rainbow and westslope cutthroat trout," *Molecular Ecology Resources*, vol. 11, no. 1, pp. 117–122, 2011.

[17] T. Derrien, J. Estellé, S. M. Sola et al., "Fast computation and applications of genome mappability," *PLoS ONE*, vol. 7, no. 1, Article ID e30377, 2012.

[18] B. Aïssani and G. Bernardi, "CpG islands, genes and isochores in the genomes of vertebrates," *Gene*, vol. 106, no. 2, pp. 185–195, 1991.

[19] A. T. Sumner, J. de la Torre, and L. Stuppia, "The distribution of genes on chromosomes: a cytological approach," *Journal of Molecular Evolution*, vol. 37, no. 2, pp. 117–122, 1993.

[20] P. J. A. Cock, C. J. Fields, N. Goto, M. L. Heuer, and P. M. Rice, "The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants," *Nucleic Acids Research*, vol. 38, no. 6, pp. 1767–1771, 2009.

[21] E. Bao, T. Jiang, I. Kaloshian, and T. Girke, "SEED: efficient clustering of next-generation sequences," *Bioinformatics*, vol. 27, no. 18, pp. 2502–2509, 2011.

[22] K. Shimizu and K. Tsuda, "Slidesort: all pairs similarity search for short reads," *Bioinformatics*, vol. 27, no. 4, pp. 464–470, 2011.

[23] Illumina Incorporation, Sequencing Analysis Software User Guide for Pipeline Version 1.4 and CASAVA Version 1.0, 2009.

[24] R. C. Shoemaker, J. Schlueter, and J. J. Doyle, "Paleopolyploidy and gene duplication in soybean and other legumes," *Current Opinion in Plant Biology*, vol. 9, no. 2, pp. 104–109, 2006.

[25] J. A. Schlueter, P. Dixon, C. Granger et al., "Mining EST databases to resolve evolutionary events in major crop species," *Genome*, vol. 47, no. 5, pp. 868–876, 2004.

[26] Z. Swigonová, J. Lai, J. Ma et al., "Close split of sorghum and maize genome progenitors," *Genome Research*, vol. 14, no. 10A, pp. 1916–1923, 2004.

[27] K. Wang, Z. Wang, F. Li et al., "The draft genome of a diploid cotton Gossypium raimondii," *Nature Genetics*, vol. 44, no. 10, pp. 1098–1103, 2012.

[28] J. W. Davey, T. Cezard, P. Fuentes-Utrilla, C. Eland, K. Gharbi, and M. L. Blaxter, "Special features of RAD Sequencing data: implications for genotyping," *Molecular Ecology*, vol. 22, no. 11, pp. 3151–3164, 2013.

[29] W. B. Upholt, "Estimation of DNA sequence divergence from comparison of restriction endonuclease digests," *Nucleic Acids Research*, vol. 4, no. 5, pp. 1257–1265, 1977.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 325697, 14 pages
http://dx.doi.org/10.1155/2014/325697

Research Article

MPINet: Metabolite Pathway Identification via Coupling of Global Metabolite Network Structure and Metabolomic Profile

Feng Li,[1] Yanjun Xu,[1] Desi Shang,[1] Haixiu Yang,[1] Wei Liu,[1,2] Junwei Han,[1] Zeguo Sun,[1] Qianlan Yao,[1] Chunlong Zhang,[1] Jiquan Ma,[1,3] Fei Su,[1] Li Feng,[1] Xinrui Shi,[1] Yunpeng Zhang,[1] Jing Li,[1] Qi Gu,[1] Xia Li,[1] and Chunquan Li[1,4]

[1] *College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, China*
[2] *Department of Mathematics, Heilongjiang Institute of Technology, Harbin 150050, China*
[3] *Department of Computer Science and Technology, Heilongjiang University, Harbin 150080, China*
[4] *Department of Medical Informatics, Harbin Medical University, Daqing Campus, Daqing 163319, China*

Correspondence should be addressed to Xia Li; lixia@ems.hrbmu.edu.cn and Chunquan Li; lcqbio@aliyun.com

Received 1 April 2014; Accepted 18 May 2014; Published 25 June 2014

Academic Editor: Li-Ching Wu

High-throughput metabolomics technology, such as gas chromatography mass spectrometry, allows the analysis of hundreds of metabolites. Understanding that these metabolites dominate the study condition from biological pathway perspective is still a significant challenge. Pathway identification is an invaluable aid to address this issue and, thus, is urgently needed. In this study, we developed a network-based metabolite pathway identification method, MPINet, which considers the global importance of metabolites and the unique character of metabolomic profile. Through integrating the global metabolite functional network structure and the character of metabolomic profile, MPINet provides a more accurate metabolomic pathway analysis. This integrative strategy simultaneously captures the global nonequivalence of metabolites in a pathway and the bias from metabolomic experimental technology. We then applied MPINet to four different types of metabolite datasets. In the analysis of metastatic prostate cancer dataset, we demonstrated the effectiveness of MPINet. With the analysis of the two type 2 diabetes datasets, we show that MPINet has the potentiality for identifying novel pathways related with disease and is reliable for analyzing metabolomic data. Finally, we extensively applied MPINet to identify drug sensitivity related pathways. These results suggest MPINet's effectiveness and reliability for analyzing metabolomic data across multiple different application fields.

1. Introduction

The development of high-throughput metabolomics technology, such as nuclear magnetic resonance and approaches based on mass chromatography, has enabled us to obtain metabolomic profiles of large numbers of metabolites [1]. The increasing availability of high-throughput data and large-scale high-quality pathway sources, such as KEGG [2] and Reactome [3], provides the potential for understanding these metabolomic data at the pathway level. Many currently available pathway-identification approaches are effective in transcriptomics [4, 5], though a method that effectively incorporates both global biological network structure and metabolomic profile is urgently needed.

Several computational approaches for metabolite pathway analysis have been developed recently, including over-representation analysis (ORA) and set enrichment analysis (SEA) [6]. ORA is widely used by researchers from a statistical perspective. It subjects a list of interesting metabolites (e.g., differential metabolites) to statistical analysis to detect whether the given metabolite set is overrepresented in a predefined pathway. For example, metabolite biological role (MBRole) [7] and metabolite pathway enrichment analysis (MPEA) [8] used ORA to perform pathway analysis based on

metabolite sets. Metabolite set enrichment analysis (MSEA) [9] is a classic method of SEA, which involves enrichment analysis based on the whole list of metabolites identified in the profile, and it is also taking into consideration metabolite concentrations. There is no doubt that pathway analysis should involve a statistical model in order to reduce the incidence of false positive identification of differential metabolites based on a computational approach. However, both ORA and SEA consider a pathway as a simple metabolite set; they ignore functional interactions among metabolites and treat all metabolites equivalently. From a biological view, some metabolites should receive more attention than others and this is defined as the nonequivalent roles of metabolites in the pathway. Recently, a topology-based pathway analysis method, metabolomics pathway analysis (MetPA), which considers the local nonequivalence of metabolites in an individual pathway, was designed by Xia and Wishart to effectively improve pathway identification [10]. However, individual pathways are components of biological networks, and metabolites in different pathways also have functional interactions [11, 12]. A network-based approach that takes account of functional interactions to evaluate the nonequivalence of metabolites from a global perspective is therefore more suitable.

Also, the currently available approaches of metabolomic functional analysis ignore several common important aspects. Firstly, from a biological point of view, some metabolites participate in many pathways; they are referred to as common metabolites, while other metabolites only participate in a few pathways, which are referred to as pathway-specific metabolites. If a pathway identification method is mainly based on common metabolites, there will be more false positive pathways due to the presence of metabolites that are involved in multiple pathways. In contrast, it is more reliable that metabolites within an identified pathway tend to be pathway-specific as these dysfunctional metabolites only particularly involved in this pathway. Thus, these pathway-specific metabolites are more important in the pathway identification.

Secondly, dysfunction of metabolites may be compensated for by their functional partners in the biological network. Metabolites involved in a number of pathways usually have many functional partners, while other metabolites only participate in a few pathways and thus have few functional-compensation partners. If metabolites within a pathway tend to be pathway-specific, its deregulated signals are likely to be amplified to the entire pathway. In contrast, if the metabolites in a pathway tend to have many functional partners in the biological network, the dysregulation signal is more likely to be alleviated. Thus, the nonequivalence roles of metabolites should be considered in order to identify pathways correctly.

Finally, from the perspective of metabolomic technology, there is bias existing in metabolite identification. Most current metabolomic technologies usually only analyze a small fraction of the entire metabolome (5–10%) [6], and the identified metabolites are not randomly selected. These identified metabolites in the profiles are preferred to be well studied. This may be due to the fact that metabolite identification depends heavily on *a priori* knowledge [13, 14]. This metabolite identification bias will lead to the subsequent pathway identification bias because these well-studied metabolites usually reside in fundamental pathways, such as glutathione metabolism, or the citrate cycle. These fundamental pathways are thus likely to be inappropriately identified when metabolites are treated equivalently in pathways. In contrast, pathways that do not contain many well-studied metabolites (i.e., metabolites that are difficult to be identified in the profile) are more likely to be ignored by most of the current methods. It is therefore essential to consider the bias from metabolomic experimental technology. Furthermore, in contrast to transcriptome analysis, which can provide thousands of differential molecules for functional analysis, metabolomics usually use only dozens of metabolites for pathway identification. However, most of the metabolite pathway analysis methods currently proposed were originally designed for transcriptomics.

In this study, we developed a network-based pathway-analysis method called MPINet that considers both the global nonequivalence of metabolites and the bias from metabolomics experimental technology. In addition, a classical statistical model is also integrated. We constructed a human metabolite functional network. The global nonequivalence of metabolites within pathways refers to the different importance of metabolites and was evaluated based on the global functional interactions of metabolites within networks. Initial bias scores of metabolites were assigned based on the metabolomic profile, and a monotonic cubic regression spline model was then fitted to integrate the global nonequivalence scores and the initial bias scores of metabolites, and weights were assigned to the metabolites (nodes) in the network. Finally, the pathway weight, which was calculated based on the global node-weighted network, was used as the parameter for Wallenius approximation [15] to evaluate the significance of the pathway. We applied MPINet to four datasets and demonstrated the ability of MPINet to identify biologically meaningful pathways successfully and reproducibly across multiple different application fields. MPINet has been implemented as a free web-based (http://bioinfo.hrbmu.edu.cn/MPINet/) and R-based tool (http://cran.r-project.org/web/packages/MPINet/), supporting 3350 human pathways across 10 databases.

2. Materials and Methods

2.1. Datasets and Processing. We analyzed one metastatic prostate cancer dataset, two type 2 diabetes datasets, and one drug-sensitivity dataset. These datasets were obtained from metabolomics experiments or manually extracted from the literature.

2.1.1. Metastatic Prostate Cancer Dataset. The metabolomic profile of prostate cancer was obtained from the study of Sreekumar et al. [16]. It included 16 tissue samples from benign adjacent prostate, 12 from localized prostate cancer, and 14 from metastatic prostate cancer samples [16]. In this study, we used the localized prostate cancer and metastatic prostate cancer samples as a case-control study. Differential

metabolites were determined by Wilcoxon's rank-sum test ($P < 0.1$). Finally, 92 metabolites were identified as metastatic prostate cancer differential metabolites and used for pathway analysis.

2.1.2. Type 2 Diabetes Datasets. We analyzed two type 2 diabetes datasets: type 2 diabetes dataset 1 and type 2 diabetes dataset 2. We extracted type 2 diabetes-associated metabolites from the HMDB database [17] and text mining (supplementary Table S1 in Supplementary Material available online at http://dx.doi.org/10.1155/2014/325697). Finally, 65 metabolites were identified as type 2 diabetes-associated metabolites.

Dataset 2 was obtained from Suhre et al. [18]. Multiplatform metabolomic profiles, including 482 metabolites, were detected from 40 individuals with type 2 diabetes and 60 control individuals. We selected the differential metabolites from the data preprocessed by Suhre et al. ($P < 0.05$) and converted the metabolite names to PubChem CIDs. Sixty-six metabolites related to type 2 diabetes were used for subsequent analysis.

2.1.3. Drug-Sensitivity Dataset. A total of 121 drugs selected from Weinstein et al. [19] were analyzed. Drug-sensitivity data based on GI50 values were obtained from the CellMiner database [20], and metabolite measurement data were obtained from Metabolon, downloaded from http://dtp.nci.nih.gov/mtargets/download.html. The drug-sensitivity dataset included the −log(GI50) data for drugs in replicated experiment across NCI-60 cell lines. For each drug, the −log(GI50) values from different experiments were averaged. The metabolite measurement dataset included 352 metabolites across 58 cell lines, of which 160 named metabolites mapped to 159 distinct PubChem CIDs.

For each drug, we calculated the Pearson correlation of the drug −log(GI50) value and the metabolite measurements obtained in the experiment by Metabolon across 58 NCI-60 cell lines. The Benjamin method was used to correct the P value. The significant drug-sensitivity-related metabolite cutoff was set at 60%. MPINet was then applied to drug-sensitivity-related metabolites to identify drug-sensitivity-associated pathways.

2.2. Methods. We have implemented MPINet as a freely available R-based (http://cran.r-project.org/web/packages/MPINet/) and web-based tool (http://bioinfo.hrbmu.edu.cn/MPINet/), supporting 3350 human pathways across 10 databases, including KEGG, Reactome, PID, and Wikipathways from ConsensusPathDB [21]. Input only requires a list of the metabolites of interest (e.g., differential metabolites). Figure 1 gives a schematic overview of MPINet.

2.2.1. Construction of the Global Edge-Weighted Human Metabolite Network. The preliminary metabolite interaction network was downloaded from the STITCH database (http://stitch.embl.de/) [22]. We constructed the global edge-weighted human metabolite network from the preliminary network as follows. First, we extracted stereospecific compound interactions from the chemical-chemical link file and used the "combined score" in STITCH, as the initial edge weight. Second, we collected human metabolites from a wide range of databases, including KEGG [2], HMDB [17], Reactome [3], MSEA [9], and SMPDB [23]. We obtained the total of 4994 human metabolites from these five databases. Third, we extracted human metabolite interactions from the preliminary network obtained in the first step. We extracted the interacting pair if both of the metabolites in the interacting pair were included in the 4994 human metabolites. Finally, a global edge-weighted human metabolite network was constructed, which contained 3764 nodes and 74667 weighted edges (Table S2).

2.2.2. Calculating the Global Nonequivalence Scores of Metabolites Based on the Network. Based on the notion that dysfunction of metabolites with strong functional-interaction partners will be more easily compensated for, we defined a global nonequivalence score (GN score) to measure the functional interaction between a metabolite and its functional partners in the global metabolite network.

Firstly, we quantified the functional interactions between each pair of metabolites in the network by calculating the global connection strength (GCS) for each metabolite pair. The GCS measure was defined as the modified version of the strength of connection (SOC) measure in Campbell et al. [24], which measured the connection strength between metabolites from a global perspective by considering both the number and length of multiple paths between two metabolites in the network. A detailed description of how the GCS values were calculated was included in the supplementary text. Higher GCS values indicate stronger functional interactions between the metabolite pairs.

We then defined the GN score of a given metabolite as the mean of the GCS values between it and the other metabolites in the network. A high GN score indicates that the metabolite has strong functional interactions with its functional partners, and dysfunction of these metabolites is thus likely to be compensated for. In contrast, metabolites with low GN scores have weaker interactions, and their dysfunction is less likely to be compensated for, which suggests that metabolites should be paid more attention.

2.2.3. Calculating the Combined Global Nonequivalence and Bias Scores of Metabolites. Our goal was to consider simultaneously the global nonequivalence of metabolites in a pathway and the bias in pathway identification based on metabolomic data. These two critical factors are not independent, given that metabolites with a high GN score are more likely to be identified in the profile and to be identified as differential. A simple sum of the values for these two factors is thus not adequate. We therefore used a monotonic cubic regression spline model [25] with six knots and a monotonicity constraint to integrate the GN scores and the initial bias scores. This monotonic model quantified the probability of metabolites being differential and compensated for as a function of GN score.

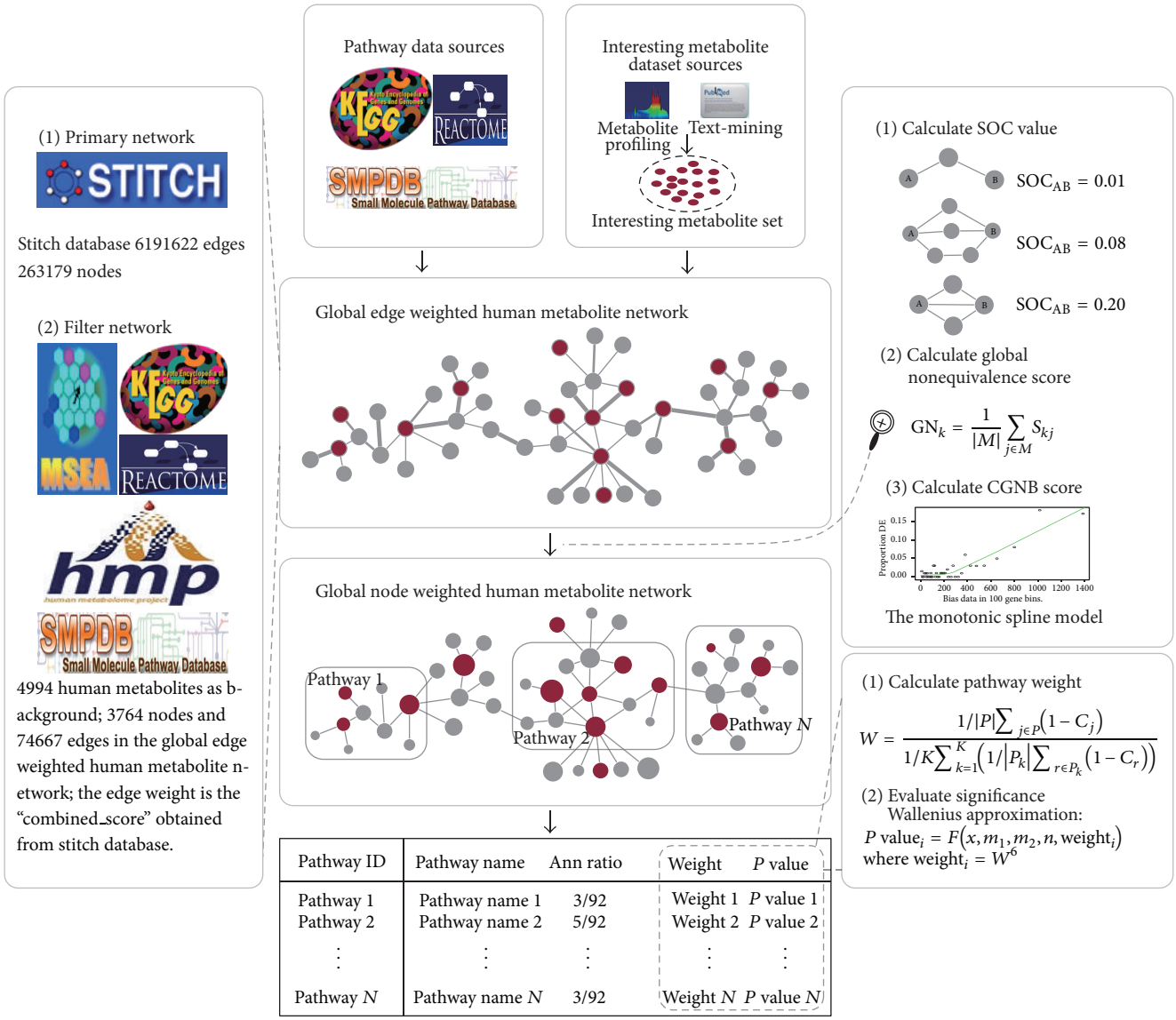


FIGURE 1: Schematic overview of MPINet.

The cubic regression spline with k knots can be represented as

$$y(x) = \beta_1 b_1(x) + \beta_2 b_2(x) + \cdots + \beta_k b_k(x), \quad (1)$$

where $b_i(x)$ is the basis cubic function at knot i, $i = 1 \cdots k$. For a given response y_j and covariates x_j,

$$y_j = y(x_j) = \beta_1 b_1(x_j) + \beta_2 b_2(x_j) + \cdots + \beta_k b_k(x_j), \quad (2)$$

where the y_j in this study is a binary value for the corresponding metabolite, 1 for differential metabolites and 0 for nondifferential metabolites in the network; x_j is the GN score for the corresponding metabolite. Suppose that the k knots are $x_1^*, x_2^*, \ldots, x_k^*$, which are placed at the quantiles of the distribution ofthe GN score vector; then the basis cubic function $b_i(x)$ can be represented as

$$b_i(x) = \varphi(x, x_i^*). \quad (3)$$

Supposing that there are M metabolites in the network, the binary initial bias score vector of these M metabolites Y can be represented as

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{bmatrix}. \quad (4)$$

The cubic spline model can then also be represented as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{bmatrix} = \begin{bmatrix} \varphi(x_1, x_1^*), \varphi(x_1, x_2^*), \ldots, \varphi(x_1, x_k^*) \\ \varphi(x_2, x_1^*), \varphi(x_2, x_2^*), \ldots, \varphi(x_2, x_k^*) \\ \vdots \\ \varphi(x_M, x_1^*), \varphi(x_M, x_2^*), \ldots, \varphi(x_M, x_k^*) \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_M \end{bmatrix}, \tag{5}$$

where y_j, x_j $(j = 1 \cdots M)$ are the initial bias score and the GN score of the corresponding metabolite, respectively. Consider $y_j = 1$ for an interesting metabolite; otherwise $y_j = 0$. The cubic spline model can be simplified as

$$Y = X\beta + \varepsilon, \tag{6}$$

where X is the model matrix, β is the parameter vector that contains $\beta_1, \beta_2, \ldots, \beta_k$. The penalized constrained least squares with monotonicity constraint are used to evaluate the parameter vector β, to minimize

$$\|Y - X\beta\|^2 + \lambda\beta^T S\beta. \tag{7}$$

Here S is a positive semidefinite matrix of coefficients. λ is the smoothing parameter which can be given by β. We then obtain the evaluated β vector, $\widehat{\beta}$. The probability of metabolites being differential and compensated can be evaluated as

$$\widehat{Y} = X\widehat{\beta}. \tag{8}$$

Finally, the combined global nonequivalence and bias (CGNB) score vector of a metabolite, C, can be calculated as

$$C = 1 - \widehat{Y}. \tag{9}$$

Thus, a high CGNB value represents a metabolite that is difficult to be identified in the profile and also not easily compensated for, which indicates that the metabolite should be paid more attention.

2.2.4. Evaluating the Significance of Pathways. We used the Wallenius approximation [15] to evaluate the significance of pathways. This method is an extended version of the hypergeometric test, which involves weighted parameters. In MPINet, we assumed that the probability of metabolites being identified in the profile and being compensated for within a pathway differed from that of metabolites within other pathways. For each pathway, a weight can be calculated based on the CGNB scores of the metabolites within it. For pathways containing metabolites that are difficult to compensate for under dysregulated conditions and are difficult to identify in the profile, MPINet will enhance their competitiveness through the pathway-weight value. For a given pathway, the following values are required for this step: (i) the number of interesting metabolites (n); (ii) the number of background metabolites (N); (iii) the number of background metabolites annotated to this pathway (m_1); (iv) the number of interesting metabolites annotated to this pathway (g); (v) the weight of this pathway (w_1). First, the relative weight of the pathway is calculated as follows:

$$W = \frac{(1/|P|)\sum_{j\in P}\left(1 - C_j\right)}{(1/K)\sum_{k=1}^{K}\left((1/|P_k|)\sum_{r\in P_k}(1 - C_r)\right)}, \tag{10}$$

where P is the metabolite set of the pathway, $|P|$ is the size of P, j is the metabolite in P, C_j is the CGNB score of the metabolite j, K is the number of pathways selected for analysis, P_k is the metabolite set in the kth pathway, $|P_k|$ is the size of P_k, r is the metabolite in P_k, and C_r is the CGNB score of the metabolite r. Finally, the significance P-value of this pathway can then be calculated as follows:

$$P \text{ value} = 1 - \sum_{x=0}^{g-1}\left(\binom{m_1}{x}\binom{m_2}{n-x}\right) \times \int_0^1 \left(1 - t^{w_1/d_x}\right)^x \left(1 - t^{w_2/d_x}\right)^{(n-x)} dt, \tag{11}$$

where $d_x = w_1^*(m_1 - x) + w_2^*(m_2 - (n - x))$, $w_1 = W^6$, $w_2 = 1$, and $m_2 = N - m_1$.

3. Results

We firstly confirmed that the bias in metabolomics experimental technologies impacts on pathway identification. Then, we applied MPINet to four datasets, including a prostate cancer metastasis dataset, diabetes dataset 1, diabetes dataset 2, and a drug-sensitivity dataset. In the analysis of the prostate cancer metastasis dataset, we aimed to compare the effectiveness of MPINet and other currently popular methods, including the hypergeometric test (ORA), MSEA (SEA), and MetPA. We then applied MPINet to a different biological phenotype (diabetes) to identify novel diabetes-related pathways. Diabetes dataset 2 was used to demonstrate the reproducibility of MPINet. Finally, we applied MPINet to a drug-sensitivity dataset to test its validity with a completely different biological problem.

3.1. Bias in Metabolomics Experimental Technologies Impacts on Pathway Identification. These identified metabolites which were provided by the metabolomic technology are small in amount and not random selected. These identified metabolites in the profiles are usually well-studied metabolites which tend to have many functional partners in the network. This will lead to the bias that pathways contain many well-studied metabolites are likely to be inappropriately identified.

In order to validate the bias in metabolite pathway identification, we performed the following analysis. First, we validated our assumption that metabolites with high GN scores were more likely to be identified in profiles and thus be considered as differential (i.e., the two main factors are

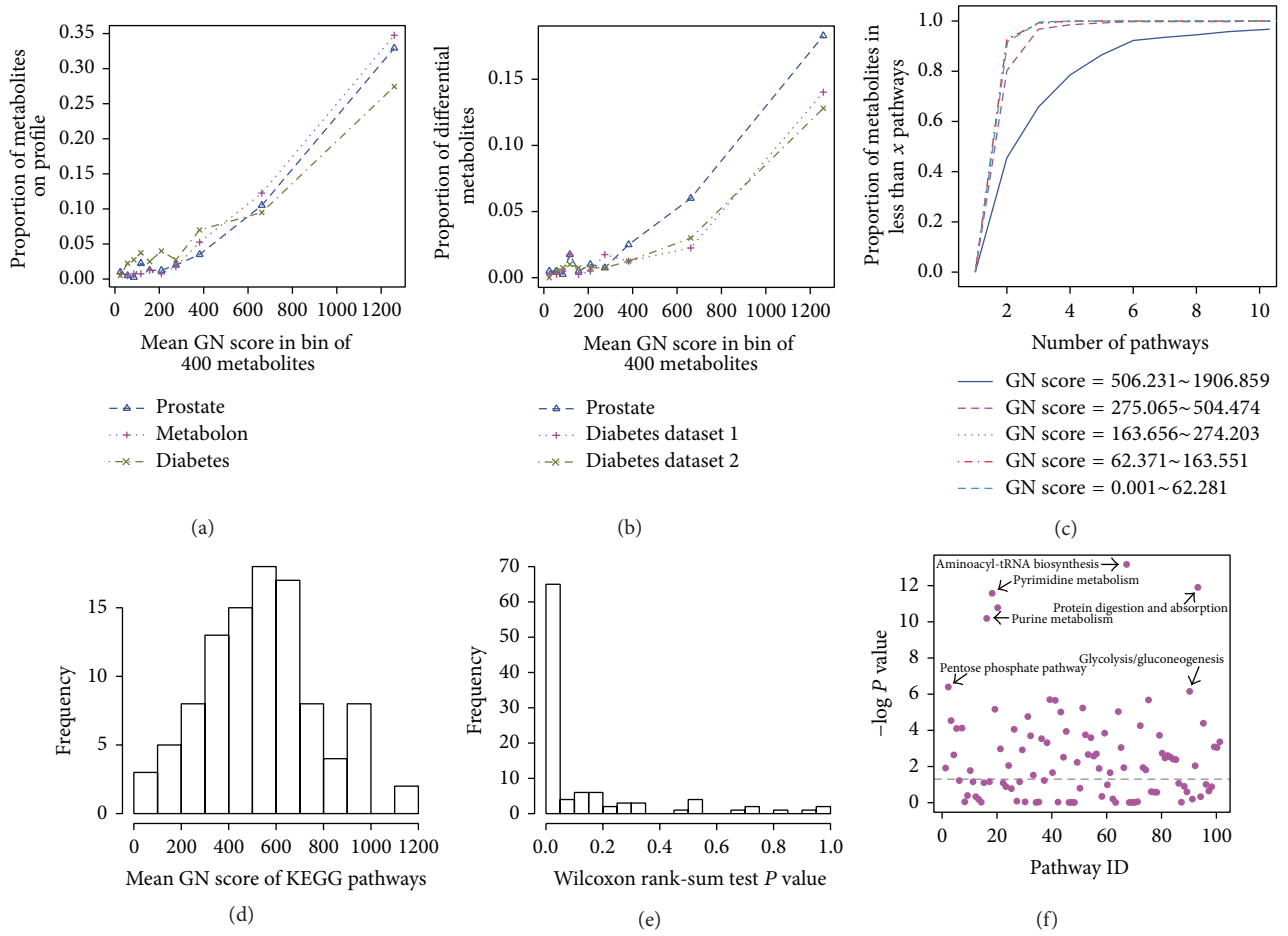


Figure 2: Validation of bias in metabolite pathway identification based on 101 pathways with more than five metabolites each. (a) The proportion of metabolites in the profile plotted against the mean GN score in a bin of 400 metabolites in the global human metabolite network across the three profiles. (b) The proportion of differential metabolites plotted across the three disease datasets. (c) Cumulative distribution of number of pathways associated with metabolites at a given GN score level. (d) Frequency of mean GN scores of metabolites in pathways. (e) P values for two-sided Wilcoxon's rank-sum test comparing the GN score of metabolites in the given pathway with that of the overall metabolites. (f) Scatter plot of pathway P value distributions. P values were calculated by one-sided Wilcoxon's rank-sum test comparing GN scores of metabolites in a given pathway with overall metabolites.

not independent). We calculated the GN scores of metabolites identified in three different metabolomic profiles and showed that metabolites identified in metabolomic profiles tended to have high GN scores (Figure 2(a)). Differential metabolites also tended to have high GN scores (Figure 2(b)). We then calculated the GN scores of metabolites in the network and inspected the number of pathways in which they participated. As expected, metabolites with high GN scores tended to participate in multiple pathways, while metabolites with low GN scores tended to reside in few pathways (i.e., pathway-specific metabolites) (Figure 2(c)).

For the pathways, we calculated the GN scores of all the metabolites in KEGG pathways and found a wide variation in the distribution of metabolite GN scores (Figure 2(d)). We then used Wilcoxon's rank-sum test to determine if the GN scores of metabolites within each pathway differed significantly from random. More than half of all pathways had significantly differential GN-score metabolites ($P < 0.05$, Figure 2(e)). We similarly analyzed all 3189 human pathways from ConsensusPathDB [21] and found consistent results (Figure S1). We also found that pathways including metabolites with high GN scores were usually fundamental pathways (Figure 2(f)). Overall, almost half of the analyzed pathways tended to contain metabolites with significantly high or low GN scores compared with random; however, the metabolites identified in the profiles and differential metabolites tended to have high GN scores (Figures 2(a) and 2(b)), suggesting that pathways containing many high-GN-score metabolites (i.e., usually fundamental pathways) were preferentially identified, while pathways containing many low-GN-score metabolites tended to be ignored by most currently used methods. A method that takes account of this bias is thus required.

3.2. Network-Based Metabolite Pathway Identification (MPINet) Can Effectively Identify Pathways Implicated in Prostate Cancer Metastasis. We determined the effectiveness

of MPINet for identifying pathways associated with prostate cancer metastasis and compared the results with those of three other popular methods, including ORA (hypergeometric test), MSEA, and MetPA.

3.2.1. Comparison of MPINet with ORA. To demonstrate the advance of MPINet, we compared pathways identified by MPINet to those by ORA using a prostate cancer dataset. MPINet identified twenty-two significant pathways, associated with prostate cancer metastasis, with a strict false discovery rate (FDR) cut-off level (FDR < 0.01). Among these pathways, up to 14 were well documented to be implicated in cancer (detailed information is provided in supplementary Table S3). After applying the ORA method to the same prostate cancer dataset, we identified 12 significant pathways under the same strict cut-off level (FDR < 0.01). Compared with ORA, MPINet identified 18 unique pathways, 11 of which were reported to associate with metastatic cancer or cancer (pathway names marked red in Figure 3(a); Table S3). These results indicate the superior ability of MPINet to identify cancer-related pathways.

The most significant additional pathway identified by MPINet was the "tryptophan metabolism" pathway. MPINet yielded a FDR of $2.18E-07$. However, this pathway was not significant even at the 10% level in the ORA method (FDR = 0.29). The degradation of tryptophan mediated by indoleamine 2,3-dioxygenase can influence the tumoral immune response [26]. Recently, Opitz et al.'s study also found that kynurenine, which is derived from tryptophan through tryptophan-2,3-dioxygenase, can suppress antitumor immune responses and promote tumor-cell survival and motility [27]. Furthermore, it has been reviewed that the subregion of tryptophan metabolism pathway which starts from tryptophan was reported to be implicated with the cell proliferation of prostate cancer [28]. An inspection of this pathway showed that there were only three differential metabolites annotated in this pathway and two of which, tryptophan and kynurenate, were presented in the network (Figure 4(a) path: 00380). Kynurenate is a catabolite generated by kynurenine which was reported to promote the survival and motility process of tumor [27]. Recent studies found that the subregion of this pathway that converted tryptophan to kynurenate which includes tryptophan, N′-formylkynurenine, kynurenine, 4-(2-aminophenyl)-2,4-dioxobutanoate, and kynurenate is closely related with the tumor progression such as survival and motility (Figure S2) [27]. Interestingly, most of metabolites in this pathway had high CGNB scores, suggesting that they were pathway-specific and thus not easily detected in the metabolomic profile. These results suggest that the nondifferential metabolites in the subpath region from tryptophan to kynurenate, including N′-formylkynurenine, kynurenine, and 4-(2-aminophenyl)-2,4-dioxobutanoate (Figure 4(a) path: 00380), may also be associated with cancer. MPINet increased the competitiveness of the "tryptophan metabolism" pathway by integrative analysis of the metabolite network structure and metabolomic profile. However, ORA ignored this pathway because a few metabolites were annotated to this pathway. ORA analysis also ignored the nonequivalence of metabolites and the bias inherent in metabolite pathway analysis.

MPINet identified the "arachidonic acid metabolism" pathway with FDR = 0.0023. Most metabolites in this pathway had high CGNB scores (Figure 4(a): path: 00590), but surprisingly, only one interesting metabolite is annotated in this pathway. ORA analysis thus disregarded this pathway with a high FDR value 0.76. However, the arachidonic acid metabolism pathway has been reported to highly associate with the progression of prostate tumor [29]. Previous studies have shown that arachidonic acid can mediate the progression of prostate cancer metastasis to bone and affect the metabolism of cancer in bone stromal cells [30]. In addition, a wide range of studies have shown that arachidonic acid can affect the progression of malignant prostate cancer through metabolites produced in the COX and LOX processes [31–33]. These metabolites further influence many cancer invasion-related activities, including proliferation, apoptosis, and angiogenesis [31–33]. For example, Yang et al. [33] demonstrated that the LOX metabolite 12-HETE was important for the progression of prostate carcinoma.

3.2.2. Comparison of MPINet with Other Methods. We also compared MPINet with MSEA and MetPA. The MSEA method identified 13 significant pathways in the prostate cancer dataset, under the strict cut-off value for significance (FDR < 0.01). Because MSEA uses the SMPDB pathways in the MSEA library as pathway databases, rather than KEGG, we also applied MPINet to the same pathway databases to ensure a fair comparison. MPINet identified 15 significant pathways (FDR < 0.01) when using SMPDB pathway databases. Most of the pathways (10 pathways) identified by MSEA were also included in the pathway list for MPINet (Figure 3(b)), and three pathways (marked red in Figure 3(b)) including "tryptophan metabolism," "tyrosine metabolism," and "steroidogenesis" pathway have been reported to be highly associated with cancer and were uniquely identified by MPINet. The tryptophan metabolism has been well documented in the literatures to be implicated with the progression of prostate cancer such as promoting the survival and motility of tumor cell and suppressing the antitumor immune response [27]. MSEA uses the whole list of metabolites in a profile and metabolite concentration changes. However, this quantitative information is unavailable for most metabolites. The "tryptophan metabolism" pathway, as discussed above that highly associates with prostate cancer, was also uniquely identified by MPINet in this comparison (Figure 3(b)). Inspection of this pathway shows that only two differential metabolites were identified in the profile, while no quantitative information was available for the other metabolites in the pathway (Figure 4(a): path 00380). In addition, MSEA treats each pathway as a simple metabolite set and ignores functional interactions among metabolites within the biological network. Two further pathways, "tyrosine metabolism" and "steroidogenesis," are also reported to be associated with cancer [34, 35]. For example, the norepinephrine is closely related with tumor and the norepinephrine metabolism which has been reported to highly

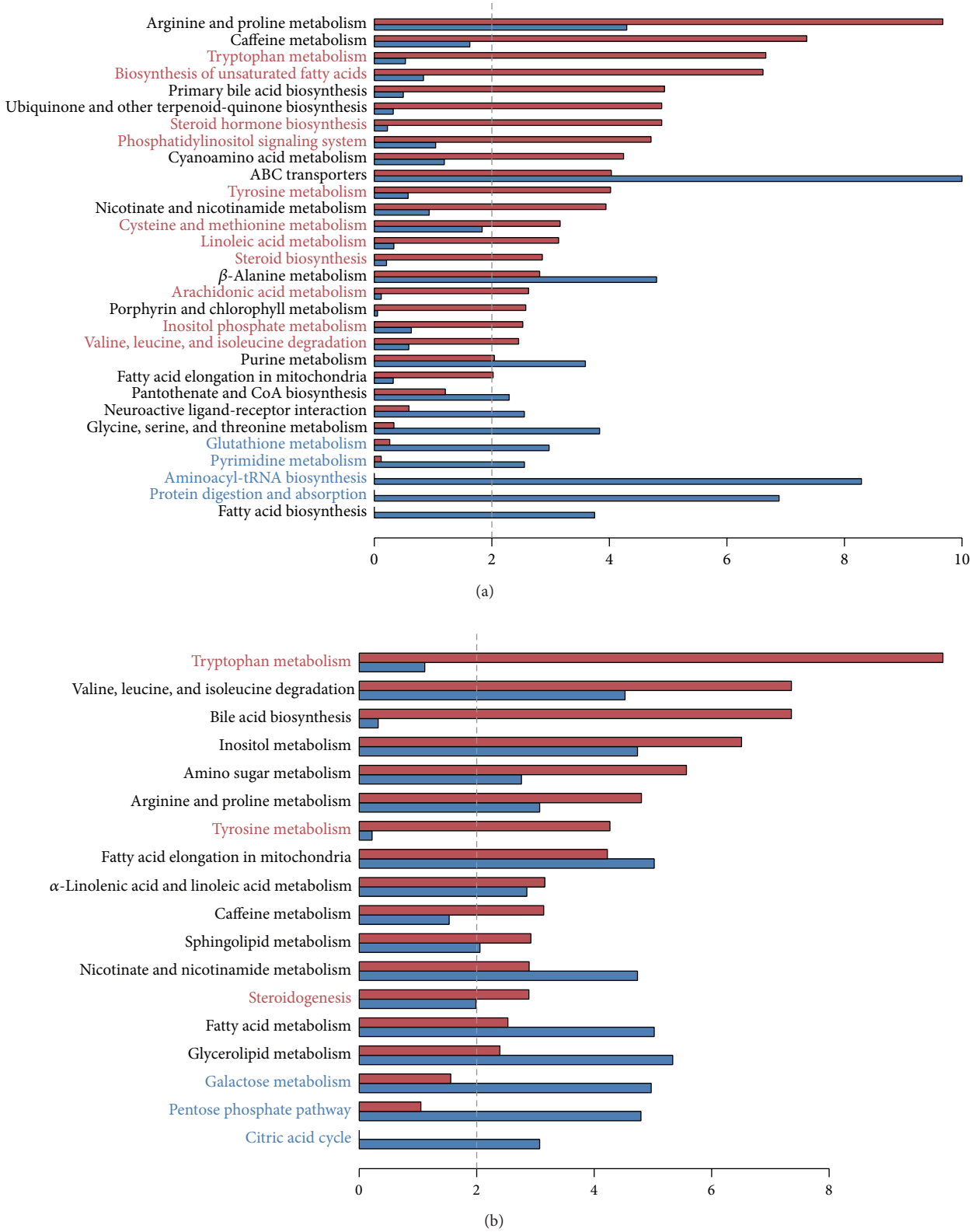


Figure 3: Comparisons between MPINet and other methods. Y-axis represents pathways, and x-axis is the –log10 transformation of FDR-values. Red bars represent pathway results identified by MPINet and blue bars represent the results of ORA (MSEA). Pathway names marked in red were uniquely identified by MPINet. Pathway names marked by blue were uniquely identified by ORA (MSEA). (a) MPINet versus ORA. (b) MPINet versus MSEA.

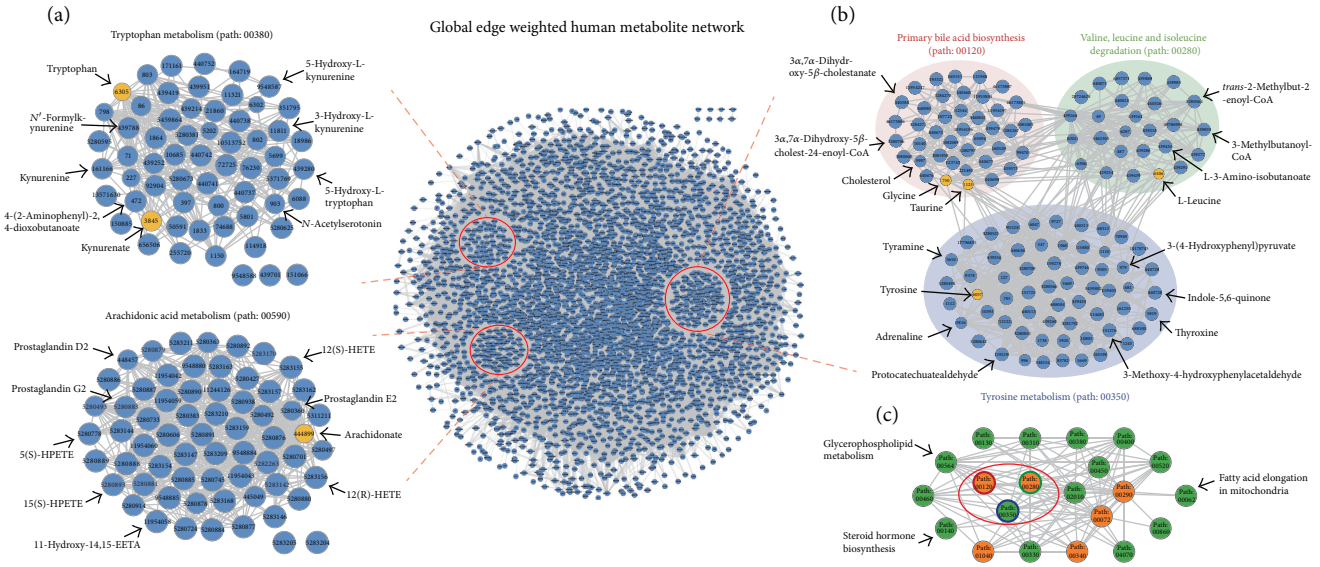


FIGURE 4: Global-weighted human metabolite network for several pathways identified by MPINet. The yellow node represents differential metabolites. Node size is proportional to the CGNB score of metabolites in (a) and (b). (a) Two metastatic prostate cancer-related pathways: "tryptophan metabolism" and "arachidonic acid metabolism." (b) Three type 2 diabetes-related pathways: "primary bile acid biosynthesis," "valine, leucine, and isoleucine degradation," and "tyrosine metabolism." (c) A global view of the interaction between the 21 type 2 diabetes-associated pathways. The edges between two pathways are displayed when the average GCS value between metabolite sets in the two pathways is greater than the median GCS value. Edge-line width is proportional to the average GCS value. Orange nodes represent pathways known to be related to type 2 diabetes. The red circle in the network corresponds to the three pathways in (b).

associate with the initiation and progression of tumor is a subprocess of the tyrosine metabolism pathway [34, 36, 37]. However, three pathways identified by MSEA, "galactose metabolism," "pentose phosphate pathway," and "citric acid cycle" (marked blue in Figure 3(b)), are more fundamental pathways.

We applied MetPA to the prostate cancer dataset. For fair comparison, 72 pathways in our KEGG pathway data, which were also in the 80 human pathway library of MetPA, were selected for comparison. The results of MetPA are given as impact scores, and we therefore compared pathway rank list from MPINet and MetPA. MPINet detected 15 pathways at a significance level of FDR < 0.01, based on this pathway dataset, up to nine of which are known to be related to cancer (Table 1). Arginine and proline metabolism is the most significant pathway identified by MPINet. Several studies have shown that arginine and proline metabolism regulated the immune responses and tumor growth and metastasis [38, 39]. We then examined their ranks in the MetPA pathway list. Among these nine pathways, the MPINet ranks of seven surpassed those in MetPA (marked by stars in Table 1). Six of these seven pathways were ranked >15 in MetPA. For example, the top ranked of the six pathways was the "tryptophan metabolism" pathway which is highly associated with the tumor progression, with a MetPA impact score of only 0.14 (rank 19). The impact score in MetPA is calculated as the normalized sum of importance measures of differential metabolites and also depends on the number of differential metabolites. However, as shown in Figure 4(a), most metabolites in this pathway were not identified in the profile. Furthermore, most of the metabolites in this pathway are pathway-specific (Figure 4(a)), suggesting that they are less likely to be compensated for by factors outside this pathway. MPINet paid more attention to this pathway by considering the global nonequivalence of metabolites and the bias associated with the experimental technology. MetPA, however, disregarded this pathway because of local importance measures and ignoring the bias.

3.3. Identifying Pathways Related to Type 2 Diabetes

3.3.1. MPINet's Potential to Identify Novel Type 2 Diabetes-Related Pathways. We applied MPINet to a different phenotype and demonstrated its power to identify novel pathways potentially related to type 2 diabetes. We initially analyzed type 2 diabetes dataset 1. MPINet identified 21 pathways with FDR < 0.01 (Table S4) and up to six of which belonged to lipid metabolism, which has been reported to play an important role in the pathogenesis of type 2 diabetes, mainly through its influence on insulin resistance [40, 41]. Furthermore, lipid management represents a useful strategy for reducing vascular risk in patients with diabetes mellitus [42]. Among these six identified lipid-metabolism pathways, "primary bile acid biosynthesis," "biosynthesis of unsaturated fatty acids," and "synthesis and degradation of ketone bodies" have previously been reported to be associated with type 2 diabetes [43–46]. The most significant additional pathway related to type 2 diabetes was the "primary bile acid biosynthesis"

Table 1: Top-ranked 15 pathways in MPINet and their ranks in MetPA.

Pathway name	FDR-N	R-P	I-P
Arginine and proline metabolism*	**8.15*E* − 09**	**6**	**0.21**
Caffeine metabolism	1.16*E* − 07	12	0.17
Tryptophan metabolism*	**5.75*E* − 07**	**17**	**0.14**
Primary bile acid biosynthesis	2.54*E* − 05	17	0.14
Ubiquinone and other terpenoid-quinone biosynthesis	2.65*E* − 05	33	0.05
Steroid hormone biosynthesis*	**2.65*E* − 05**	**38**	**0.03**
Cyanoamino acid metabolism	0.00019	12	0.17
Tyrosine metabolism*	**0.00034**	**34**	**0.04**
Nicotinate and nicotinamide metabolism	0.0004	26	0.07
Linoleic acid metabolism	**0.0012**	**1**	**0.62**
Cysteine and methionine metabolism	**0.0039**	**9**	**0.19**
Arachidonic acid metabolism*	**0.0039**	**—**	**—**
Porphyrin and chlorophyll metabolism	0.0043	34	0.04
Inositol phosphate metabolism*	**0.007**	**17**	**0.14**
Valine, leucine, and isoleucine degradation*	**0.0081**	**26**	**0.07**

FDR-N: FDR values of pathways in MPINet; R-P and I-P: ranks and impact scores for MetPA. Bold pathways have been well reported to be related with cancer. Ranks of pathways marked by asterisk in MPINet surpass that in MetPA.

pathway, with an FDR value of $1.45E-05$ (rank 3), compared with an FDR value of 0.216 (rank 34) yielded by ORA, and an impact value of only 0.07 (ranked 23) reported by MetPA. Most metabolites in this pathway tended to be pathway-specific and thus difficult to be identified (Figure 4(b)). Bile acid sequestration maintains lipid concentrations by mediating bile acid metabolism, and bile acid has the ability to influence systemic glucose metabolism [43]. Furthermore, Kobayashi et al. [44] also found that bile acids impacted on glucose metabolism and proposed the bile acid metabolism pathway as a novel potential therapeutic target pathway in type 2 diabetes. Therefore, bile acid sequestrants such as cholestyramine have been developed for the treatment of type 2 diabetes [47]. Biosynthesis of unsaturated fatty acids pathway is also highly associated with type 2 diabetes. The animal model experimental study of Krishna Mohan and Das suggested that polyunsaturated fatty acids have the ability of suppressing the occurrence of diabetes mellitus induced by chemical [48]. Moreover, the similar conclusion was review in that elevating the consumption of n-3 long chain polyunsaturated fatty acids may prevent type 2 diabetes and increasing evidences support this conclusion [45]. Three other possible novel type 2 diabetes-related lipid-metabolism pathways were identified by MPINet (FDR < 0.01), including "steroid hormone biosynthesis," "fatty acid elongation in mitochondria," and "glycerophospholipid metabolism."

MPINet also identified some additional pathways, including "valine, leucine, and isoleucine degradation," "valine, leucine, and isoleucine biosynthesis," and "histidine metabolism," which are related to type 2 diabetes [49–51]. The most significant of these was the "valine, leucine, and isoleucine degradation" pathway, which was identified by MPINet with an FDR value of $3.33E-05$ (rank 4), compared with an FDR value of 0.042 (rank 14) in the ORA method, and an impact score of only 0.09 in MetPA (rank 19). Inspection of this pathway showed that it included many metabolites with high CGNB scores (Figure 4(b)). Leucine-mediated mTORC1 signaling activation and mTORC1 together with its downstream target are important regulators of insulin resistance, which is a key feature of type 2 diabetes [50, 51]. The "histidine metabolism" pathway may be implicated with type 2 diabetes because histidine has the ability to suppress the production of hepatic glucose, and Kimura et al. suggest that histidine or the suppression of hepatic glucose production mediated by histidine is a potential therapeutic target of type 2 diabetes [49].

Finally, we investigated the crosstalk between the 21 pathways identified by MPINet with a significance level of FDR < 0.01. Interestingly, the additional "tyrosine metabolism" (path: 00350) pathway was closely connected with "primary bile acid biosynthesis" (path: 00120) and "valine, leucine, and isoleucine degradation" (path: 00280) in the network (Figure 4(b)). These results indicate that this pathway may also be associated with type 2 diabetes. Furthermore, several pathways not associated with type 2 diabetes in the literature were found to interact closely with well-reported type 2 diabetes-related pathways (Figure 4(c)).

3.3.2. Reproducibility of MPINet. The type 2 diabetes dataset 2 was used to evaluate the reproducibility of MPINet. We selected 66 differential metabolites from the data of Suhre et al. (see Section 2). Only 30.7% of metabolites in diabetes dataset 1 overlapped with those in diabetes dataset 2. However, up to 20 (76.9%) pathways identified in diabetes dataset 2 overlapped with diabetes dataset 1 (FDR < 0.05). The overlapped pathways were highly significant $P = 2.36e-11$,

hypergeometric test. Some critical pathways such as "valine, leucine, and isoleucine degradation," "valine, leucine, and isoleucine biosynthesis," "primary bile acid biosynthesis," and "histidine metabolism," which are implicated in the progression of type 2 diabetes, were also identified in diabetes dataset 2. These results suggest that MPINet performs reliably in metabolomics pathway identification.

3.4. Identifying Drug-Sensitivity-Associated Pathways. We investigated the effectiveness of MPINet for identifying drug-sensitivity-related pathways. Platinum-based drugs are widely used to treat many types of tumors, such as colon cancer, lung cancer, breast cancer, and ovarian cancer. However, resistance to platinum-based drugs is a major bottleneck in cancer therapy. The identification of consistent pathways relevant to platinum-based drug sensitivity is therefore of considerable importance. We analyzed the platinum-based drugs tetraplatin, iproplatin, and cisplatin and identified 55, 17, and 44 significant metabolites (FDR < 60%) associated with these three drugs, respectively. Through inputting these metabolites, MPINet identified 19, 9, and 19 significant pathways as drug-sensitivity-related pathways, respectively (FDR < 0.01). As we expected, most of the pathways identified by MPINet were shared by at least two of the three drugs (Figure 5(a)). These results indicate that MPINet was able to identify highly consistent pathways associated with the sensitivities of these three platinum-based drugs. MPINet identified five sensitivity-related pathways shared by all three drugs, including the three lipid metabolism pathways, "steroid biosynthesis," "primary bile acid biosynthesis," and "steroid hormone biosynthesis." Some previous studies have shown that lipid metabolism may influence the effects of platinum-based drugs [52], and Hendrich and Michalak [53] suggested that membrane-lipid composition might be associated with multidrug resistance. These results demonstrate the ability of MPINet to identify effectively pathways associated with platinum-based drug sensitivity. Furthermore, we expansively applied MPINet to all 121 drugs (see Section 2) and identified the sensitivity pathways for each drug. From a global point of view, drugs assigned to the same mode of action tended to be associated with sensitivity pathways from the same class (Figure 5(b)). For example, amino acid metabolic pathways tended to be related to topoisomerase I inhibitor sensitivity (Figure 5(c)). Overall, these results indicate that MPINet provides a robust and effective method for identifying drug-sensitivity pathways.

4. Conclusion and Discussion

The major innovations of MPINet include the simultaneous evaluation of the global nonequivalence of metabolites based on the global network structure and consideration of the bias associated with metabolomic experimental technology. From a biological perspective, dysfunctions of metabolites with strong functional-interaction partners are more likely to be compensated for than those with weak functional-interaction partners. Taking account of these functional interactions (network structure) to evaluate the nonequivalence of metabolites will help to improve the power of pathway identification. From the experimental technology perspective, the low coverage of metabolomics [9] and metabolite identification bias can have a great impact on subsequent pathway analysis (Figure 2). Based on these considerations, we developed a novel metabolite pathway identification method that takes account of the above aspects by allowing the integrated analysis of functional interactions between metabolites in the global biological network and the character of the metabolomic profile. We demonstrated the effectiveness and reproducibility of MPINet and its wide applicability across different application fields by analyzing several real datasets.

Metabolites involved in fundamental pathways have often been well studied and tend to be easily detected in the profile (Figure 2). These fundamental pathways may thus be preferentially identified if this bias is not adjusted for. In contrast, pathways that contain many pathway-specific metabolites (i.e., metabolites with low GN score) (specific pathways) will be more easily ignored. However, specific pathways have often been reported to be associated with diseases. If the metabolites in a pathway tend to be pathway-specific, dysregulation signals will be amplified and the entire pathway may even be shut down. An example of this is the arachidonic acid metabolism pathway, which includes many metabolites with high CGNB scores (Figure 4(a)). The metabolites are influenced by dysfunctional arachidonic acid and may not be easily compensated for by functional-interaction partners. MPINet pays more attention to pathway-specific metabolites by considering the nonequivalence of metabolites from a global perspective; MPINet increases the competitiveness of pathways containing many metabolites with high CGNB scores and decreases the competitiveness of pathways that contain many well-studied metabolites.

Compared with other current methods such as ORA, MSEA, and MetPA, MPINet can integrate not only metabolites from metabolomics experiments, but also the global metabolite network structure to enhance the ability of pathway identification. One of the limitations of current metabolomics technology is that many metabolites cannot be identified. MPINet utilizes the global metabolite functional network to compensate for the lack of information. Thus, in the case of pathways with weak signals from the annotation number point of view, MPINet can also identify these by considering the additional global network structure. MSEA can also detect some weak pathways [54, 55] by considering the additional nondifferential metabolites in the metabolomic profile; however, compared with MSEA, the additional information in MPINet is based on global functional level rather than on the concentration of a single metabolite. Moreover, MPINet only requires simple input consisting of a list of interesting metabolites. Taken together, MPINet has the potential to complement ORA, MSEA, and MetPA and may also be a useful tool in subsequent studies, such as studies of disease classification based on biologically meaningful pathways [56].

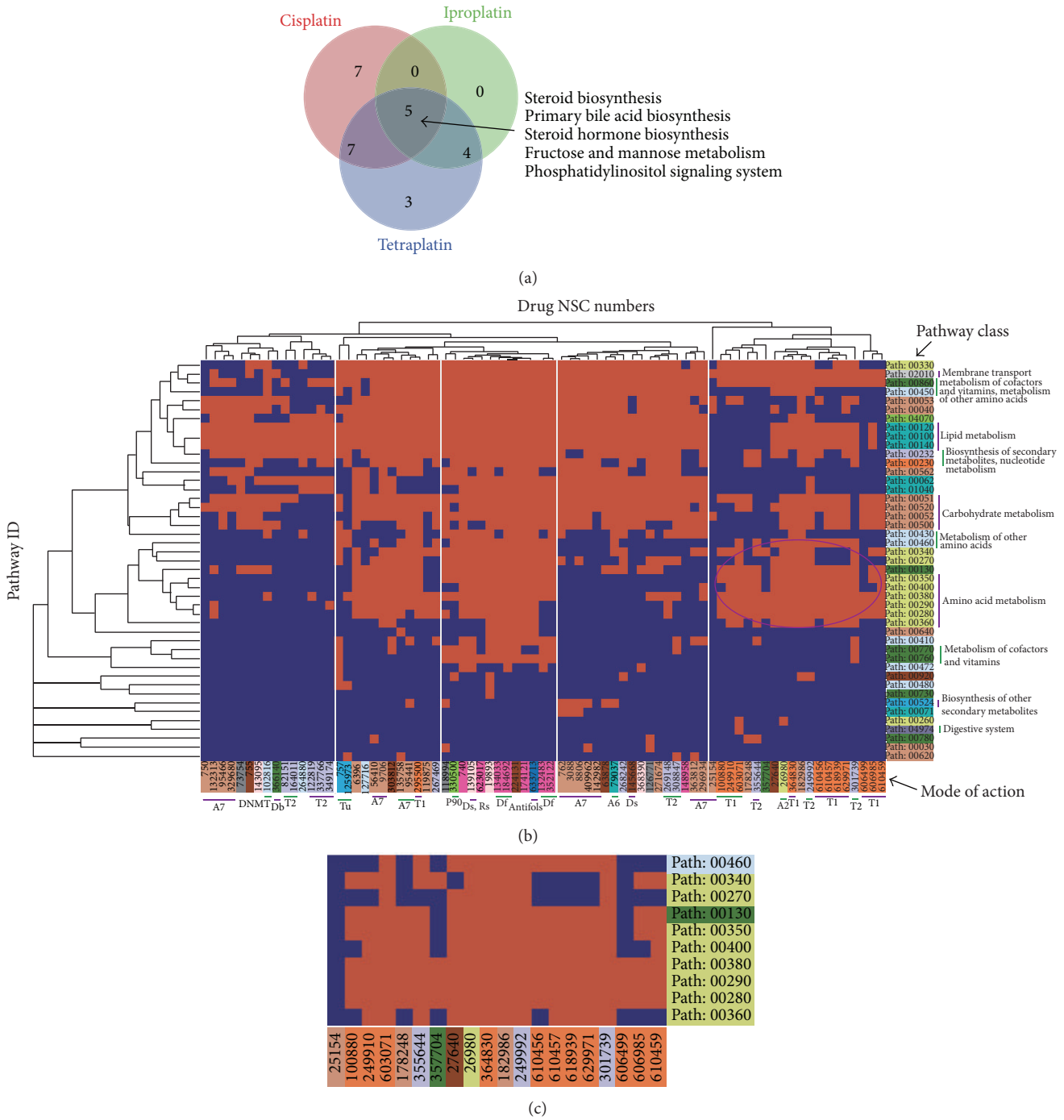


FIGURE 5: Identification of drug-sensitivity-related pathways. (a) Venn plot of pathways related to platinum-based-drug sensitivity. (b) Hierarchical clustering of drugs and sensitivity-related pathways. The corresponding cell was colored orange if the pathway was significantly associated with drug sensitivity (FDR < 0.01). (c) Zoom-in plot of the circle region in (b).

Conflict of Interests

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

Authors' Contribution

Feng Li and Yanjun Xu contributed equally to this work.

Acknowledgments

This work was supported in part by the Funds for Creative Research Groups of the National Natural Science Foundation of China (Grant No. 81121003), the National Natural Science Foundation of China (Grant Nos. 61170154, 61073136, and 31200996), the National Program on Key Basic Research Project (973 Program, Grant No. 2014CB910504),

and the Innovation Fund for Graduate Students of Heilongjiang province (Grant No. YJSCX2012-227HLJ).

References

[1] K. Hollywood, D. R. Brison, and R. Goodacre, "Metabolomics: current technologies and future trends," *Proteomics*, vol. 6, no. 17, pp. 4716–4723, 2006.

[2] M. Kanehisa, S. Goto, M. Hattori et al., "From genomics to chemical genomics: new developments in KEGG," *Nucleic Acids Research*, vol. 34, pp. D354–D357, 2006.

[3] G. Joshi-Tope, M. Gillespie, I. Vastrik et al., "Reactome: a knowledgebase of biological pathways," *Nucleic Acids Research*, vol. 33, pp. D428–D432, 2005.

[4] C. Li, X. Li, Y. Miao et al., "SubpathwayMiner: a software package for flexible identification of pathways," *Nucleic Acids Research*, vol. 37, no. 19, article e131, 2009.

[5] Z. Fang, W. Tian, and H. Ji, "A network-based gene-weighting approach for pathway analysis," *Cell Research*, vol. 22, pp. 565–580, 2012.

[6] M. Chagoyen and F. Pazos, "Tools for the functional interpretation of metabolomic experiments," *Briefings in Bioinformatics*, vol. 14, no. 6, pp. 737–744., 2013.

[7] M. Chagoyen and F. Pazos, "MBRole: Enrichment analysis of metabolomic data," *Bioinformatics*, vol. 27, no. 5, Article ID btr001, pp. 730–731, 2011.

[8] M. Kankainen, P. Gopalacharyulu, L. Holm, and M. Orešič, "MPEA-metabolite pathway enrichment analysis," *Bioinformatics*, vol. 27, no. 13, Article ID btr278, pp. 1878–1879, 2011.

[9] J. Xia and D. S. Wishart, "MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data," *Nucleic Acids Research*, vol. 38, no. 2, Article ID gkq329, pp. W71–W77, 2010.

[10] J. Xia and D. S. Wishart, "MetPA: a web-based metabolomics tool for pathway analysis and visualization," *Bioinformatics*, vol. 26, no. 18, Article ID btq418, pp. 2342–2344, 2010.

[11] L. Pham, L. Christadore, S. Schaus, and E. D. Kolaczyk, "Network-based prediction for sources of transcriptional dysregulation using latent pathway identification analysis," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 32, pp. 13347–13352, 2011.

[12] Y. Li, P. Agarwal, and D. Rajagopalan, "A global pathway crosstalk network," *Bioinformatics*, vol. 24, no. 12, pp. 1442–1447, 2008.

[13] M. Heinonen, H. Shen, N. Zamboni, and J. Rousu, "Metabolite identification and molecular fingerprint prediction through machine learning," *Bioinformatics*, vol. 28, no. 18, Article ID bts437, pp. 2333–2341, 2012.

[14] E. Werner, J.-F. Heilier, C. Ducruix, E. Ezan, C. Junot, and J.-C. Tabet, "Mass spectrometry for the identification of the discriminating signals from metabolomics: Current status and future trends," *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences*, vol. 871, no. 2, pp. 143–163, 2008.

[15] W. KT, *Biased sampling: the non-central hypegeometric probability distribution [Ph.D. thesis]*, Stanford University, 1963.

[16] A. Sreekumar, L. M. Poisson, T. M. Rajendiran et al., "Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression," *Nature*, vol. 457, pp. 910–914, 2009.

[17] D. S. Wishart, D. Tzur, C. Knox et al., "HMDB: the human metabolome database," *Nucleic Acids Research*, vol. 35, no. 1, pp. D521–D526, 2007.

[18] K. Suhre, C. Meisinger, A. Döring et al., "Metabolic footprint of diabetes: a multiplatform metabolomics study in an epidemiological setting," *PLoS ONE*, vol. 5, no. 11, Article ID e13953, 2010.

[19] J. N. Weinstein, K. W. Kohn, M. R. Grever et al., "Neural computing in cancer drug development: predicting mechanism of action," *Science*, vol. 258, no. 5081, pp. 447–451, 1992.

[20] W. C. Reinhold, M. Sunshine, H. Liu et al., "CellMiner: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set," *Cancer Research*, vol. 72, no. 14, pp. 3499–3511, 2012.

[21] A. Kamburov, U. Stelzl, H. Lehrach, and R. Herwig, "The ConsensusPathDB interaction database: 2013 Update," *Nucleic Acids Research*, vol. 41, no. 1, pp. D793–D800, 2013.

[22] M. Kuhn, D. Szklarczyk, A. Franceschini, C. von Mering, L. J. Jensen, and P. Bork, "STITCH 3: zooming in on protein-chemical interactions," *Nucleic Acids Research*, vol. 40, no. 1, pp. D876–D880, 2012.

[23] A. Frolkis, C. Knox, E. Lim et al., "SMPDB: The small molecule pathway database," *Nucleic Acids Research*, vol. 38, no. 1, Article ID gkp1002, pp. D480–D487, 2009.

[24] C. Campbell, J. Thakar, and R. Albert, "Network analysis reveals cross-links of the immune pathways activated by bacteria and allergen," *Physical Review E*, vol. 84, Article ID 031929, 2011.

[25] S. N. Wood, *Generalized Additive Models: An Introduction With R*, Chapman & Hall/CRC, New York, NY, USA, 2006.

[26] C. Uyttenhove, L. Pilotte, I. Théate et al., "Evidence for a tumoral immune resistance mechanism based on tryptophan degradation by indoleamine 2,3-dioxygenase," *Nature Medicine*, vol. 9, no. 10, pp. 1269–1274, 2003.

[27] C. A. Opitz, U. M. Litzenburger, F. Sahm et al., "An endogenous tumour-promoting ligand of the human aryl hydrocarbon receptor," *Nature*, vol. 478, no. 7368, pp. 197–203, 2011.

[28] E. J. Siddiqui, C. S. Thompson, D. P. Mikhailidis, and F. H. Mumtaz, "The role of serotonin in tumour growth (Review)," *Oncology Reports*, vol. 14, no. 6, pp. 1593–1597, 2005.

[29] D. Nie, M. Che, D. Grignon, K. Tang, and K. V. Honn, "Role of eicosanoids in prostate cancer progression," *Cancer and Metastasis Reviews*, vol. 20, no. 3-4, pp. 195–206, 2001.

[30] A. Angelucci, S. Garofalo, S. Speca et al., "Arachidonic acid modulates the crosstalk between prostate carcinoma and bone stromal cells," *Endocrine-Related Cancer*, vol. 15, no. 1, pp. 91–100, 2008.

[31] M. Hughes-Fulford, C. F. Li, J. Boonyaratanakornkit, and S. Sayyah, "Arachidonic acid activates phosphatidylinositol 3-kinase signaling and induces gene expression in prostate cancer," *Cancer Research*, vol. 66, no. 3, pp. 1427–1433, 2006.

[32] J. Ghosh, "Rapid induction of apoptosis in prostate cancer cells by selenium: reversal by metabolites of arachidonate 5-lipoxygenase," *Biochemical and Biophysical Research Communications*, vol. 315, no. 3, pp. 624–635, 2004.

[33] P. Yang, C. A. Cartwright, J. Li et al., "Arachidonic acid metabolism in human prostate cancer," *International Journal of Oncology*, vol. 41, no. 4, pp. 1495–1503, 2012.

[34] G. Eisenhofer, I. J. Kopin, and D. S. Goldstein, "Catecholamine metabolism: a contemporary view with implications for physiology and medicine," *Pharmacological Reviews*, vol. 56, no. 3, pp. 331–349, 2004.

[35] E. Thysell, I. Surowiec, E. Hörnberg et al., "Metabolomic characterization of human prostate cancer bone metastases reveals increased levels of cholesterol," *PLoS ONE*, vol. 5, no. 12, Article ID e14175, 2010.

[36] J. S. Fowler, J. Logan, G.-J. Wang et al., "Comparison of monoamine oxidase A in peripheral organs in nonsmokers and smokers," *Journal of Nuclear Medicine*, vol. 46, no. 9, pp. 1414–1420, 2005.

[37] G. Eisenhofer, T. T. Huynh, M. Hiroi, and K. Pacak, "Understanding Catecholamine Metabolism as a Guide to the Biochemical Diagnosis of Pheochromocytoma," *Reviews in Endocrine and Metabolic Disorders*, vol. 2, no. 3, pp. 297–311, 2001.

[38] D. S. Lind, "Arginine and cancer," *Journal of Nutrition*, vol. 134,, pp. 2837S–2853S, 2004.

[39] V. Bronte and P. Zanovello, "Regulation of immune responses by L-arginine metabolism," *Nature Reviews Immunology*, vol. 5, no. 8, pp. 641–654, 2005.

[40] I. Raz, R. Eldor, S. Cernea, and E. Shafrir, "Diabetes: insulin resistance and derangements in lipid metabolism. Cure through intervention in fat transport and storage," *Diabetes/Metabolism Research and Reviews*, vol. 21, no. 1, pp. 3–14, 2005.

[41] D. B. Savage, K. F. Petersen, and G. I. Shulman, "Disordered lipid metabolism and the pathogenesis of insulin resistance," *Physiological Reviews*, vol. 87, pp. 507–520, 2007.

[42] D. J. Betteridge, "Lipid control in patients with diabetes mellitus," *Nature Reviews Cardiology*, vol. 8, pp. 278–290, 2011.

[43] B. Staels and V. A. Fonseca, "Bile acids and metabolic regulation: mechanisms and clinical responses to bile acid sequestration," *Diabetes Care*, vol. 32, supplement 2, pp. S237–S245, 2009.

[44] M. Kobayashi, H. Ikegami, T. Fujisawa et al., "Prevention and treatment of obesity, insulin resistance, and diabetes by bile acid-binding resin," *Diabetes*, vol. 56, no. 1, pp. 239–247, 2007.

[45] J. A. Nettleton and R. Katz, "n-3 long-chain polyunsaturated fatty acids in type 2 diabetes: a review," *Journal of the American Dietetic Association*, vol. 105, no. 3, pp. 428–440, 2005.

[46] A. Avogaro, C. Crepaldi, M. Miola et al., "High blood ketone body concentration in Type 2 non-insulin dependent diabetic patients," *Journal of Endocrinological Investigation*, vol. 19, no. 2, pp. 99–105, 1996.

[47] Y. Handelsman, "Role of bile acid sequestrants in the treatment of type 2 diabetes," *Diabetes Care*, vol. 34, supplement 2, pp. S244–S250, 2011.

[48] I. Krishna Mohan and U. N. Das, "Prevention of chemically induced diabetes mellitus in experimental animals by polyunsaturated fatty acids," *Nutrition*, vol. 17, no. 2, pp. 126–151, 2001.

[49] K. Kimura, Y. Nakamura, Y. Inaba et al., "Histidine augments the suppression of hepatic glucose production by central insulin action," *Diabetes*, vol. 62, pp. 2266–2277, 2013.

[50] B. C. Melnik, "Leucine signaling in the pathogenesis of type 2 diabetes and obesity," *World Journal of Diabetes*, vol. 3, pp. 38–53, 2012.

[51] M. Krebs and M. Roden, "Nutrient-induced insulin resistance in human skeletal muscle," *Current Medicinal Chemistry*, vol. 11, no. 7, pp. 901–908, 2004.

[52] A. Kozubik, A. Vaculova, K. Soucek, J. Vondracek, J. Turanek, and J. Hofmanova, "Novel anticancer platinum(IV) complexes with adamantylamine: their efficiency and innovative chemotherapy strategies modifying lipid metabolism," *Metal-Based Drugs*, vol. 2008, Article ID 417897, 15 pages, 2008.

[53] A. B. Hendrich and K. Michalak, "Lipids as a target for drugs modulating multidrug resistance of cancer cells," *Current Drug Targets*, vol. 4, no. 1, pp. 23–30, 2003.

[54] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists," *Nucleic Acids Research*, vol. 37, no. 1, pp. 1–13, 2009.

[55] C. Li, J. Han, Q. Yao et al., "Subpathway-GM: identification of metabolic subpathways via joint power of interesting genes and metabolites and their topologies within pathways," *Nucleic Acids Research*, vol. 41, no. 9, article e101, 2013.

[56] W. Liu, C. Li, Y. Xu et al., "Topologically inferring risk-active pathways toward precise cancer classification by directed random walk," *Bioinformatics*, vol. 29, no. 17, pp. 2169–2177, 2013.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 462135, 8 pages
http://dx.doi.org/10.1155/2014/462135

Research Article

miRSeq: A User-Friendly Standalone Toolkit for Sequencing Quality Evaluation and miRNA Profiling

Cheng-Tsung Pan,[1] Kuo-Wang Tsai,[2] Tzu-Min Hung,[3] Wei-Chen Lin,[4] Chao-Yu Pan,[3] Hong-Ren Yu,[5] and Sung-Chou Li[3,6]

[1] *Institute of Bioinformatics and Systems Biology, National Chiao Tung University, Hsinchu, Taiwan*
[2] *Department of Medical Education and Research, Kaohsiung Veterans General Hospital, Kaohsiung, Taiwan*
[3] *Genomics & Proteomics Core Laboratory, Department of Medical Research, Kaohsiung Chang Gung Memorial Hospital and Chang Gung University College of Medicine, Kaohsiung, Taiwan*
[4] *Department of Parasitology, National Cheng Kung University College of Medicine, Tainan, Taiwan*
[5] *Departments of Pediatrics, Chang Gung Memorial Hospital-Kaohsiung Medical Center, Graduate Institute of Clinical Medical Science, Chang Gung University College of Medicine, Kaohsiung, Taiwan*
[6] *Children's Hospital, 12th Floor, No. 123, Dapi Road, Niaosong District, Kaohsiung 83301, Taiwan*

Correspondence should be addressed to Sung-Chou Li; raymond.pinus@gmail.com

Received 7 April 2014; Revised 24 May 2014; Accepted 25 May 2014; Published 24 June 2014

Academic Editor: Li-Ching Wu

MicroRNAs (miRNAs) present diverse regulatory functions in a wide range of biological activities. Studies on miRNA functions generally depend on determining miRNA expression profiles between libraries by using a next-generation sequencing (NGS) platform. Currently, several online web services are developed to provide small RNA NGS data analysis. However, the submission of large amounts of NGS data, conversion of data format, and limited availability of species bring problems. In this study, we developed miRSeq to provide alternatives. To test the performance, we had small RNA NGS data from four species, including human, rat, fly, and nematode, analyzed with miRSeq. The alignments results indicate that miRSeq can precisely evaluate the sequencing quality of samples regarding percentage of self-ligation read, read length distribution, and read category. miRSeq is a user-friendly standalone toolkit featuring a graphical user interface (GUI). After a simple installation, users can easily operate miRSeq on a PC or laptop by using a mouse. Within minutes, miRSeq yields useful miRNA data, including miRNA expression profiles, 3′ end modification patterns, and isomiR forms. Moreover, miRSeq supports the analysis of up to 105 animal species, providing higher flexibility.

1. Introduction

MicroRNAs (miRNAs) are non-protein coding RNAs. The mature products of miRNA genes are approximately 22-nt RNA fragments, rather than polypeptides. Because of intramolecular base pairing, the full-length primary transcripts form a hairpin structure plus unpaired ends, which is processed by the Drosha complex, trimming out the unpaired ends and releasing the hairpin structure into cytoplasm. The hairpin structure is further processed by Dicer, trimming out the terminal loop and releasing the RNA duplex. Either one or both strands of the RNA duplex are incorporated into RISC, functioning as mature miRNAs. By complimentary base pairing with the 3′ UTR, miRNAs guide RISC to the target gene's mRNAs, downregulating the target gene through either mRNA degradation or translational repression [1].

In addition to the biogenesis mechanisms and novel miRNA identification, many studies have focused on the regulatory functions of miRNAs. Since miRNA was discovered and characterized in *C. elegans*, numerous studies have reported that miRNAs play regulatory functions in a wide range of biological activities. Typically, the miRNA regulatory roles in cancer pathogenesis are investigated. According to the previous studies, miRNAs may function as tumor repressors, repressing tumor growth, tumor cell migration, or tumor cell proliferation [2, 3]. MiRNA may also play the roles

of onco-miR, promoting tumor growth or tumor cell migration [4, 5]. MiRNAs are also involved in development regulation, including axon regeneration [6], sarcomere formation [7], and embryo development [8]. Moreover, miRNAs also serve as biomarkers of numerous diseases, such as Alzheimer's disease [9], liver pathology [10], heart failure [11], and graft-versus-host disease [12].

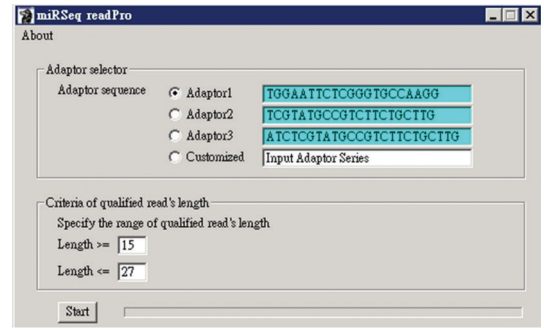
With more and more miRNAs identified in model organisms, miRNA-related studies focus on investigating miRNA regulation in diseases. Such studies depend on determining miRNA expression profiles between libraries, treatment versus control or normal versus disease. Therefore, next-generation sequencing (NGS) is usually applied to sequence miRNA, providing not only qualitative but also quantitative measurement of miRNA expression. However, the sequence data produced using NGS platforms typically require large disk space and are, therefore, difficult to be analyzed by the traditional biological researchers.

Currently, there are several online web services available for small RNA NGS data analysis [13, 14]. The user must first submit the raw sequence data in a fastq file to the online services and then request an analysis job. However, submitting large amounts of sequence data dramatically increases the Internet workload. Without a broadband network, transferring sequence data is time-consuming and becomes impeded. Although collapsed sequence data in the fasta format is also acceptable, converting fastq format into fasta format is generally a difficult task for biological researchers. Moreover, the requested jobs are typically held in a queue on the server rather than analyzed immediately. The server analyzes the requested jobs according to a first-come-first-served rule, making it difficult to estimate when the jobs are completed. Additionally, the online web services generally allow users to analyze small RNA NGS data of only a few species. So, an alternative is required.

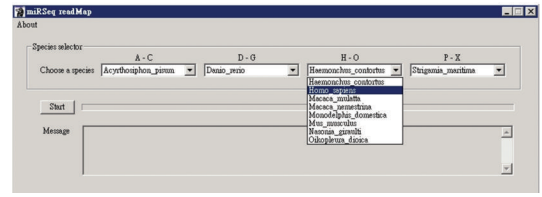
Therefore, we developed miRSeq as an alternative. miRSeq is compatible with Windows operating systems. Following step-by-step instructions, users can easily install and operate miRSeq. From the initial raw NGS data in fastq format, miRSeq can evaluate the sequencing quality regarding percentage of self-ligation read, read length distribution, and read category. Within minutes, miRSeq yields useful miRNA data, including miRNA expression profiles, 3′ end modification patterns, and isomiR forms. Moreover, miRSeq supports the analysis on up to 105 animal species, providing higher flexibility.

2. Materials and Methods

2.1. Data Resources. miRSeq classifies sequence reads into categories and determines miRNA expression profiles by mapping the sequence reads back to known annotated transcripts, downloaded as follows. The miRNA data belong to miRBase 20. The sequences of mRNAs and ncRNAs were derived from the RefSeq 60 [15]. The sequences of tRNAs were downloaded from the Genomic tRNA database [16]. The sequences of rRNAs were provided by the SILVA database [17]. The sequence reads not belonging to any of the previous classes were classified into the unknown class.



(a)



(b)

Figure 1: The operation interface of miRSeq. miRSeq is composed of readPro and readMap. (a) readPro deals with raw sequence reads in fastq format by collapsing raw reads into unique reads, tabulating read count, and trimming 3′ adaptor. (b) readMap is responsible for mapping the reads back to known annotations, classifying reads into different categories, determining miRNA expression profile, reporting 3′ end modification patterns, and analyzing isomiR forms.

miRSeq was developed with Perl and is compatible with Windows operating systems. The user can download miRSeq package via https://docs.google.com/forms/d/1WMsHS8jlxL-k3cL_UHTEKQGssno4ru6CuJMsGbPOqtk/viewform. Then, the user should first install the Perl complier and related modules by following the instructions in the readMe file. miRSeq is composed of two modules, including readPro and readMap, both of which can operate independently.

2.2. readPro. The operation interface of readPro is illustrated in Figure 1(a). readPro deals with raw sequence reads in fastq format. readPro first collapses the raw reads into unique sequence tags with the read count of each unique sequence tag tabulated. Then, readPro trims the 3′ adaptor by referring to the sequence of the specified 3′ adaptor (Figure 1(a)). The sequence tags with 3′ adaptor detected and trimmed are named "clean reads" and further analyzed. However, the sequence tags without 3′ adaptor detected are discarded. The clean reads are further collapsed into unique clean reads and the read counts of those are also retabulated. readPro also analyzes the length distribution of the clean reads, by which the user can evaluate the sequence quality of the NGS data. Next, readPro sifts the qualified clean reads according to the user-specified length criteria (Figure 1(a)), and the qualified clean reads are presented in fasta format. In addition, only the unique clean reads with read count $\geqq 2$ are included in the fasta file for further analysis. The results of each readPro alignment are available.

Table 1: The detailed information of analyzed libraries. SRA in the "Source" column denotes that the small RNA libraries were downloaded from NCBI SRA database. The SRA IDs of the libraries were provided. CGMH denotes that the libraries were prepared and sequenced by our research team in Kaohsiung Chang Gung Memorial Hospital.

Library	Organism	Source	Details
L1	*Homo sapiens*	CGMH	Human breast cancer cell line: MDA-MB-361
L2	*Homo sapiens*	CGMH	Human prostate cancer cell line: PC3
L3	*Rattus norvegicus*	CGMH	Normal lung tissue from 4-month rat
L4	*Rattus norvegicus*	CGMH	Normal lung tissue from 4-month rat
L5	*Caenorhabditis elegans*	SRA: SRR1175721	Synchronized adult population of nematodes
L6	*Caenorhabditis elegans*	SRA: SRR1139598	Stage-matched population of nematodes
L7	*Drosophila melanogaster*	SRA: SRR513989	Abdomens and thoraxes from w1118 male flies infected by Nora virus
L8	*Drosophila melanogaster*	SRA: SRR351332	Whole bodies of 2-3-day-old wild-type flies

2.3. readMap. The output result of readPro is the input data for readMap. The operation interface of readMap is illustrated in Figure 1(b). readMap maps the clean reads to pre-miRNAs with bowtie [18]. miRSeq collects the index data of miRNA and other annotated transcripts for 105 species (Supplementary Table 1 in Supplementary Material available online at http://dx.doi.org/10.1155/2014/462135). The miRNA expression profile, namely, the read count of each mature miRNA, can thereby be determined. The read count of miRNA is presented in transcript per million (TPM). In addition, the non-miRNA reads are further mapped in order to the rRNAs, tRNAs, mRNAs, ncRNAs, and genome (if applicable) of the specified species (Figure 1(b)). Thereby, clean reads are classified into categories. Using this information, users can evaluate whether the sample preparation protocol performed well.

When mapping reads back to pre-miRAs, variations preferentially occur at the 3′ terminal ends of the reads. Such variation could be caused by either RNA editing or nucleotide addition. It currently remains debatable which of the two accounts for the 3′ terminal variations. To bypass the debate, we named such variations "3′ end modifications." Using the output of readMap, the users can compare the 3′ end modification patterns between libraries. MiRNA NGS reads usually exist as isomiRs, namely, miRNA isoforms exhibiting a length difference or position shift compared with the reference mature miRNAs. Previous studies have showed that isomiRs may perform regulatory functions between different libraries [19, 20]. readMap reports all isomiR forms of mature miRNAs. The results of each readPro alignment are also available.

2.4. Small RNA Sample Preparation. We prepared small RNA samples with the standard or nonstandard protocols. The standard small RNA sample preparation protocol applies in-gel size fraction to enrich the RNA fragments with size of approximately 22 nucleotides, excluding most of the non-miRNA fragments. For the L1 library (Table 1), we applied the nonstandard sample preparation protocol. During the in-gel size fraction step, we avoided extracting the RNA fragments exhibiting a length of approximately 22 nucleotides. Instead, we extracted the RNA fragments with a length adjacent to 22 nucleotides, picking up the RNA fragments with more diverse range in length. By doing so, we expect more non-miRNA contaminants are included.

3. Results

3.1. Library Summary. To illustrate the output data by miRSeq and to test the performance of miRSeq, we had small RNA NGS data of eight libraries from four species analyzed with miRSeq. The detailed information of the eight libraries is listed in Table 1. Among the eight libraries, L1, L2, L3, and L4 ones (from human and rat) are prepared and sequenced by our laboratory. To evaluate whether miRSeq can examine sequencing quality and detect sequencing quality variations between libraries, the L1 library was prepared with the non-standard sample preparation protocol (Section 2) to apply variations between libraries. The remaining L2, L3, and L4 libraries were prepared with the standard sample preparation protocol. Then, the RNA samples were sequenced using the Illumina HiSeq platform. The L5, L6, L7, and L8 libraries (from nematode and fly) were downloaded from NCBI SRA database to examine miRSeq performance with public domain data [21]. When operating readPro (a tool of miRSeq package), the length constraint parameter was specified as 15 to 30.

3.2. Self-Ligation Reads. Small RNA samples are generally subject to 5′ adaptor and 3′ adaptor ligation at both ends before PCR amplification. Generally, the 5′ adaptor can ligate the 3′ adaptor without an RNA fragment inserted. Such self-ligation reads provide no useful information but garbage only, wasting sequencing consumables. As shown in Table 2, the self-ligation reads accounted for less than 1% in most libraries, except for L5 library (from *C. elegans*). Higher percentage of self-ligation read resulted in lower percentages of clean and qualified reads, leading to wasting budget. Thus, the percentage of self-ligation read is a crucial index for sequencing quality evaluation.

3.3. Length Distribution. The lengths of animal miRNAs are highly enriched at nucleotides 21, 22, and 23. Therefore, the lengths of clean reads from a well-prepared library should also be highly enriched at nucleotides 21, 22, and 23 without even scattering. In Figure 2(a), the length distribution of L2 library's clean reads was highly enriched at 22-nt as miRBase miRNAs. However, the lengths of clean reads in L1 library scatter with much less enrichment, which is consistent with the experimental design of this study. The less enrichment

Table 2: Alignment results of readPro. The small RNA NGS data of eight libraries were analyzed with miRSeq. Raw reads were classified into clean, nonclean, or self-ligation after 3′ adaptor trimming step. The clean reads following specified criteria are classified as qualified reads for further analysis. The sequences of adaptors 1, 2, and 3 are TGGAATTCTCGGGTGCCAAGG, TCGTATGCCGTCTTCTGCTTG, and ATCTCGTATGCCGTCTTCTGCTTG, respectively. The sequence of adaptor C (CTGTAGGCACCATCAATCGT) is based on the information in the corresponding SRA page.

Library	All	Self-ligation	Nonclean	Clean	Qualified	Adaptor version
L1	11,229,160	0.72%	4.65%	94.62%	86.15%	Adaptor 1
L2	11,501,087	0.02%	3.34%	96.65%	86.12%	Adaptor 1
L3	6,314,030	0.04%	3.94%	96.02%	85.37%	Adaptor 1
L4	6,235,528	0.03%	3.71%	96.26%	87.98%	Adaptor 1
L5	22,634,033	15.08%	7.59%	77.33%	68.37%	Adaptor 2
L6	9,023,339	0.23%	3.73%	96.04%	74.08%	Adaptor 1
L7	10,314,488	0.00%	13.61%	86.39%	83.08%	Adaptor C
L8	22,435,248	1%	7%	92%	83%	Adaptor 3

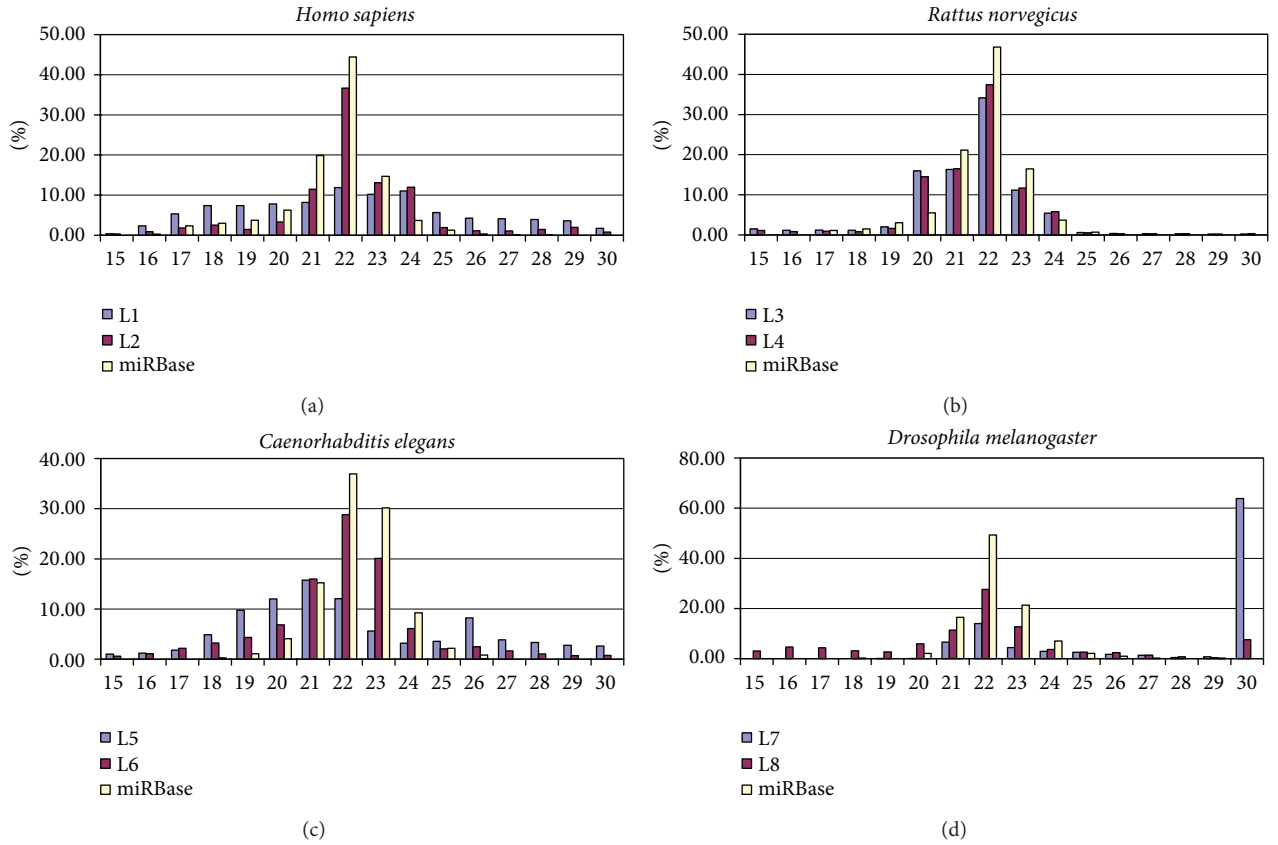
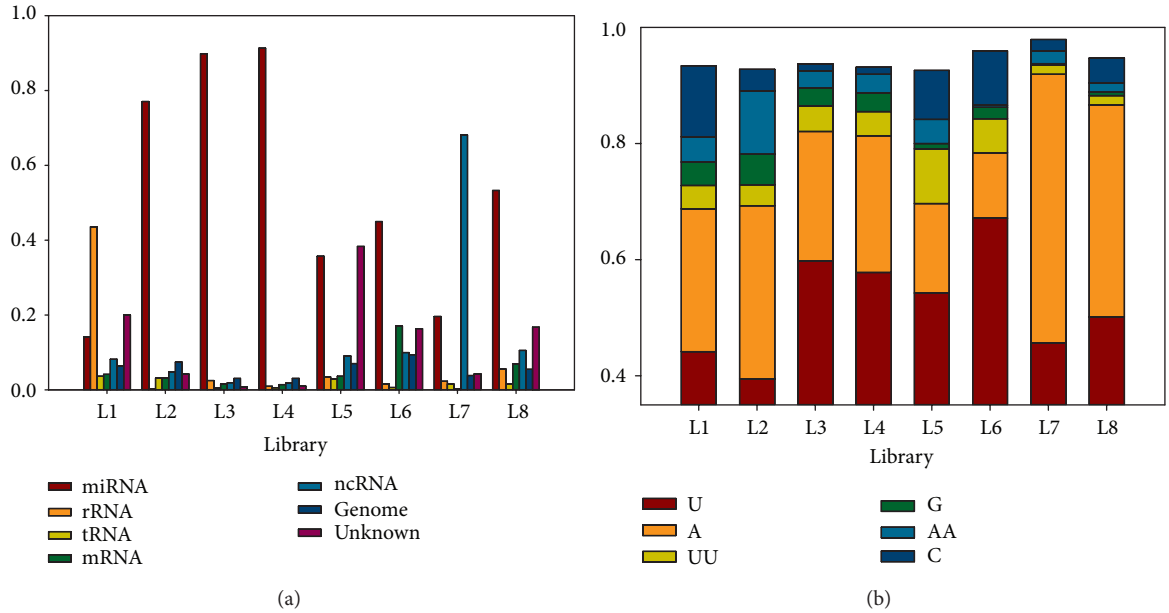


Figure 2: Length distribution comparisons of clean reads between libraries. From the output results of readPro, we may compare the length distribution of clean reads, examining if length enrichment occurs. (a) The length distribution pattern of the well-prepared L2 library was more similar to the one of miRBase miRNAs. (b) L3 and L4 libraries had similar distribution patterns. (c) The read length of L5 library scattered without enrichment. (d) The reads with length 30-nt dominated L7 library.

of L1 library implied that L1 library contained more non-miRNA contaminants. Both L3 and L4 libraries were prepared according to the standard protocol, resulting in the same high enrichment as miRBase miRNAs (Figure 2(b)).

In Figure 2(c), different degree of length enrichments was also observed between L5 and L6 libraries. The result of L5 and L6 was much similar to the one of L1 and L2 in Figure 2(a). In Figure 2(d), although length enrichment was observed in L8 library, the enrichment was much less than other good examples such as L2, L3, L4, and L6. In addition, most clean reads in L7 library belong to the length 30-nt, implying more than 60% non-miRNA contaminants. The results of length distribution imply the proportions of miRNAs and can be confirmed in the next section.

3.4. Read Categories. At the small RNA sample preparation step, the in-gel size fraction is typically applied to remove contaminants from other RNA molecules. Thus, a high percentage of non-miRNA reads reflects the poor performance




```
>hsa-mir-2110 5p:8–29 L2
caggggUUGGGGAAACGGCCGCUGAGUGaggcgucggcuguguuucucaccgcggucuuuuccucccacucuug
-------UUGGGGAAACGGCCGCUGAGU---------------------------------------------- 3   0, −1
-------UUGGGGAAACGGCCGCUGAGUGA--------------------------------------------- 5   0, 1
-------UUGGGGAAACGGCCGCUGAGUGAG-------------------------------------------- 21  0, 2
-------UUGGGGAAACGGCCGCUGAGUGAGa------------------------------------------- 9   0, 2
-------UUGGGGAAACGGCCGCUGAGUGAGu------------------------------------------- 9   0, 2
-----------------------------------------------UCACCGCGGUCUUUUCCUCCC------- 2
-----------------------------------------------UCACCGCGGUCUUUUCCUCCCu------ 19
-----------------------------------------------UCACCGCGGUCUUUUCCUCCCA------ 8
-----------------------------------------------UCACCGCGGUCUUUUCCUCCCAu----- 6
-----------------------------------------------UCACCGCGGUCUUUUCCUCCCAC----- 6
----------------------------------------------CACCGCGGUCUUUUCCUCCCu------ 2
----------------------------------------------CACCGCGGUCUUUUCCUCCCA------ 12
----------------------------------------------CACCGCGGUCUUUUCCUCCCAa----- 4
----------------------------------------------CACCGCGGUCUUUUCCUCCCAaa---- 5
----------------------------------------------CACCGCGGUCUUUUCCUCCCAu----- 7
----------------------------------------------CACCGCGGUCUUUUCCUCCCAC----- 86
----------------------------------------------CACCGCGGUCUUUUCCUCCCACa---- 12
----------------------------------------------CACCGCGGUCUUUUCCUCCCACg---- 2
----------------------------------------------CACCGCGGUCUUUUCCUCCCACU---- 24
----------------------------------------------ACCGCGGUCUUUUCCUCCCACU----
```

(c)

FIGURE 3: Illustration of readMap alignment output. From the output results of readMap, we may compare read category, 3′ end modification, and isomiR patterns among libraries. (a) readMap reports the read categories between libraries. High percentage of rRNA reads in L1 library resulted from the nonstandard protocol in the in-gel size fraction procedure. The non-miRNA reads accounted for much higher proportions in the libraries with less enriched read length at 22-nt. (b) readMap reports the 3′ modification patterns of miRNA reads. Such information is consistent between the libraries from the same species. Here, only the patterns more frequent than 1% are shown. (c) The top sequence denotes pre-miRNA sequence with mature miRNA marked in upper case. All isomiR forms are shown according to their relative position in pre-miRNA. The 3′ end modification patterns are presented in lower case. The middle digits and right-hand side information containing comma denote the read count and position shift of each isomiR.

of the in-gel size fraction. Figure 2 demonstrated high enrichment and low enrichment of clean reads' lengths. We further examined the categories of the clean reads, connecting length distribution and read category. As shown in Figure 3(a), miRNA dominated other molecules in L2 library. However, the miRNA reads accounted for only 16% in L1 library, much less than the rRNA reads did. Thus, the read category result was consistent with the experimental design of this study and also consistent with the result of the length distribution survey.

Table 3: MiRNA expression profile. MiRNA expression profiles of libraries were presented in the unit transcript per million (TPM). Here, only the data of the five most abundant miRNAs is shown.

Library	1st miRNA	2nd miRNA	3rd miRNA	4th miRNA	5th miRNA	%
L1	hsa-miR-30a-5p	hsa-miR-21-5p	hsa-miR-181a-5p	hsa-miR-92a-3p	hsa-miR-22-3p	51.69%
L2	hsa-miR-92a-3p	hsa-miR-22-3p	hsa-miR-143-3p	hsa-miR-10a-5p	hsa-miR-21-5p	46.41%
L3	rno-miR-143-3p	rno-miR-30a-5p	rno-miR-26a-5p	rno-miR-181a-5p	rno-miR-22-3p	52.77%
L4	rno-miR-143-3p	rno-miR-30a-5p	rno-miR-26a-5p	rno-miR-22-3p	rno-miR-10a-5p	51.43%
L5	cel-miR-58-3p	cel-miR-70-3p	cel-miR-71-5p	cel-miR-65-5p	cel-miR-241-5p	83.53%
L6	cel-miR-80-3p	cel-miR-35-3p	cel-miR-52-5p	cel-miR-72-5p	cel-miR-229-5p	38.84%
L7	dme-miR-1-3p	dme-miR-317-3p	dme-miR-276a-3p	dme-miR-263a-5p	dme-miR-184-3p	63.56%
L8	dme-miR-1-3p	dme-miR-8-3p	dme-miR-184-3p	dme-let-7-5p	dme-miR-263a-5p	86.06%

Table 4: The numbers of detected miRNAs and pre-miRNAs. The values in brackets denote the numbers of mature miRNAs and pre-miRNAs of the corresponding species according to miRBase 20 annotation. In addition to the individual miRNA expression profile, the information of all miRNA profiles is also provided.

Category	Detected miRNA	Detected pre-miRNA	Detected opp-miRNA
L1	533 (2,578)	411 (1,872)	28
L2	1098 (2,578)	824 (1,872)	86
L3	425 (728)	272 (449)	15
L4	444 (728)	288 (449)	15
L5	229 (368)	167 (223)	10
L6	259 (368)	165 (223)	9
L7	257 (426)	158 (238)	4
L8	269 (426)	167 (238)	6

Consistent with length distribution, miRNA accounted for equal proportions in both L3 and L4 libraries. The lower miRNA proportion in L5 library resulted from higher proportion of unknown reads leading to less enrichment in length distribution. Finally, for L7 library, the reads with length 30 nucleotides belonged to ncRNAs, largely lowering down the proportion of miRNA reads. Comparing Figure 2 and Figure 3(a), the results of length distribution and read category were consistent with each other and both of them are critical indices for overall NGS quality. In summary, miRSeq is able to evaluate the sequencing quality of small RNA NGS data by using self-ligation read, length distribution, and read category.

3.5. miRNA Expression Profile. The major purpose of small RNA NGS is to acquire a miRNA expression profile. By mapping reads back to pre-miRNAs, readMap yields the miRNA expression profile. The expression profiles of the five most abundant miRNAs in libraries are illustrated in Table 3. The miRNA expression profile was presented in transcript per million (TPM) so that expression abundance could be compared between different libraries exhibiting unequal miRNA reads. The expression profiles of all miRNAs in libraries are available. In addition to expression profile, the number of detected pre-miRNA and mature miRNA was also provided, as shown in Table 4. Because L1 library exhibited considerably fewer miRNA reads than L2 library did (1.36 versus 7.63 millions), fewer pre-miRNAs and mature miRNAs were detected in L1 library. Such result resulted from and is consistent with the experimental design of this study. Thus, the results of the length distribution, read category, and miRNA numbers were consistent with each other.

3.6. 3′ End Modification Patterns. In addition to miRNA expression profiles, recently, 3′ end modification patterns of miRNA reads have also caught the attentions of researches [19]. miRSeq also provides the 3′ end modification patterns of miRNA reads. As shown in Figure 3(b), the 3′ end modification patterns between libraries from the same species were pretty similar. U dominated over other patterns in frequency, followed by A. In addition, AU-rich patterns accounted for more than 70% of all patterns.

3.7. IsomiRs Forms. When mapped back to pre-miRNA, miRNA reads are generally observed to exhibit position shift or length variation compared with reference mature miRNAs. Such a phenomenon is named "isomiR." Previous studies using NGS data have showed that isomiRs performed specific regulatory functions [22, 23]. Therefore, all isomiR forms of mature miRNAs are provided by miRSeq. As illustrated in Figure 3(c), hsa-miR-2110 consisted of three isomiRs, the position shifts of which are shown. In addition, the 3′ end modification patterns are represented in lower case (i.e., a and u). According to miRBase 20, hsa-mir-2110 encodes mature miRNA only at the 5p arm, ranging from positions 8 to 29. However, applying NGS data, additional mature miRNAs can be detected at the opposite arm, named "opp-miRNA" in Table 4 [19, 23]. The opp-miRNA also consisted of isomiRs and 3′ end modification. Moreover, the newly identified opp-miRNAs could have higher expression abundance than the originally annotated miRNAs.

3.8. Time Needed for a miRSeq Alignment. miRSeq is compatible with Windows operating systems. To estimate how much time is needed for analyzing one library of NGS data, we processed the L1 NGS data by using miRSeq. As demonstrated in Table 5, miRSeq totally required 10 minutes to analyze 11 million NGS reads on the Windows 7 platform. Thus, the analysis of small RNA NGS data can be completed in little time. Even a large set of NGS data can be analyzed overnight.

Table 5: The time needed for a miRSeq alignment. Input data: L1 small RNA NGS data, totally 11,229,160 reads and accounting for approximately 1.7 GB disk space.

OS	CPU	Memory	readPro	readMap
32 bit Win. XP	Intel Pentium 4, 3.0 GHz	1.0 GB	6 min.	40 min.
32 bit Win. XP	Intel Atom D525, 1.0 GHz	3.0 GB	12 min.	52 min.
64 bit Win. 7	Intel Core i5, 1.7 GHz	4.0 GB	5 min.	4 min.
64 bit Win. Server 2008	Intel Xeon E5-2620, 2.0 GHz	16.0 GB	5 min.	7 min.

4. Conclusion

To examine whether miRSeq can precisely evaluate sequencing quality, we prepared RNA samples well or poorly. The variation in in-gel size fraction resulted in several variations between the libraries, including the percentage of qualified reads, length distribution, and read category. The alignment results proved that miRSeq is competent for sequencing quality evaluation. To demonstrate the applicability of miRSeq in animal species, we analyzed the small RNA NGS data from a wide range of animal species, including primate, rodent, insect, and nematode. The alignment results proved that miRSeq is applicable for the 105 animal species. miRSeq is a user-friendly standalone toolkit for sequencing quality evaluation and miRNA profiling from NGS data. Following step-by-step instructions, users can easily install miRSeq and can then analyze small RNA NGS data on their own PC or laptop.

5. Discussion

With the prevalence of NGS application on miRNA study, more and more toolkits, including commercial and free ones, were developed for small RNA NGS data analysis. The current free toolkits are based on online web services. So, the user must first either submit the raw sequence data in fastq format or convert the raw data into the collapsed fasta format, which usually increases the workload of internet connection and biological researchers. With the progress of PC and laptop's equipment, analyzing small RNA NGS data in PC or laptop is now applicable. miRSeq provides an alternative for the analysis and quality evaluation of small RNA NGS data, saving the network connection and biological researchers much workload and troubles.

When operating readPro, the sequence of 3′ adaptor must be specified. There are several versions of 3′ adaptor available, confusing the users. In addition, the people preparing the small RNA samples could design their own special 3′ adaptor, making it difficult to choose the correct 3′ adaptor. The user may first use a small fraction of the fastq file for test, for example, the first 10,000 records. The 3′ adaptor by referring to which readPro yields higher proportion of clean reads is usually the correct choice. Otherwise, the user may also refer to the introduction information in the corresponding SRA page for the information of 3′ adaptor. In addition, the user may also consult the people generating the NGS data.

The time needed for each miRSeq alignment depends on the size of input data and PC or laptop's CPU and memory. Of course, the faster the CPU, the shorter the time. For memory, we strongly suggest that the memory installed be double the size of input data to avoid IO errors.

Conflict of Interests

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

Authors' Contribution

Cheng-Tsung Pan and Kuo-Wang Tsai contribute equally to this paper.

Acknowledgments

The authors thank miRBase and NCBI for sharing the miRNA and mRNA data, respectively. The authors acknowledge the Genomic tRNA and SILVA database for providing tRNA and rRNA data, respectively. The authors thank the Genomics & Proteomics Core Laboratory, Department of Medical Research, Kaohsiung Chang Gung Memorial Hospital, for technical supports. This study is financially supported by the Grants from Kaohsiung Chang Gung Memorial Hospital (CLRPG8C0091 and CMRPG8D0031).

References

[1] D. P. Bartel, "MicroRNAs: genomics, biogenesis, mechanism, and function," *Cell*, vol. 116, no. 2, pp. 281–297, 2004.

[2] J. Hou, L. Lin, W. Zhou et al., "Identification of miRNomes in human liver and hepatocellular carcinoma reveals miR-199a/b-3p as therapeutic target for hepatocellular carcinoma," *Cancer Cell*, vol. 19, no. 2, pp. 232–243, 2011.

[3] D. Sun, Y. S. Lee, A. Malhotra et al., "miR-99 family of microRNAs suppresses the expression of prostate-specific antigen and prostate cancer cell proliferation," *Cancer Research*, vol. 71, no. 4, pp. 1313–1324, 2011.

[4] C. M. Zhang, J. Zhao, and H. Y. Deng, "MiR-155 promotes proliferation of human breast cancer MCF-7 cells through targeting tumor protein 53-induced nuclear protein 1," *Journal of Biomedical Science*, vol. 20, no. 1, article 79, 2013.

[5] P. Chen, C. Price, Z. Li et al., "miR-9 is an essential oncogenic microRNA specifically overexpressed in mixed lineage leukemia-rearranged leukemia," *Proceedings of the National Academy of Sciences of the United States of America USA*, vol. 110, no. 28, pp. 11511–11516, 2013.

[6] D. Wu and A. K. Murashov, "MicroRNA-431 regulates axon regeneration in mature sensory neurons by targeting the Wnt antagonist *Kremen1*," *Frontiers in Molecular Neuroscience*, vol. 6, article 35, 2013.

[7] A. Heidersbach, C. Saxby, K. Carver-Moore et al., "microRNA-1 regulates sarcomere formation and suppresses smooth muscle gene expression in the mammalian heart," *eLife*, vol. 2, Article ID e01323, 2013.

[8] G. Yao, M. Liang, N. Liang et al., "MicroRNA-224 is involved in the regulation of mouse cumulus expansion by targeting Ptx3," *Molecular and Cellular Endocrinology*, vol. 382, no. 1, pp. 244–253, 2013.

[9] P. Leidinger, C. Backes, S. Deutscher et al., "A blood based 12-miRNA signature of Alzheimer disease patients," *Genome Biology*, vol. 14, no. 7, article R78, 2013.

[10] J. Gui, Y. Tian, X. Wen et al., "Serum microRNA characterization identifies miR-885-5p as a potential marker for detecting liver pathologies," *Clinical Science*, vol. 120, no. 5, pp. 183–193, 2011.

[11] A. J. Tijsen, E. E. Creemers, P. D. Moerland et al., "MiR423-5p as a circulating biomarker for heart failure," *Circulation Research*, vol. 106, no. 6, pp. 1035–1039, 2010.

[12] B. Xiao, Y. Wang, and W. Li, "Plasma microRNA signature as a noninvasive biomarker for acute graft-versus-host disease," *Blood*, vol. 122, no. 19, pp. 3365–3375, 2013.

[13] E. Zhu, F. Zhao, G. Xu et al., "MirTools: microRNA profiling and discovery based on high-throughput sequencing," *Nucleic Acids Research*, vol. 38, supplement 2, pp. W392–W397, 2010.

[14] P. J. Huang, Y. C. Liu, C. C. Lee et al., "DSAP: deep-sequencing small RNA analysis pipeline," *Nucleic Acids Research*, vol. 38, no. 2, pp. W385–W391, 2010.

[15] K. D. Pruitt, T. Tatusova, W. Klimke, and D. R. Maglott, "NCBI reference sequences: current status, policy and new initiatives," *Nucleic Acids Research*, vol. 37, no. 1, pp. D32–D36, 2009.

[16] P. P. Chan and T. M. Lowe, "GtRNAdb: a database of transfer RNA genes detected in genomic sequence," *Nucleic Acids Research*, vol. 37, no. 1, pp. D93–D97, 2009.

[17] E. Pruesse, C. Quast, K. Knittel et al., "SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB," *Nucleic Acids Research*, vol. 35, no. 21, pp. 7188–7196, 2007.

[18] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biology*, vol. 10, no. 3, article R25, 2009.

[19] S. C. Li, K. W. Tsai, H. W. Pan, Y. M. Jeng, M. R. Ho, and W. H. Li, "MicroRNA 3' end nucleotide modification patterns and arm selection preference in liver tissues," *BMC Systems Biology*, vol. 6, supplement 2, article S14, 2012.

[20] S. L. Fernandez-Valverde, R. J. Taft, and J. S. Mattick, "Dynamic isomiR regulation in *Drosophila* development," *RNA*, vol. 16, no. 10, pp. 1881–1888, 2010.

[21] D. L. Wheeler, T. Barrett, D. A. Benson et al., "Database resources of the National Center for Biotechnology Information," *Nucleic Acids Research*, vol. 36, no. 1, pp. D13–D21, 2008.

[22] L. Guo, H. Zhang, Y. Zhao, S. Yang, and F. Chen, "Selected isomiR expression profiles via arm switching?" *Gene*, vol. 533, no. 1, pp. 149–155, 2014.

[23] S. Li, Y. Liao, M. Ho, K. Tsai, C. Lai, and W. Lin, "miRNA arm selection and isomiR distribution in gastric cancer," *BMC genomics*, vol. 13, supplement 1, p. S13, 2012.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 386124, 10 pages
http://dx.doi.org/10.1155/2014/386124

Research Article

Ultrasonographic Fetal Growth Charts: An Informatic Approach by Quantitative Analysis of the Impact of Ethnicity on Diagnoses Based on a Preliminary Report on Salentinian Population

Andrea Tinelli,[1] Mario Alessandro Bochicchio,[2] Lucia Vaira,[2] and Antonio Malvasi[3]

[1] *Division of Experimental Endoscopic Surgery, Imaging, Technology and Minimally Invasive Therapy, Centre for Interdisciplinary Research Applied to Medicine and Department of Obstetrics and Gynecology, Vito Fazzi Hospital, Piazza Muratore, 73100 Lecce, Italy*
[2] *Department of Engineering for Innovation, Software Engineering and Telemedia LAB (SET-Lab), University of Salento, 73100 Lecce, Italy*
[3] *Department of Obstetrics and Gynecology, Santa Maria Hospital, 70124 Bari, Italy*

Correspondence should be addressed to Andrea Tinelli; andreatinelli@gmail.com

Received 27 February 2014; Accepted 24 May 2014; Published 18 June 2014

Academic Editor: Li-Ching Wu

Clear guidance on fetal growth assessment is important because of the strong links between growth restriction or macrosomia and adverse perinatal outcome in order to reduce associated morbidity and mortality. Fetal growth curves are extensively adopted to track fetal sizes from the early phases of pregnancy up to delivery. In the literature, a large variety of reference charts are reported but they are mostly up to five decades old. Furthermore, they do not address several variables and factors (e.g., ethnicity, foods, lifestyle, smoke, and physiological and pathological variables), which are very important for a correct evaluation of the fetal well-being. Therefore, currently adopted fetal growth charts are inadequate to support the melting pot of ethnic groups and lifestyles of our society. Customized fetal growth charts are needed to provide an accurate fetal assessment and to avoid unnecessary obstetric interventions at the time of delivery. Starting from the development of a growth chart purposely built for a specific population, in the paper, authors quantify and analyse the impact of the adoption of wrong growth charts on fetal diagnoses. These results come from a preliminary evaluation of a new open service developed to produce personalized growth charts for specific ethnicity, lifestyle, and other parameters.

1. Introduction

In current clinical medicine, data coming from medical records and analysis are often used to document diagnostic issue, giving the opportunity of a systematic data meta-analysis to improve patient care and to develop new health-assessment techniques.

Correct assessment of gestational age and fetal growth is essential for optimal obstetric management. For this purpose, ultrasound obstetric scans in pregnancy are routinely used to track fetal growth and to assess fetal health.

Fetal size charts are used to compare the size of a fetus (of known gestational age) with reference data and to compare it on two or more different circumstances.

This can be performed using look-up tables or charts, but, as it is easier to identify any deviation from normal by plotting measurements on charts, the use of charts is recommended and the clinical evidence supports their efficacy.

The detection of a potential abnormal growth by means of intrauterine fetal parameters during pregnancy was proposed by serial US scans by Lubchenco [1], Usher and McLean [2], and Babson and Benda [3], more than five decades ago, and fetal growth assessment is a well-established and mature research field in obstetrics and gynecology [4–6].

Fetal growth charts are compared to statistical data (i.e., reference charts with fetal growth curves, showing average values of biometric parameters as a function of the gestational

age) so that clinicians may detect fetal growth associated to fetal intrauterine anomalies [7].

Numerous studies have been conducted to derive reference charts for fetal size. Many, however, have a suboptimal design, using a hospital-based population or having an inappropriate sample size.

The proliferation of further studies on specific subgroups of patients [8–12] and the related proposal of an ever increasing number of reference charts were characterized by a considerable methodological heterogeneity, making them difficult to use for diagnostic purposes.

As a consequence in clinical practice, generic charts are preferred to specific ones or to more complex approaches based on suitable mathematic models [11], because of their feasibility.

Moreover, the World Health Organization (WHO) standards are still commonly based on generic reference charts; they do not differentiate by ethnic origin and are not subject to frequent update, so they are unsuitable to assess the biometric parameters in several cases of practical interest.

To preserve the feasibility of the approach without losing diagnostic power, some authors proposed the adoption of purposely developed software tools (Web Applications, Mobile Application, etc.) allowing us to create customized growth charts [13, 14], based on a regression model fitted to a very large group of newborns.

Medical literature clearly showed its main drawbacks: (A) the number of patients considered in the studies (some thousandth) is low with respect to the total number of newborn per year (about 160 Millions in 2013) in the world; (B) patients considered in the studies are not representative of the variety of anthropometrical factors due to ethnicity, familial aspects, and other relevant internal and external factors; (C) the commonly used growth curves are up to five decades old; they are not updated for the current population and they are not suitable to investigate temporal trends and dynamic aspects in fetal growth curves.

Nevertheless, fetal growth is influenced by a variety of factors, racial, social, and economic among others, as well as specific medical conditions that may preexist or that may develop during pregnancy.

Hence, it is not surprising that fetal biometric parameters show high degree of variation in evaluated population from country to country and from area to area, within the same country. Beyond ethnicity, many other factors affect fetal growth including fetus gender, physiological and pathological variables, maternal height and weight, drug or tobacco exposure, genetic syndromes, congenital anomalies, and placental failure [15–18].

In this context, it is necessary to have personalized charts for fetal growth in order to provide an accurate fetal assessment and to make the presence of false positive and false negative potentially avoidable.

The adoption of wrong reference curves on specific fetuses could cause an incorrect evaluation of fetal biometric parameters, identifying for example cases known in literature as Small for Gestational Age (SGA) or Large for Gestational Age (LGA). So, using personalized growth curves would result in a considerable decline in the rate of a false-positive diagnosis of SGA/LGA.

In this scenario, authors quantify and analyse the impact of the adoption of such wrong growth charts on fetal diagnoses. As initial results, authors show how much different are values and boundaries of certain biometric parameters according to ethnicity. Salentinian population (southeast of Italy) has been analysed and its samples have been compared with the reference curves adopted for Italian [19] and European [20] fetuses.

These preliminary results have been obtained by adopting a new "online service" in charge to develop personalized growth charts, which take into account differences due to ethnicity, lifestyle, familial aspects, and other parameters.

2. Material and Methods

The study includes a population of about 500 Italian women undergoing ultrasound examination between the 11th and 41th weeks of gestation, between November 2012 and September 2013.

All pregnant women were enrolled in a previously defined area, southeast of Italy, in the Vito Fazzi Hospital, Italy, and Departments of Obstetrics and Gynecology assessed the investigation.

Gestational age was established by using US imaging during the first visit, at study enrolment. All patients received written and oral information about the study, and they signed the informed consent.

3. Data Harvesting Methodology

Before enrolment, authors defined, in the setup study, the inclusion and exclusion criteria Inclusion criteria were: singleton pregnancies, known first day of the Last Menstrual Period (LMP), regular cycle (lasting 28 ± 4 days) The date of the LMP was confirmed with the pregnant woman at the first obstetric visit, and additional information on regularity and duration of the cycle was collected during visit. Cases with low birth weight, preterm delivery, or other prenatal complications were not excluded from analysis. Gestational age was based on the last menstrual period and in all cases adjusted according to the CRL measured in the first trimester ultrasound.

Pregnant women were excluded from analysis if they joined the study after the 24th week of pregnancy, because reliable dating of pregnancy is more difficult as pregnancy proceeds. Bidimensional (2D) US scans were conducted either with a Logic 7 Pro US system (GE-Kretz, Zipf, Austria), an IU 22 xMATRIX US system (Philips Healthcare, Eindhoven, The Netherlands), or a Voluson 730 US system (GE-Kretz, Zipf, Austria) equipped with a 3.8–5.2 MHz transabdominal transducer by resident clinicians well-trained in obstetric US. All machines had a standard US setting of Doppler and grey scale, provided by companies. Measurements of the biparietal diameter (BPD) and head circumference (HC) were obtained from a transverse axial plane of the fetal head showing a central midline echo broken in the anterior third by the cavum of septum pellucidi and demonstrating the anterior and posterior horns of the lateral ventricle. The BPD was measured from the outer margin

of the proximal skull to the inner margin of the distal skull. The HC was measured fitting a computer-generated ellipse to include the outer edges of the calvarial margins of the fetal skull. The abdominal circumference (AC) was measured fitting a computer-generated ellipse through a transverse section of the fetal abdomen at the level of the stomach and bifurcation of the main portal vein into its right and left branches. The femur length (FL) was measured in a longitudinal scan where the whole femural diaphysis was seen almost parallel to the transducer and measured from the greater trochanter to the lateral condyle. In the third trimester, particular care was taken not to include the epiphysis.

4. Statistical Methods

Each interval of gestational age was centred on a week, so that from 13 weeks and 4 days up to 14 weeks and 3 days has been considered as 14th week.

Statistical analysis has been performed using appropriate packages of R Software (http://www.r-project.org).

The normality of measurements at each week of gestation was assessed using the Shapiro-Wilk test [21], which is one of the most powerful tests to use for the normality assessment, especially for small samples. It tests the null hypothesis that a given sample came from a normally distributed population.

In order to obtain normal ranges for fetal measurements, a multistep procedure based on regression model has been used, according to the recommended methodology for this type of data [22, 23].

Assuming that, at each gestational age, the measurement of interest has a Gaussian distribution with a mean and a standard deviation (SD) and that, in general, both vary smoothly with gestational age, a centile curve has been calculated using the well-known formula:

$$\text{Centile} = \text{mean} + K * \text{SD}, \tag{1}$$

where K is the corresponding centile of the standard Gaussian distribution (e.g., determination of 10th and 90th centile curves requires that $K = \pm 1.28$), mean is the mean, and SD is the standard deviation of the mean of the fetal measurements for each gestational age.

The mean has been estimated by the fitted values from an appropriate polynomial regression curve of the measurement of interest on gestational age.

Several curve-fitting and smoothing techniques have been tested for the mean estimation of the different biometric parameters and the goodness of fit for each regression model has been carefully assessed. The polynomial model that better satisfies the experimental data is the cubic one, since it better fulfils the fractional polynomial and the logarithmic transformations.

The adopted equation is

$$y = a + (b * \text{GA}) + \left(c * \text{GA}^2\right) + \left(d * \text{GA}^3\right). \tag{2}$$

When the measurement has approximately a Gaussian distribution, the fitted values following regression of the "scaled absolute residuals" on age are estimate of the SD curve. These residuals are the difference between the measurements and the estimated curve for the mean with the sign removed and multiplied by a corrective constant equal to $\sqrt{(\pi/2)} = 1.253$.

Generally, if the scaled absolute residuals appear to show no trend with gestational age, the SD is estimated as the standard deviation of the unscaled residuals (measurements minus the estimated mean curve). If there is a trend, then polynomial regression analysis is needed to estimate an appropriate curve in the same way of the mean.

For BPD, HC, and AC biometric parameters, the residuals were regressed on gestational ages by using a linear model in the form of

$$y_{\text{BPD,HC,AC}} = a + (b * \text{GA}), \tag{3}$$

While, considering the FL parameter, the quadratic regression seems to better fulfil the linear one. The adopted equation is

$$y_{\text{FL}} = a + (b * \text{GA}) + \left(c * \text{GA}^2\right). \tag{4}$$

Finally, these predictive mean and SD equations allow calculating any required centile, replacing the value in the centile formula.

5. Results

Full biometric measurements (AC, BPD, FL, and HC) were obtained for about 500 fetuses.

Data analysis showed that neither the use of fractional polynomials (the greatest power of the polynomials being 3) nor the logarithmic transformation improved the fitting of the curves. Therefore, the data were kept in their original scale. The best-fitted regression model to describe the relationships between HC, AC, BPD, and FL and gestational age was a cubic one, whereas other studies proved that a simple quadratic model fitted BPD and FL [24].

Models fitting the SD were straight lines for BPD, HC, and AC and quadratic line for FL.

To choose the best fitting model, we have taken into consideration primarily the R^2 index (which is the linear determination index: in the ideal case its value should be equal to 1; in real cases it is near to 1 if the interpolating curve is a good approximation of the real data set) but the value of R^2 alone is not the only factor that we have considered in choosing the best model. Other factors we have considered include the validity and the effectiveness of the model.

There will be an improvement in fit as higher-order terms are added, but because these terms are not theoretically justified, the improvement will be sample-specific.

Unless the sample is very small, the fits of higher-order polynomials are unlikely to be very different from those of a quadratic over the main part of the data range.

Consider that, for example, the R^2 for the quadratic specification of BPD parameter is 0.98081 and for the cubic and quartic curves it is 0.98229 and 0.98242, relatively small improvements.

Further, the cubic and quartic curves both exhibit implausible strange twists at the extremities (Figures 1 and 2).

The scatter of absolute residuals from the regression for estimation of the standard deviation of femur length as a function of gestational age is shown in Figure 3.

Table 1: Regression equations for the mean and the standard deviation of AC, BPD, FL, and HC.

Fetal parameter	Regression equations	R^2 (%)
Abdominal circumference (AC)		
Mean	$2{,}2783 - 0{,}07057\ GA + 0{,}05214\ GA^2 - 0{,}0007706\ GA^3$	94
SD	$0{,}0284 + 0{,}354 * GA$	
Biparietal diameter (BPD)		
Mean	$-0{,}0377 + 0{,}10476\ GA + 0{,}012021\ GA^2 - 0{,}0002124\ GA^3$	98
SD	$0{,}0762 + 0{,}0042 * GA$	
Femur length (FL)		
Mean	$-1{,}2394 + 0{,}13735\ GA + 0{,}007537\ GA^2 - 0{,}000132\ GA^3$	98
SD	$0{,}3186 - 0{,}0162 * GA + 0{,}0004 * GA^2$	98
Head circumference (HC)		
Mean	$-9{,}8027 + 1{,}4258\ GA + 0{,}006556\ GA^2 - 0{,}0003765\ GA^3$	98
SD	$0{,}453 + 0{,}0102 * GA$	

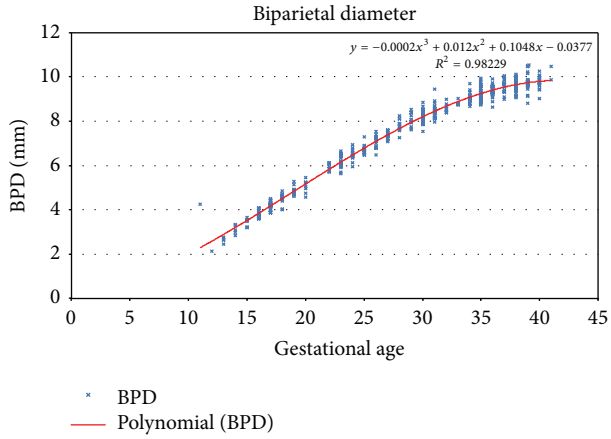


Figure 1: Third order polynomial regression for biparietal diameter.

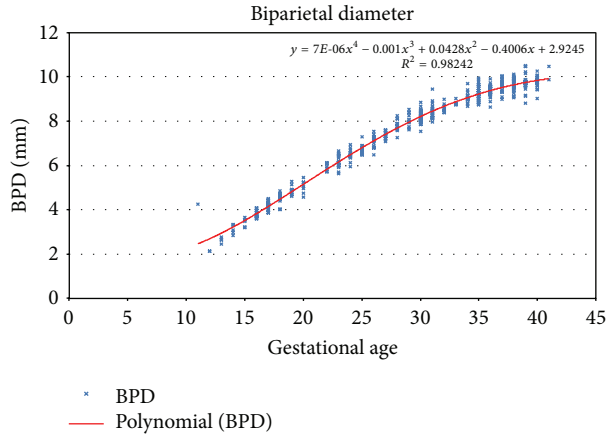


Figure 2: Fourth order polynomial regression for biparietal diameter.

The corresponding regression equations, with the respective R^2 index for the mean and the standard deviation, are illustrated in Table 1.

Table 1 shows regression equations for the mean and the standard deviation of AC, BPD, FL, and HC.

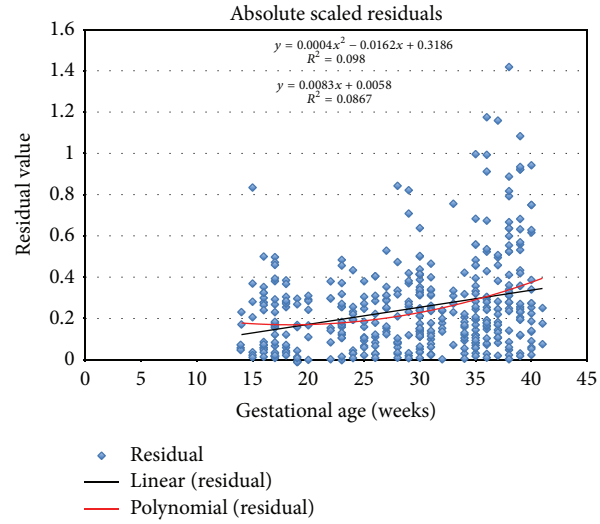


Figure 3: Absolute scaled residuals for FL measurement.

The relevant centile (5th, 10th, 50th, 90th, and 95th), representing, respectively, the HC, the BPD, the A,C and the FL, are reported in Tables 2, 3, 4, and 5. In each table, it is also indicated that the sample number, the mean, and the standard deviation are related to each gestational week.

6. Discussion

In order to validate the system, authors have performed an initial technical test with a growth curve simulator able to respect the mean and the standard deviation that characterize the Gaussian distribution for a specific patient age. The generated data allowed authors to prove the correctness of the elaboration of the fetal growth curves model.

After this preliminary analysis, authors have performed a test on the field considering about 500 US pictures related to Italian women undergoing ultrasound examination between the 11th and 41th weeks of gestation at Vito Fazzi Hospital, Lecce, between November 2012 and September 2013. Measurements of biparietal diameter (BPD), head circumference

Table 2: Fitted centiles of head circumference (mm).

GA (weeks)	Sample size (n)	Mean HC (mm)	SD	Percentile				
				5th	10th	50th	90th	95th
14	5	104,10	0,60	94,31	96,48	104,10	111,73	113,89
15	7	117,89	0,61	107,93	110,13	117,89	125,64	127,84
16	17	131,46	0,62	121,34	123,57	131,46	139,35	141,59
17	30	144,81	0,63	134,52	136,79	144,81	152,83	155,10
18	13	157,90	0,64	147,44	149,75	157,90	166,05	168,36
19	12	170,72	0,65	160,09	162,44	170,72	179,00	181,35
20	7	183,24	0,66	172,44	174,83	183,24	191,65	194,03
21	—	195,44	0,67	184,47	186,89	195,44	203,98	206,40
22	7	207,29	0,68	196,16	198,62	207,29	215,96	218,42
23	20	218,78	0,69	207,48	209,98	218,78	227,58	230,08
24	13	229,88	0,70	218,41	220,95	229,88	238,81	241,35
25	13	240,57	0,71	228,94	231,51	240,57	249,63	252,20
26	14	250,83	0,72	239,02	241,63	250,83	260,02	262,63
27	12	260,63	0,73	248,66	251,30	260,63	269,95	272,60
28	12	269,95	0,74	257,81	260,49	269,95	279,40	282,08
29	13	278,77	0,75	266,46	269,18	278,77	288,35	291,07
30	42	287,06	0,76	274,59	277,34	287,06	296,78	299,54
31	23	294,81	0,77	282,17	284,96	294,81	304,66	307,45
32	8	301,99	0,78	289,18	292,01	301,99	311,97	314,80
33	12	308,58	0,79	295,60	298,47	308,58	318,69	321,56
34	15	314,55	0,80	301,41	304,31	314,55	324,79	327,70
35	32	319,89	0,81	306,58	309,52	319,89	330,26	333,20
36	27	324,57	0,82	311,09	314,06	324,57	335,07	338,05
37	19	328,56	0,83	314,91	317,93	328,56	339,20	342,21
38	28	331,85	0,84	318,04	321,09	331,85	342,62	345,67
39	25	334,42	0,85	320,43	323,52	334,42	345,31	348,40
40	15	336,23	0,86	322,08	325,20	336,23	347,25	350,38
41	3	337,27	0,87	322,95	326,11	337,27	348,43	351,59

(HC), abdominal circumference (AC), and femur length (FL) were obtained during the clinical practice.

The obtained curves were then compared with those developed by Giorlandino et al. [19] as reference growth curves for the Italian population and those developed by Johnsen et al. [20] as reference growth curves for the European population, in order to verify possible differences due to statistic methodology, selection criteria, or, possibly, true genetic variability of the studied population.

The AC and HC biometric parameters seem to follow more or less the same Italian and European trend according to the gestational age. In fact, no significant differences were observed in the values measured during the different growth stages. Considering the BPD and the FL parameters, instead, they present a little variability.

As shown in Figures 4 and 5, the Salentinian BPD values are always up for about 6 mm and FL ones are always greater than 7 mm.

This variability may be better presented by means of scatterplot of Salentinian samples overlapped with the centile

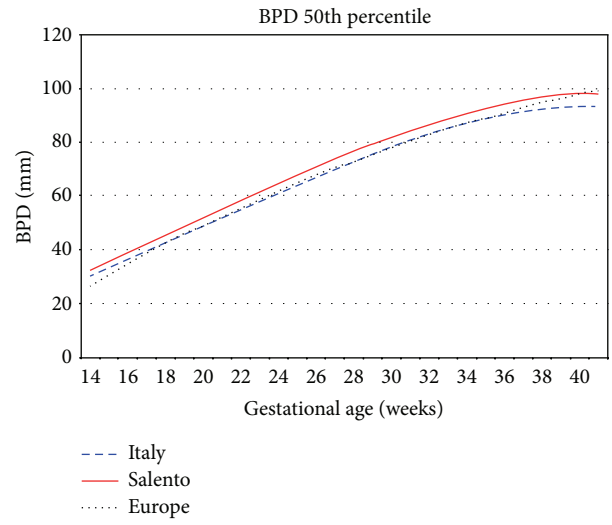


Figure 4: Biparietal diameter 50th percentile Salento versus Italy versus Europe.

Table 3: Fitted centiles of biparietal diameter (mm).

GA (weeks)	Sample size (n)	Mean BPD (mm)	SD	Percentile				
				5th	10th	50th	90th	95th
14	5	32,02	0,14	29,80	30,29	32,02	33,75	34,24
15	7	35,22	0,14	32,93	33,43	35,22	37,00	37,51
16	17	38,46	0,14	36,10	36,62	38,46	40,30	40,82
17	30	41,74	0,15	39,31	39,85	41,74	43,63	44,17
18	14	45,04	0,15	42,54	43,10	45,04	46,99	47,54
19	14	48,35	0,16	45,79	46,36	48,35	50,35	50,92
20	7	51,67	0,16	49,03	49,61	51,67	53,72	54,30
21	—	54,96	0,16	52,26	52,86	54,96	57,07	57,67
22	7	58,24	0,17	55,46	56,07	58,24	60,40	61,01
23	20	61,47	0,17	58,62	59,25	61,47	63,68	64,31
24	13	64,64	0,18	61,73	62,38	64,64	66,91	67,56
25	13	67,76	0,18	64,78	65,43	67,76	70,08	70,74
26	13	70,79	0,19	67,74	68,42	70,79	73,17	73,84
27	13	73,73	0,19	70,62	71,30	73,73	76,16	76,85
28	12	76,57	0,19	73,39	74,09	76,57	79,06	79,76
29	13	79,30	0,20	76,04	76,76	79,30	81,84	82,55
30	43	81,89	0,20	78,57	79,30	81,89	84,48	85,22
31	22	84,34	0,21	80,95	81,70	84,34	86,99	87,74
32	8	86,64	0,21	83,18	83,94	86,64	89,34	90,11
33	6	88,77	0,21	85,24	86,02	88,77	91,53	92,31
34	15	90,72	0,22	87,12	87,92	90,72	93,53	94,32
35	33	92,48	0,22	88,81	89,62	92,48	95,34	96,15
36	30	94,03	0,23	90,29	91,12	94,03	96,95	97,77
37	18	95,36	0,23	91,56	92,40	95,36	98,33	99,17
38	30	96,47	0,24	92,59	93,44	96,47	99,49	100,35
39	26	97,33	0,24	93,38	94,25	97,33	100,40	101,27
40	17	97,93	0,24	93,91	94,80	97,93	101,06	101,94
41	2	98,26	0,25	94,17	95,08	98,26	101,44	102,35

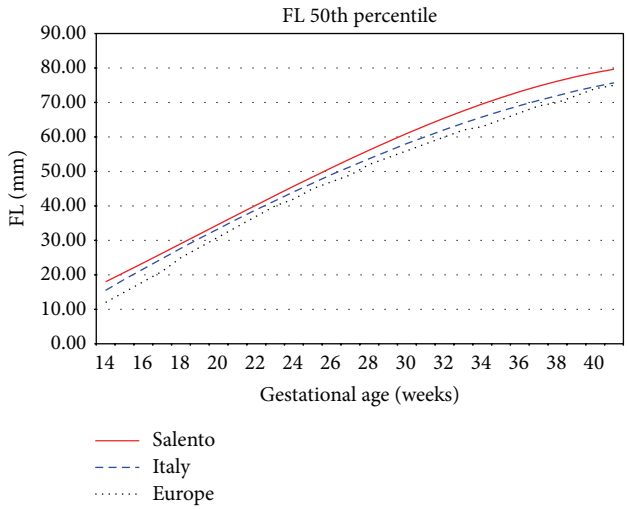


Figure 5: Femur length 50th percentile Salento versus Italy versus Europe.

curves to verify the amount and the density of the samples that are outside the considered range.

Considering the Italian reference centile curves depicted in Figure 6, which represent, respectively, the 5th, 10th, 25th, 50th, 75th, 90th, and 95th, the Salentinian samples are always above the upper limit, especially in the last weeks of gestation.

Samples above the 95th centile are traditionally used to define large for gestational age (LGA), and the usage of such Italian reference curves on a Salentinian fetus could lead to misdiagnosis.

To examine in a quick way one or more sets of data graphically, box plots can be used. They can be useful to indicate the degree of dispersion (spread) and skewness in the data and to identify outliers. Each plot depicts the five-number summaries for each biometric parameter, namely, the minimum and maximum values, the upper (Q3) and lower (Q1) quartiles, and the median.

The variability present in the FL parameter can be also observed in this kind of graph, which considers more population groups.

Table 4: Fitted centiles of abdominal circumference (mm).

GA (weeks)	Sample size (n)	Mean AC (mm)	SD	Percentile				
				5th	10th	50th	90th	95th
14	5	93,95	0,52	85,33	87,24	93,95	100,67	102,57
15	6	103,50	0,56	94,30	96,34	103,50	110,67	112,71
16	17	113,41	0,59	103,62	105,78	113,41	121,03	123,19
17	29	123,61	0,63	113,24	115,53	123,61	131,69	133,98
18	13	134,07	0,67	123,12	125,54	134,07	142,60	145,02
19	12	144,74	0,70	133,21	135,76	144,74	153,73	156,28
20	7	155,58	0,74	143,47	146,14	155,58	165,02	167,69
21	—	166,54	0,77	153,84	156,64	166,54	176,43	179,23
22	7	177,56	0,81	164,28	167,22	177,56	187,91	190,84
23	19	188,61	0,84	174,75	177,81	188,61	199,41	202,47
24	13	199,64	0,88	185,20	188,39	199,64	210,90	214,09
25	12	210,61	0,91	195,58	198,90	210,61	222,32	225,63
26	14	221,46	0,95	205,85	209,30	221,46	233,62	237,07
27	12	232,15	0,98	215,96	219,54	232,15	244,77	248,34
28	12	242,64	1,02	225,87	229,57	242,64	255,71	259,41
29	13	252,87	1,06	235,52	239,35	252,87	266,39	270,23
30	43	262,81	1,09	244,87	248,84	262,81	276,78	280,75
31	22	272,40	1,13	253,88	257,97	272,40	286,83	290,92
32	7	281,60	1,16	262,50	266,72	281,60	296,49	300,70
33	6	290,37	1,20	270,69	275,03	290,37	305,70	310,05
34	14	298,65	1,23	278,39	282,86	298,65	314,44	318,92
35	33	306,40	1,27	285,56	290,16	306,40	322,65	327,25
36	25	313,58	1,30	292,15	296,88	313,58	330,28	335,01
37	19	320,14	1,34	298,12	302,99	320,14	337,29	342,15
38	29	326,02	1,37	303,43	308,42	326,02	343,63	348,62
39	26	331,20	1,41	308,02	313,14	331,20	349,26	354,37
40	18	335,61	1,44	311,85	317,10	335,61	354,12	359,37
41	3	339,22	1,48	314,88	320,25	339,22	358,18	363,56

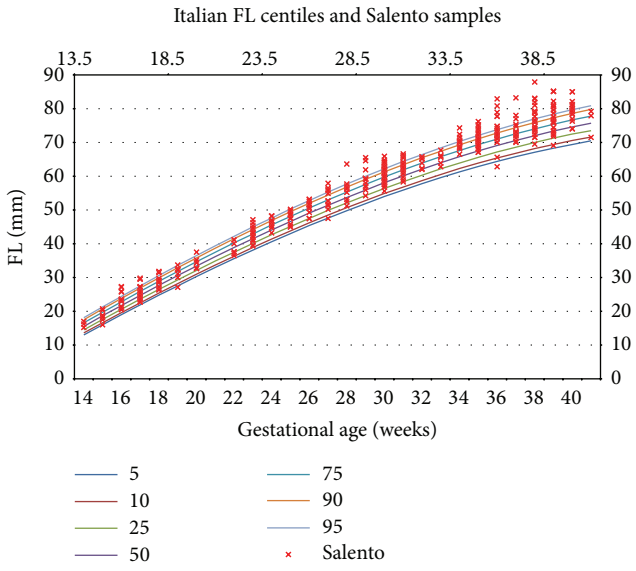


Figure 6: Femur length: Italian centiles and Salentinian samples.

As can be seen in Figure 7, an average length that is similar to that of Germany characterizes the Salentinian femur. Its maximum value is rather close to that of the UK.

This variability has to be medically investigated since it can be due to several reasons: equipment or measurement errors, genetic variability of the analysed population, racial factors, and so on.

In any case, the measured variability is useful to demonstrate the effectiveness of the proposed approach.

The complete set of curves obtained from the mentioned dataset and the complete description of the mathematical procedure adopted for the analysis are published and described at http://www.fpgt.unisalento.it/FPGT/Projects/scientificFoundations.php.

In order to quantify the impact of the adoption of wrong growth charts on fetal diagnoses, authors have analysed the samples' trend for each biometric parameter and have then compared it with the Italian and European standard.

Authors found significant differences between Salentinian FL growth plots and those reported by Giorlandino et al. [19] for Italy and Johnsen et al. [20] for Europe.

Table 5: Fitted centiles of femur length (mm).

GA (weeks)	Sample size (n)	Mean AC (mm)	SD	Percentile				
				5th	10th	50th	90th	95th
14	5	17,99	0,17	15,19	15,80	17,99	20,17	20,79
15	7	20,71	0,17	17,99	18,59	20,71	22,83	23,44
16	17	23,47	0,16	20,81	21,40	23,47	25,54	26,13
17	29	26,25	0,16	23,64	24,22	26,25	28,29	28,86
18	13	29,05	0,16	26,47	27,04	29,05	31,06	31,63
19	12	31,86	0,16	29,30	29,87	31,86	33,85	34,41
20	7	34,66	0,15	32,12	32,68	34,66	36,65	37,21
21	—	37,46	0,15	34,92	35,48	37,46	39,45	40,01
22	7	40,25	0,16	37,68	38,25	40,25	42,24	42,81
23	21	43,01	0,16	40,41	40,99	43,01	45,03	45,60
24	12	45,74	0,16	43,10	43,68	45,74	47,79	48,37
25	13	48,42	0,16	45,73	46,33	48,42	50,52	51,12
26	14	51,07	0,17	48,31	48,92	51,07	53,22	53,83
27	12	53,65	0,17	50,81	51,44	53,65	55,87	56,50
28	12	56,18	0,18	53,24	53,89	56,18	58,47	59,12
29	13	58,63	0,19	55,58	56,26	58,63	61,00	61,68
30	42	61,00	0,19	57,84	58,54	61,00	63,47	64,17
31	21	63,29	0,20	59,99	60,72	63,29	65,86	66,59
32	7	65,48	0,21	62,03	62,79	65,48	68,17	68,93
33	6	67,57	0,22	63,96	64,76	67,57	70,39	71,18
34	14	69,55	0,23	65,76	66,60	69,55	72,50	73,34
35	31	71,41	0,24	67,44	68,32	71,41	74,51	75,39
36	25	73,15	0,25	68,97	69,89	73,15	76,40	77,32
37	19	74,75	0,27	70,36	71,33	74,75	78,16	79,13
38	29	76,20	0,28	71,59	72,61	76,20	79,80	80,82
39	28	77,51	0,30	72,65	73,73	77,51	81,29	82,36
40	18	78,66	0,31	73,55	74,68	78,66	82,64	83,77
41	3	79,64	0,33	74,27	75,45	79,64	83,83	85,02

Table 6: Synthetic values for biparietal diameter (BPD).

	Biparietal diameter					
	Salento		Italy		Europe	
	Samples	%	Samples	%	Samples	%
>95th centile	49/448	10,93%	99/448	22%	69/448	15,40%
<5th centile	41/448	9,15%	0/448	0%	2/448	0,44%

From Tables 6, 7, 8, and 9, we describe this difference, representing the sample number and the percentage value for each biometric parameters (BPD, HC, AC, and FL) which exceed the upper limit (95th centile) and the lower one (5th centile) considering the Italian and European reference curves.

7. Conclusions

The fetal growth assessment is a relevant problem, since it concerns about 160 ML of newborns per year. The population reshuffling and the increased mobility of families push for a new assessment approach based on dynamic and individualized fetal growth curves.

The importance of the growth curves is proven by the fact that they are commonly used in neonatal units today. They serve as standard references to classify neonates as SGA, LGA, and AGA. In order to evaluate the applicability of these standards to current patients, we compared data accumulated in our research data system to determine whether our patients were categorized appropriately.

Our findings require that we should carefully reexamine the appropriateness of continued use of currently adopted reference growth curves to classify neonates SGA, LGA, and AGA.

Table 7: Synthetic values for abdominal circumference (AC).

	Abdominal circumference					
	Salento		Italy		Europe	
	Samples	%	Samples	%	Samples	%
>95th centile	34/436	7,79%	7/436	1,60%	20/436	4,58%
<5th centile	21/436	4,81%	13/436	2,98%	13/436	2,98%

Table 8: Synthetic values for head circumference (HC).

	Head circumference					
	Salento		Italy		Europe	
	Samples	%	Samples	%	Samples	%
>95th centile	17/444	3,82%	4/444	0,90%	65/444	14,60%
<5th centile	23/444	5,18%	20/444	4,50%	1/444	0,22%

Table 9: Synthetic values for femur length (FL).

	Femur length					
	Salento		Italy		Europe	
	Samples	%	Samples	%	Samples	%
>95th centile	42/437	9,61%	115/437	26,00%	202/437	46%
<5th centile	30/437	6,86%	1/437	0,20%	0/437	0%

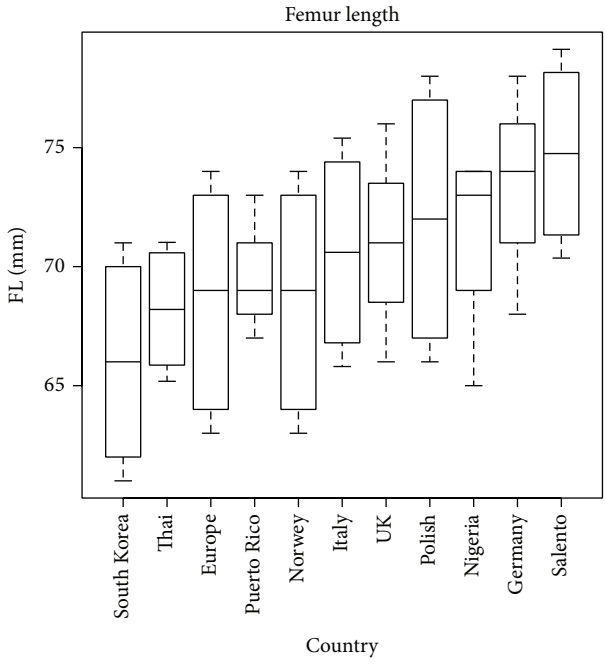


Figure 7: Femur length box plot.

In fact, considering, for example, the femur length parameter, Salentinian fetuses present bigger values with respect to those of Italy (26% of Salentinian samples are above the 95th centile) and Europe (46% of Salentinian samples are above the 95th centile).

This is a preliminary approach, which does not represent the development and publication of new reference curves for Salentinian population but rather represents the introduction of a new method to construct the fetal growth curves which has to take into consideration several information about ethnicity, foods, lifestyle, drugs assumption, and other internal or external factors influencing growth.

Conflict of Interests

Authors certify that there is no actual or potential conflict of interests in relation to this paper and they reveal no financial interests or connections, direct or indirect, or other situations that might raise the question of bias in the work reported or the conclusions, implications, or opinions stated—including pertinent commercial or other sources of funding for the individual authors or for the associated departments or organizations, personal relationships, or direct academic competition.

References

[1] L. O. Lubchenco, C. Hansman, and E. Boyd, "Intrauterine growth in length and head circumference as estimated from live births at gestational ages from 26 to 42 weeks," *Pediatrics*, vol. 37, no. 3, pp. 403–408, 1966.

[2] R. Usher and F. McLean, "Intrauterine growth of live-born Caucasian infants at sea level: standards obtained from measurements in 7 dimensions of infants born between 25 and 44 weeks," *Journal of Pediatrics*, vol. 74, no. 6, pp. 901–910, 1969.

[3] S. G. Babson and G. I. Benda, "Growth graphs for the clinical assessment of infants of varying gestational age," *Journal of Pediatrics*, vol. 89, no. 5, pp. 814–820, 1976.

[4] J. Gardosi, M. Mongelli, M. Wilcox, and A. Chang, "An adjustable fetal weight standard," *Ultrasound in Obstetrics & Gynecology*, vol. 6, no. 3, pp. 168–174, 1995.

[5] S. Bonellie, J. Chalmers, R. Gray, I. Greer, S. Jarvis, and C. Williams, "Centile charts for birthweight for gestational age for Scottish singleton births," *BMC Pregnancy and Childbirth*, vol. 8, article 5, 2008.

[6] T. R. Fenton, "A new growth chart for preterm babies: Babson and Benda's chart updated with recent data and a new format," *BMC Pediatrics*, vol. 3, article 13, 2003.

[7] F. P. Hadlock, R. L. Deter, R. J. Carpenter, and S. K. Park, "Estimating fetal age: effect of head shape on BPD," *American Journal of Roentgenology*, vol. 137, no. 1, pp. 83–85, 1981.

[8] M. Mongelli and J. Gardosi, "Longitudinal study of fetal growth in subgroups of a low-risk population," *Ultrasound in Obstetrics & Gynecology*, vol. 6, no. 5, pp. 340–344, 1995.

[9] M. S. Kramer, R. W. Platt, S. W. Wen et al., "A new and improved population-based Canadian reference for birth weight for gestational age," *Pediatrics*, vol. 108, no. 2, article e35, 2001.

[10] L. McCowan, A. W. Stewart, A. Francis, and J. Gardosi, "A customised birthweight centile calculator developed for a New Zealand population," *Australian and New Zealand Journal of Obstetrics and Gynaecology*, vol. 44, no. 5, pp. 428–431, 2004.

[11] L. J. Salomon, M. Duyme, J. Crequat et al., "French fetal biometry: reference equations and comparison with other charts," *Ultrasound in Obstetrics and Gynecology*, vol. 28, no. 2, pp. 193–198, 2006.

[12] B. O. Verburg, E. A. P. Steegers, M. de Ridder et al., "New charts for ultrasound dating of pregnancy and assessment of fetal growth: longitudinal data from a population-based cohort study," *Ultrasound in Obstetrics and Gynecology*, vol. 31, no. 4, pp. 388–396, 2008.

[13] J. Gardosi, A. Chang, B. Kalyan, D. Sahota, and E. M. Symonds, "Customised antenatal growth charts," *The Lancet*, vol. 339, no. 8788, pp. 283–287, 1992.

[14] E. Bertino, E. Spada, L. Occhi et al., "Neonatal anthropometric charts: the Italian neonatal study compared with other European

studies," *Journal of Pediatric Gastroenterology & Nutrition*, vol. 51, no. 3, pp. 353–361, 2010.

[15] J. C. Drooger, J. W. M. Troe, G. J. J. M. Borsboom et al., "Ethnic differences in prenatal growth and the association with maternal and fetal characteristics," *Ultrasound in Obstetrics and Gynecology*, vol. 26, no. 2, pp. 115–122, 2005.

[16] A. Niklasson and K. Albertsson-Wikland, "Continuous growth reference from 24th week of gestation to 24 months by gender," *BMC Pediatrics*, vol. 8, article 8, 2008.

[17] E. Oken, K. P. Kleinman, J. Rich-Edwards, and M. W. Gillman, "A nearly continuous measure of birth weight for gestational age using a United States national reference," *BMC Pediatrics*, vol. 3, article 6, 2003.

[18] P. Schwärzler, J. M. Bland, D. Holden, S. Campbell, and Y. Ville, "Sex-specific antenatal reference growth charts for uncomplicated singleton pregnancies at 15–40 weeks of gestation," *Ultrasound in Obstetrics and Gynecology*, vol. 23, no. 1, pp. 23–29, 2004.

[19] M. Giorlandino, F. Padula, P. Cignini et al., "Reference interval for fetal biometry in Italian population," *Journal of Prenatal Medicine*, vol. 3, no. 4, pp. 62–65, 2009.

[20] S. L. Johnsen, T. Wilsgaard, S. Rasmussen, R. Sollien, and T. Kiserud, "Longitudinal reference charts for growth of the fetal head, abdomen and femur," *European Journal of Obstetrics Gynecology and Reproductive Biology*, vol. 127, no. 2, pp. 172–185, 2006.

[21] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3-4, pp. 591–611, 1965.

[22] D. G. Altman and L. S. Chitty, "New charts for ultrasound dating of pregnancy," *Ultrasound in Obstetrics and Gynecology*, vol. 10, no. 3, pp. 174–191, 1997.

[23] P. Royston and E. M. Wright, "How to construct "normal ranges" for fetal variables," *Ultrasound in Obstetrics and Gynecology*, vol. 11, no. 1, pp. 30–38, 1998.

[24] D. Paladini, M. Rustico, E. Viora et al., "Fetal size charts for the Italian population. Normative curves of head, abdomen and long bones," *Prenatal Diagnosis*, vol. 25, no. 6, pp. 456–464, 2005.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 623078, 8 pages
http://dx.doi.org/10.1155/2014/623078

Research Article

Large-Scale Investigation of Human TF-miRNA Relations Based on Coexpression Profiles

Chia-Hung Chien,[1] Yi-Fan Chiang-Hsieh,[1] Ann-Ping Tsou,[2] Shun-Long Weng,[3,4,5] Wen-Chi Chang,[1] and Hsien-Da Huang[6,7,8,9]

[1] *Institute of Tropical Plant Sciences, National Cheng Kung University, Tainan 701, Taiwan*
[2] *Department of Biotechnology and Laboratory Science in Medicine, National Yang-Ming University, Taipei 112, Taiwan*
[3] *Mackay Medicine, Nursing and Management College, Taipei 112, Taiwan*
[4] *Department of Medicine, Mackay Medical College, New Taipei City 252, Taiwan*
[5] *Department of Obstetrics and Gynecology, Hsinchu Mackay Memorial Hospital, Hsinchu 300, Taiwan*
[6] *Institute of Bioinformatics and Systems Biology, National Chiao Tung University, Hsinchu 300, Taiwan*
[7] *Department of Biological Science and Technology, National Chiao Tung University, Hsinchu 300, Taiwan*
[8] *Department of Biomedical Science and Environmental Biology, Kaohsiung Medical University, Kaohsiung 807, Taiwan*
[9] *Institute of Biomedical Engineering, National Chiao Tung University, Hsinchu 300, Taiwan*

Correspondence should be addressed to Wen-Chi Chang; sarah321@mail.ncku.edu.tw
and Hsien-Da Huang; bryan@mail.nctu.edu.tw

Received 11 March 2014; Revised 2 May 2014; Accepted 18 May 2014; Published 9 June 2014

Academic Editor: Tzong-Yi Lee

Noncoding, endogenous microRNAs (miRNAs) are fairly well known for regulating gene expression rather than protein coding. Dysregulation of miRNA gene, either upregulated or downregulated, may lead to severe diseases or oncogenesis, especially when the miRNA disorder involves significant bioreactions or pathways. Thus, how miRNA genes are transcriptionally regulated has been highlighted as well as target recognition in recent years. In this study, a large-scale investigation of novel *cis*- and *trans*-elements was undertaken to further determine TF-miRNA regulatory relations, which are necessary to unravel the transcriptional regulation of miRNA genes. Based on miRNA and annotated gene expression profiles, the term "coTFBS" was introduced to detect common transcription factors and the corresponding binding sites within the promoter regions of each miRNA and its coexpressed annotated genes. The computational pipeline was successfully established to filter redundancy due to short sequence motifs for TFBS pattern search. Eventually, we identified more convinced TF-miRNA regulatory relations for 225 human miRNAs. This valuable information is helpful in understanding miRNA functions and provides knowledge to evaluate the therapeutic potential in clinical research. Once most expression profiles of miRNAs in the latest database are completed, TF candidates of more miRNAs can be explored by this filtering approach in the future.

1. Introduction

Among functional noncoding RNAs (ncRNAs), microRNAs (miRNAs) are tiny molecules (~21–23 nt) with giant roles. miRNAs participate in gene regulation by targeting messenger RNAs (mRNAs) and influencing their stability and the initiation of translation. It is implied that if the expression of miRNA is aberrant, miRNA-mediated gene circuitries will be disordered, resulting in homeostatic imbalance, pathogenesis, and oncogenesis [1, 2]. In recent years, elucidating transcriptional regulatory mechanisms of miRNA genes has been highlighted when studying miRNA function. Core promoters of miRNA genes were promptly identified for depicting full-length primary transcripts [3–5]. High-throughput sequencing datasets derived from epigenetic signature and TSS-relevant experiments unfold transcriptional start sites (TSSs) of miRNAs and offer a practical strategy to determine miRNA promoters [6–9].

To further understand upstream regulatory elements controlling miRNA expression, transcription factors (TFs)

and their binding sites of miRNA promoters were deciphered by either literature survey or computational prediction [10–13]. By integrating the information of TF-miRNA regulatory relations and miRNA target interactions, regulatory networks that revolved around miRNAs provide a biological insight into how miRNAs dominate functional processes in biochemical reactions or metabolic pathways [14–17]. However, most of the foregoing studies regarded the upstream regions of pre-miRNAs as promoters (e.g., 10 kb upstream or −900~+100 of the 5′-start of the pre-miRNA). More convinced miRNA promoters should be used for searching putative TF-binding sites. In addition, the experimentally verified TF-miRNA relations are insufficient for current miRNAs. According to the statistics in the latest version 1.2 of TransmiR, only 735 entries, which include ~201 transcriptional factors, ~209 miRNAs, and 16 organisms from 268 publications, were curated. A large-scale investigation of novel *cis*- and *trans*-elements is necessary to fulfill the unmet need for locating transcription factor binding sites (TFBSs) within miRNA promoter regions.

Since sequence-specific TFs possess DNA-binding domains (DBDs) to recognize specific motifs in miRNA promoter sequences, potential binding sites can be detected by sequence-based computational approaches, for example, position weight matrix (PWM). A position weight matrix (also called position specific scoring matrix, PSSM) infers a pattern of DNA segment and is widely applied in searching TFBSs [18–21]. Two well-known databases collecting matrix information are TRANSFAC [22] and JASPAR [23]. To avoid excess false positives due to short sequence motifs used for TFBS pattern search, *cis*-regulatory analysis of coregulated gene sets was executed to determine overrepresented TFBSs [24, 25]. Although previously published web server Pscan [26] and oPOSSUM [27] can search transcription factor binding sites in coexpressed gene promoters to remit the impact of false positive issue, users have to define their coexpressed gene groups with valid gene IDs. In addition, miRNA promoter sequences are not included in these two web servers.

Therefore, the purpose of this study is to create a computational pipeline to undertake large-scale investigation of novel *cis*- and *trans*-elements for human miRNA genes based on coexpression strategy. First, we constructed 255 coexpressed gene groups of human miRNAs. Moreover, instead of grapping the upstream regions of pre-miRNAs as promoter sequences, we exploited more concrete human miRNA promoters by following the processes as previously described [7]. Through the detection of transcription factor binding sites within promoter regions of human miRNAs and their coexpressed genes by using matrix information from TRANSFAC, the common TFBSs were identified. We then filtered the redundancy by not only the occurrence of common TFBSs but also the expression correlation between TF-encoded genes and corresponding human miRNA gene groups. Finally, more reliable TF-miRNA regulatory relations of 225 human miRNAs were provided. Furthermore, the liver-specific hsa-miR-122 was also selected as the case study to demonstrate the usage of this filtering approach and its practicability.

2. Materials and Methods

2.1. Human miRNA Promoter Sequences. The latest genomic coordinates of human miRNAs were retrieved from miRBase release 19 [28]. Human miRNA TSSs were identified by reproducing the computational procedures described in miRStart resource [7]. The formula which determines tag-enriched loci was modified by weighted tag density (tag density multiply by tag number) in the SVM step to precisely reveal the effect of tag intensity. Then, 1 kb upstream sequences of human miRNA TSSs were acquired from UCSC Genome Browser [29] (hg19/GRCh37 assembly) in FASTA format. The updated intergenic miRNA TSSs in human genome are listed in supplementary Table S1 (available online at http://dx.doi.org/10.1155/2014/623078) in additional file for reference.

2.2. Expression Profiles of Human miRNAs and Annotated Genes. In order to determine human genes that are coexpressed with specific miRNA genes, GDS596 record, the Affymetrix gene expression profiles from 79 physiologically human normal tissues, was downloaded from Gene Expression Omnibus (GEO) [30, 31]. Among annotated human genes in GDS596, genes were filtered out if their HGNC symbols are invalid or have been withdrawn. On the other hand, the expression data of 345 miRNAs in 40 normal human tissues generated by a new type of real time reverse transcription- (RT-) PCR-based miRNA assays were also collected [32]. Mature miRNAs with eliminated miRBase IDs (release 19) were discarded. In both expression datasets, only 17 tissues (adrenal, brain, heart, kidney, liver, lung, lymph node, ovary, pancreas, placenta, prostate, skeletal muscle, testicle, trachea, thymus, thyroid, and uterus) are in common and were considered to be integrated expression profiles of annotated genes and miRNA genes with 17 conditions. To standardize expression levels across extensive range of both datasets, the raw intensity was Z-score transformed [33].

2.3. Coexpressed Gene Groups of Human miRNAs. After the transformation process between two expression datasets, Pearson's correlation coefficient (PCC) was calculated to estimate which annotated genes are coexpressed with specific miRNAs. The formula of Pearson's correlation coefficient (usually using the letter r) is as follows:

$$r_{xy} = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2} \sqrt{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2}}, \quad (1)$$

where X_i and Y_i denote the expression level of i condition (tissue) of two genes x and y that are calculated for r, whereas $\overline{X}$ and $\overline{Y}$ represent the average of corresponding expression levels in total n conditions. Since the value of Pearson's correlation coefficient ranges from +1 (positively correlated) to −1 (negatively correlated), we defined that two genes of interest are coexpressed if their Pearson's correlation coefficient is more than 0.8. Subsequently, coexpressed genes of human miRNAs were determined, and their promoter

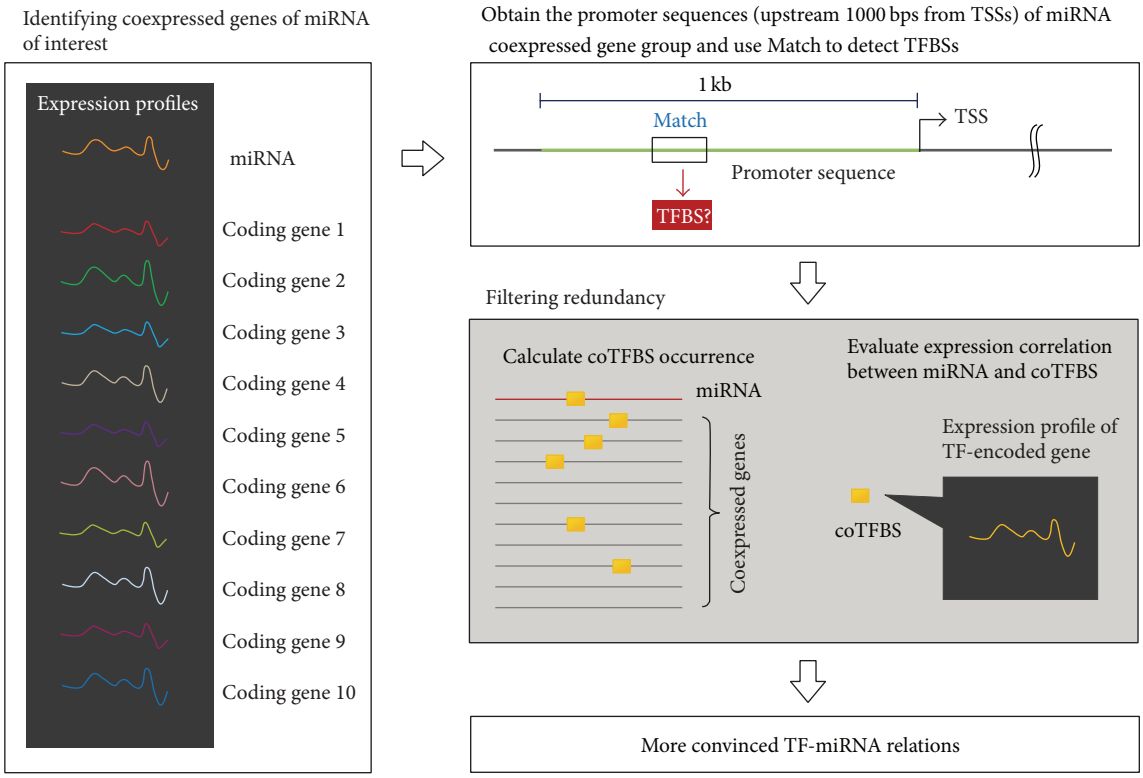


Figure 1: The summary of detecting TF-miR regulatory relations in human genome.

sequences (1 kb upstream from TSS) were obtained from BioMart of Ensembl release 69 [34].

2.4. CoTFBSs in Human miRNA Promoters. In this work, we defined *coTFBS* as a common TFBS located in promoter regions of coexpressed genes. The TRANSFAC database [22], comprising data on transcription factors, their target genes, and regulatory binding sites, has been widely used when studying eukaryotic transcriptional regulation. After acquiring 1 kb promoter sequences of miRNA genes and their coexpressed genes in FASTA format, the Match program [35] was executed for sequence motif search of transcription factor binding sites according to the matrix information provided by TRANSFAC (version 2011.4). A customized profile which specified human matrices in TRANSFAC library was used for Match to minimize both false positive and false negative rates with core similarity and matrix similarity cutoff values for each matrix. As defined in the publication of Match, the core of each matrix refers to the first five most conserved consecutive positions of a matrix. A putative TFBS with core similarity less than 1 was filtered out. The occurrence of each coTFBS and its expression correlation with miRNA in coexpressed gene groups was calculated finally. Only those whose encoded genes are coexpressed with corresponding miRNA were considered.

3. Results

3.1. Human miRNAs with Coexpressed Gene Group for Detecting TF-miRNA Relations. Figure 1 summarizes the workflow to investigate human TF-miRNA regulatory relations based on expression profiles of miRNA and annotated genes. Pearson's correlation coefficient (PCC) was applied to measure the similarity of expression patterns across 17 human normal tissues, which represents the coexpressed level between a specific miRNA gene and an annotated gene of interest. Due to the reason that Pearson's r cannot be calculated if the expression levels in all 17 conditions are identical, 29 mature miRNAs with such expression profiles were excluded. After discarding the mature miRNAs whose miRBase IDs are eliminated, 289 human miRNAs were selected to estimate PCC values with annotated human genes in GDS596 record.

Among them, however, only 255 human miRNAs have more than one coexpressed genes (PCC > 0.8). Moreover, 25 out of 255 human miRNAs have no identified TSSs by following the previous strategy [7] and have to be filtered. In total, 230 human miRNAs were qualified to discover putative *cis*- and *trans*-elements in their promoter regions. The number of members of each miRNA coexpressed gene group ranges from several to thousand (see Table S2 in additional file for details).

3.2. Putative TF-miRNA Relations Were Explored according to the Occurrence of CoTFBS. Theoretically, a group of genes that are coexpressed may be regulated by common transcription factors. Based on this concept, a specific transcription factor binding site located in the promoter regions of most genes in each coexpressed group implies that its corresponding TF is the most possible one controlling the expression of these genes. Here, we introduce the term "coTFBS" which

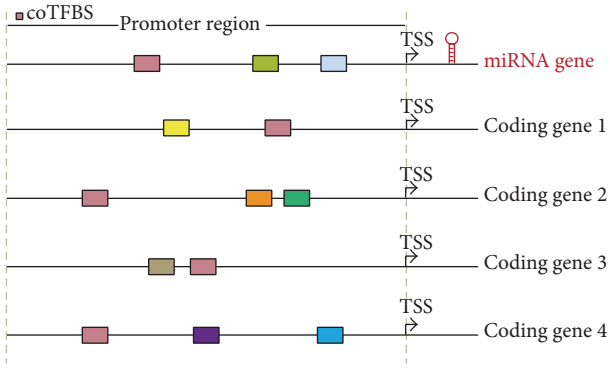


Figure 2: The concept of coTFBS. Protein-coding genes 1, 2, 3, and 4 are coexpressed with the miRNA gene. Colored rectangles represent transcription factor binding motifs in each promoter region. The common TFBS (pink) located within the promoters of miRNA gene and other four genes represents the "coTFBS" of this gene group.

represents the common TFBS of a coexpressed gene group to filter redundant TF candidate of miRNA genes. For example, the "pink rectangle" binding motifs were detected in the proposed miRNA and other four gene promoters in Figure 2. The occurrence of this coTFBS is five. By scanning TFBSs in promoter regions of each miRNA gene and its coexpressed gene group using Match based on TRANSFAC library (version 2011.4) and following the filtering process, the putative TFs that regulate a specific miRNA were determined. In this work, we successfully identified putative TF-miRNA relations of 225 human miRNAs. The full list of putative TFs for each human miRNA promoter can be accessed in Table S3.

3.3. Case Study: hsa-miR-122. hsa-miR-122 is one of the intergenic miRNAs whose TSS and promoter have been experimentally characterized [7]. Previous studies reported that this liver-specific miRNA is significantly downregulated in hepatocellular carcinoma and profoundly affects carcinogenesis [36]. Because of the explicit promoter and biological importance, hsa-miR-122 was selected as the case study to investigate which TFs may regulate its gene expression. 261 annotated genes coexpressed with miR-122 were identified according to their expression levels with Pearson's correlation coefficient (PCC) more than 0.8. Figure 3 compares the expression patterns of 261 coexpressed genes (the pink line represents the average value of their expression) with hsa-miR-122 among 17 human normal tissues, indicating that remarkable peaks appear in liver for all the coexpressed genes. The expression image (see Figure S1 in additional file) also reveals the similar trends between hsa-miR-122 and its coexpressed genes.

Then, the promoter sequences (1 kb upstream from TSS) of hsa-miR-122 gene and 261 coexpressed genes were collected to identify coTFBSs using Match. The occurrence of putative transcription factors of miR-122 is listed in Table 1. Among TF candidates regulating hsa-miR-122, the TF binding motif of HNF-4alpha can be found in 191 coexpressed gene promoters. In 2011, Li et al. reported that HNF-4alpha is a key regulator positively controlling the expression of miR-122 in liver [37], proving our computational finding. They performed not only the luciferase reporter gene assay to detect the *trans*-activation effect of HNF-4alpha in miR-122 promoter but also the ChIP and EMSA assays to determine HNF-4alpha binding of miR-122 promoter in vitro and in vivo. Moreover, other liver-enriched transcription factors including HNF-1alpha, HNF-3alpha, HNF-3beta, and HNF-6 showed a strong positive correlation with miR-122. The knockout of HNF1A, FOXA1, and FOXA2 by RNAi assay reduces the expression of miR-122, suggesting that these transcription factors may bind to miR-122 promoter and transcriptionally regulate miR-122 [38]. Importantly, HNF-1alpha, HNF-3alpha, and HNF-3beta were identified in our list of TF candidates, and a HNF-6 binding site was determined (−2720 from miR-122 TSS) if the 3 kb promoter sequence of miR-122 was used.

In addition, the TF binding site of NR1B2 can be found in 226 coexpressed gene promoters. A previous research article published in 1987 indicated that the inappropriate expression of HAP gene (the official HGNC symbol is RARB) may relate to the hepatocellular carcinogenesis, and hap protein may directly participate in the hepatocellular transformation [39]. It is implied that transcription factor NR1B2 may bind to miR-122 promoter and regulate its expression.

4. Discussion

For large-scale investigation of human TF-miRNA relations in this study, the expression profiles of human miRNAs and over ten thousands genes from normal human tissues facilitate the acquirement of miRNA coexpressed gene groups. However, the latest usable RT-qPCR miRNA expression data for human normal tissues were published in 2007 [32]. The currently available data are almost derived from cancer cell lines or tumor tissues. Totally 230 coexpressed gene groups limit the coTFBS analysis of most miRNAs in existence. Besides, based on the cut-off 0.8 PCC value applied to define coexpression between a miRNA and an annotated gene, 34 miRNAs have no coexpressed genes and 50 miRNAs have less than ten coexpressed genes. Although each miRNA has sufficient coexpressed genes to analyze its putative *cis*- and *trans*-elements if the lower PCC cut-off was used, it may cause the trade-off of specificity when determining miRNA coexpressed genes. For example, hsa-let-7a-1 has no coexpressed gene when using 0.8 PCC value but has 29 coexpressed genes if 0.6 PCC value was applied.

It is noteworthy that the normalization process between two raw expression datasets was only Z-score transformed, without using $\log_{10}$ transformation before Z-score standardization. According to the definition of formula, Pearson's correlation coefficient calculates linear correlation between two variables and has an invariant property in statistics. In fact, the $\log_{10}$ transformation tends to alter the original expression patterns and sharply reduces the scale of raw intensity, resulting in unexpected affection of PCC values. Figure S2 illustrates an example of hsa-miR-122 expression patterns with and without $\log_{10}$ transformation. In addition

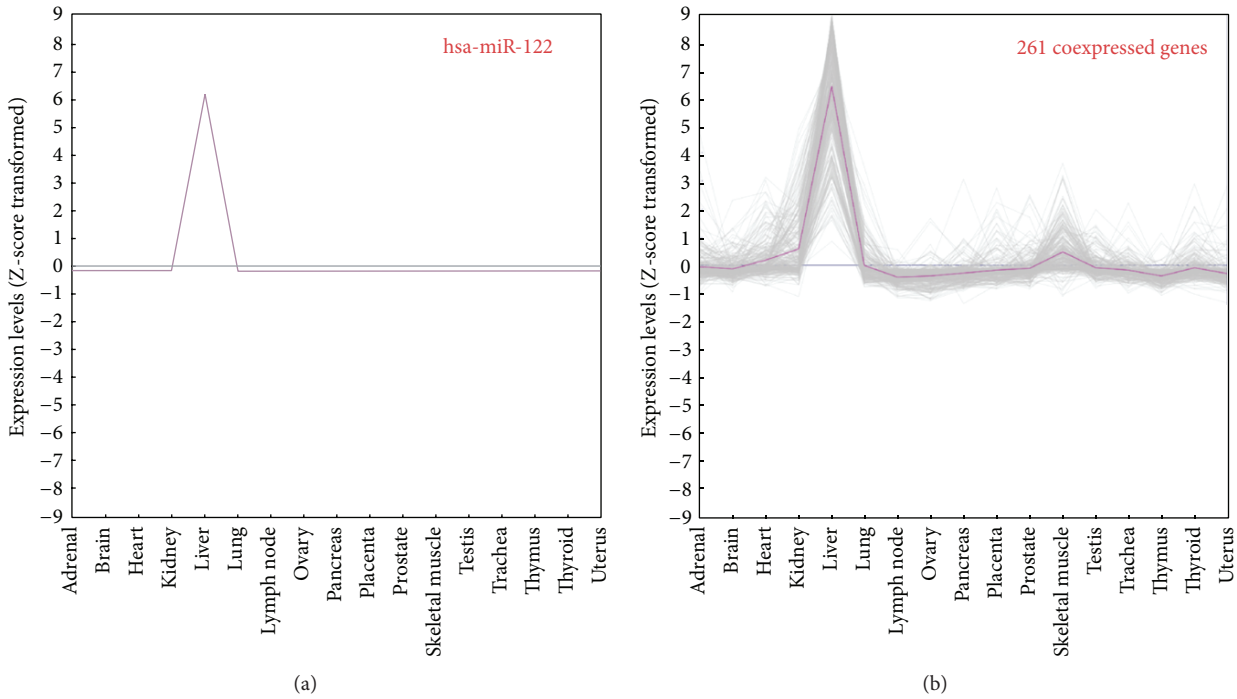


FIGURE 3: 261 human protein-coding genes are coexpressed with hsa-miR-122. The pink curve in 3(b) represents the average value of miR-122 coexpressed genes among 17 human normal tissues. The remarkable peaks appear in liver.

to the high expression level in liver, two obvious peaks of brain and thymus appear in the $\log_{10}$ transformed expression profile of hsa-miR-122. The members of hsa-miR-122 coexpression group also reflect the variation. Unlike more than two hundred coexpression genes by using Z-score transformed expression data, merely five coexpressed genes left based on the cut-off 0.8 PCC value by Z-score and $\log_{10}$ transformed data.

Another limitation of TFBS/TF detection depends on present transcription factor binding matrices collected in TRANSFAC. As is well known, sequence motif search of transcription factor binding sites is executed by the Match program according to the matrix information in TRANSFAC library. A specific TFBS in miRNA promoter will not be detected if the corresponding matrix has not been obtained by in vitro selection studies. Here is an example. Unlike hsa-miR-122, miR-224 is upregulated in HCC through epigenetic mechanisms and controls several crucial cellular processes [40]. Wang et al. indicated that miR-224 expression is reciprocally regulated by HDAC1, HDAC3, and EP300. Our result shows that P300 (encoded by EP300) is the TF candidate of miR-224, and its TFBS can be found in 227 coexpressed gene promoters. Because matrices of histone deacetylases 1 and 3 are not available in TRANSFAC, no such transcription factors were predicted.

In conclusion, rather than traditional PWM search characterizing putative *cis*-elements in promoter regions, the coTFBS strategy was developed to determine more confident TFBSs for human miRNAs. The investigation was restricted by the incomplete expression profiles of present human miRNAs. Once most expression profiles of miRNAs in the latest database are available, TF candidates of more miRNAs can be explored in the future. Furthermore, although more and more ChIP-seq data were generated and were useful to identify transcription factor binding sites [41–43], the corresponding ChIP-seq data of specific TFs are still the minority. It is expected that large-scale ChIP analysis of general TFs contributes more confident TFBSs by observing the aggregated peaks within promoter regions of human miRNAs.

5. Conclusions

In organisms, not all of ribonucleic acids (RNAs) are translated to proteins. miRNAs are such noncoding RNAs which play critical roles in gene regulation, even if it is generally believed that proteins convey vital information from genes and execute biological functions to maintain life processes. Although target prediction has been the mainstream when studying miRNA functions for a while, researchers start to explore TF-miRNA interactions and study the transcriptional regulation of miRNAs, which are necessary to depict how miRNAs participate in diverse biological processes. To determine putative TFs and TFBSs located in human miRNA promoters, we created a computational pipeline which not only allows large-scale investigation as long as the expression profiles of miRNAs are available, but also filters the redundancy when searching short sequence. This valuable information is helpful in understanding miRNA functions and provides knowledge to evaluate the therapeutic potential in clinical research.

Table 1: Putative transcription factors regulating miR-122 gene. The bold items indicate the experimental-supported TFs. PCC represents the Pearson's correlation coefficient between each TF-encoded gene and miR-122.

Matrix_ID	Transcription factor	Gene	PCC	Occurrence
V$ELF1_Q6	Elf-1	ELF1	0.192827	242
V$MAFB_01	MAFB	MAFB	0.383476	239
V$TBX5_02	TBX5	TBX5	0.132419	227
V$NR1B2_Q6	**NR1B2**	**RARB**	**0.117271**	**226**
V$CDX2_Q5_02	CDX-2	CDX2	0.105817	215
V$GATA1_01	GATA-1	GATA1	0.0128229	215
V$SMAD3_Q6_01	Smad3	SMAD3	0.0915205	203
V$HNF4A_Q6_01	**HNF-4alpha**	**HNF4A**	**0.316168**	**191**
V$SOX5_01	SOX5	SOX5	0.149287	185
V$SPI1_03	SPI1	SPI1	0.268782	178
V$ERBETA_Q5	ER-beta	ESR2	0.303769	176
V$GATA1_02	GATA-1	GATA1	0.0128229	173
V$GATA1_05	GATA-1	GATA1	0.0128229	168
V$GATA1_06	GATA-1	GATA1	0.0128229	168
V$CDX2_Q5_01	Cdx-2	CDX2	0.105817	167
V$MAZ_Q6	MAZ	MAZ	0.414832	166
V$MYOD_Q6_01	MyoD	MYOD1	0.306424	164
V$PITX3_Q2	PITX3	PITX3	0.127413	160
V$ING4_01	ING4	ING4	0.234372	158
V$HNF3B_Q6	**HNF-3beta**	**FOXA2**	**0.44036**	**154**
V$CDX2_01	Cdx-2	CDX2	0.105817	152
V$CRX_Q4	Crx	CRX	0.243011	149
V$HNF3A_01	**HNF3A**	**FOXA1**	**0.140429**	**148**
V$GATA1_04	GATA-1	GATA1	0.0128229	139
V$ERR1_Q3	ERR1	ESRRA	0.18197	135
V$CEBPE_Q6	CEBPE	CEBPE	0.0949125	133
V$HNF1_02	**HNF-1alpha**	**HNF1A**	**0.363153**	**122**
V$CRX_02	Crx	CRX	0.243011	119
V$NEUROD_02	NeuroD	NEUROD1	0.0402891	114
V$MAZ_Q6_01	MAZ	MAZ	0.414832	110
V$OC2_Q3	OC-2	ONECUT2	0.124899	102
V$IRF7_Q3	IRF-7	IRF7	0.13082	100
V$PIT1_Q6	Pit-1	POU1F1	0.319886	99
V$DBP_Q6_01	DBP	DBP	0.043001	76
V$CEBPG_Q6_01	C/EBPgamma	CEBPG	0.027117	73
V$MYOD_01	MyoD	MYOD1	0.306424	72
V$CREL_01	c-Rel	REL	0.29444	44
V$CEBPG_Q6	C/EBPgamma	CEBPG	0.027117	43
V$ATF4_Q6	ATF-4	ATF4	0.0251237	40
V$HOXA7_01	HOXA7	HOXA7	0.397453	31
V$E2F1_Q4	E2F-1	E2F1	0.110614	29
V$ATF5_01	ATF5	ATF5	0.912782	26

Conflict of Interests

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

Authors' Contribution

Chia-Hung Chien, Shun-Long Weng, Wen-Chi Chang, Ann-Ping Tsou, and Hsien-Da Huang conceived and designed the experiments. Chia-Hung Chien and Yi-Fan Chiang-Hsieh analyzed the data and performed the experiments. Chia-Hung Chien wrote the paper.

Acknowledgments

Thanks are due to Pei-Wen Jiang for the assistance of programming and system maintenance and to Hsi-Yuan Huang for experience sharing in expression profile analysis. This research was supported by a grant from National Science Council of the Republic of China for financially supporting this research under Contract nos. NSC 102-2313-B-006-004, NSC 101-2311-B-009-003-MY3, NSC 100-2627-B-009-002, NSC 102-2911-I-009-101, and NSC 99-2628-B-006-016-MY3.

References

[1] S. M. Khoshnaw, A. R. Green, D. G. Powe, and I. O. Ellis, "MicroRNA involvement in the pathogenesis and management of breast cancer," *Journal of Clinical Pathology*, vol. 62, no. 5, pp. 422–428, 2009.

[2] R. Schickel, B. Boyerinas, S.-M. Park, and M. E. Peter, "MicroRNAs: key players in the immune system, differentiation, tumorigenesis and cell death," *Oncogene*, vol. 27, no. 45, pp. 5959–5974, 2008.

[3] H. K. Saini, S. Griffiths-Jones, and A. J. Enright, "Genomic analysis of human microRNA transcripts," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 45, pp. 17719–17724, 2007.

[4] H. K. Saini, A. J. Enright, and S. Griffiths-Jones, "Annotation of mammalian primary microRNAs," *BMC Genomics*, vol. 9, article 564, 2008.

[5] J. Qian, Z. Zhang, J. Liang et al., "The full-length transcripts and promoter analysis of intergenic microRNAs in *Drosophila melanogaster*," *Genomics*, vol. 97, no. 5, pp. 294–303, 2011.

[6] A. Barski, R. Jothi, S. Cuddapah et al., "Chromatin poises miRNA- and protein-coding genes for expression," *Genome Research*, vol. 19, no. 10, pp. 1742–1751, 2009.

[7] C.-H. Chien, Y.-M. Sun, W.-C. Chang et al., "Identifying transcriptional start sites of human microRNAs based on high-throughput sequencing data," *Nucleic Acids Research*, vol. 39, no. 21, pp. 9345–9356, 2011.

[8] M. Bhattacharyya, L. Feuerbach, T. Bhadra, T. Lengauer, and S. Bandyopadhyay, "MicroRNA transcription start site prediction with multi-objective feature selection," *Statistical Applications in Genetics and Molecular Biology*, vol. 11, no. 1, article 6, 2012.

[9] Y. Saito, H. Suzuki, T. Taya et al., "Development of a novel microRNA promoter microarray for ChIP-on-chip assay to identify epigenetically regulated microRNAs," *Biochemical and Biophysical Research Communications*, vol. 426, no. 1, pp. 33–37, 2012.

[10] J. Wang, M. Lu, C. Qiu, and Q. Cui, "TransmiR: a transcription factor microRNA regulation database," *Nucleic Acids Research*, vol. 38, supplement 1, pp. D119–D122, 2009.

[11] R. Shalgi, D. Lieber, M. Oren, and Y. Pilpel, "Global and local architecture of the mammalian microRNA-transcription factor regulatory network," *PLoS Computational Biology*, vol. 3, no. 7, article e131, 2007.

[12] C.-Y. Chen, S.-T. Chen, C.-S. Fuh, H.-F. Juan, and H.-C. Huang, "Coregulation of transcription factors and microRNAs in human transcriptional regulatory network," *BMC Bioinformatics*, vol. 12, supplement 1, article S41, 2011.

[13] A. Re, D. Core, D. Taverna, and M. Caselle, "Genome-wide survey of microRNA-transcription factor feed-forward regulatory circuits in human," *Molecular BioSystems*, vol. 5, no. 8, pp. 854–867, 2009.

[14] O. Friard, A. Re, D. Taverna, M. de Bortoli, and D. Corá, "CircuitsDB: a database of mixed microRNA/transcription factor feed-forward regulatory circuits in human and mouse," *BMC Bioinformatics*, vol. 11, article 435, 2010.

[15] A. le Bechec, E. Portales-Casamar, G. Vetter et al., "MIR@NT@N: a framework integrating transcription factors, microRNAs and their targets to identify sub-network motifs in a meta-regulation network model," *BMC Bioinformatics*, vol. 12, article 67, 2011.

[16] H. Naeem, R. Kuffner, and R. Zimmer, "MIRTFnet: analysis of miRNA regulated transcription factors," *PLoS ONE*, vol. 6, no. 8, article e22519, 2011.

[17] X. Li, W. Jiang, W. Li et al., "Dissection of human miRNA regulatory influence to subpathway," *Briefings in Bioinformatics*, vol. 13, no. 2, pp. 175–186, 2012.

[18] M. Beckstette, R. Homann, R. Giegerich, and S. Kurtz, "Fast index based algorithms and software for matching position specific scoring matrices," *BMC Bioinformatics*, vol. 7, article 389, 2006.

[19] M. C. Frith, J. L. Spouge, U. Hansen, and Z. Weng, "Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences," *Nucleic Acids Research*, vol. 30, no. 14, pp. 3214–3224, 2002.

[20] S. Rahmann, T. Muller, and M. Vingron, "On the power of profiles for transcription factor binding site detection," *Statistical Applications in Genetics and Molecular Biology*, vol. 2, article 7, 2003.

[21] Y. M. Oh, J. K. Kim, S. Choi, and J.-Y. Yoo, "Identification of co-occurring transcription factor binding sites from DNA sequence using clustered position weight matrices," *Nucleic Acids Research*, vol. 40, no. 5, article e38, 2012.

[22] 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.

[23] E. Portales-Casamar, S. Thongjuea, A. T. Kwon et al., "JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles," *Nucleic Acids Research*, vol. 38, supplement 1, pp. D105–D110, 2010.

[24] S. Aerts, G. Thijs, B. Coessens, M. Staes, Y. Moreau, and B. de Moor, "Toucan: deciphering the cis-regulatory logic of coregulated genes," *Nucleic Acids Research*, vol. 31, no. 6, pp. 1753–1764, 2003.

[25] T. T. Marstrand, J. Frellsen, I. Moltke et al., "Asap: a framework for over-representation statistics for transcription factor binding sites," *PLoS ONE*, vol. 3, no. 2, article e1623, 2008.

[26] F. Zambelli, G. Pesole, and G. Pavesi, "Pscan: finding over-represented transcription factor binding site motifs in sequences from co-regulated or co-expressed genes," *Nucleic Acids Research*, vol. 37, supplement 2, pp. W247–W252, 2009.

[27] S. J. H. Sui, J. R. Mortimer, D. J. Arenillas et al., "oPOSSUM: identification of over-represented transcription factor binding sites in co-expressed genes," *Nucleic Acids Research*, vol. 33, no. 10, pp. 3154–3164, 2005.

[28] A. Kozomara and S. Griffiths-Jones, "miRBase: integrating microRNA annotation and deep-sequencing data," *Nucleic Acids Research*, vol. 39, supplement 1, pp. D152–D157, 2011.

[29] T. R. Dreszer, D. Karolchik, A. S. Zweig et al., "The UCSC Genome Browser database: extensions and updates 2011," *Nucleic Acids Research*, vol. 40, supplement 1, pp. D918–D923, 2012.

[30] A. I. Su, T. Wiltshire, S. Batalov et al., "A gene atlas of the mouse and human protein-encoding transcriptomes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 16, pp. 6062–6067, 2004.

[31] 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.

[32] Y. Liang, D. Ridzon, L. Wong, and C. Chen, "Characterization of microRNA expression profiles in normal human tissues," *BMC Genomics*, vol. 8, article 166, 2007.

[33] C. Cheadle, M. P. Vawter, W. J. Freed, and K. G. Becker, "Analysis of microarray data using Z score transformation," *The Journal of Molecular Diagnostics*, vol. 5, no. 2, pp. 73–81, 2003.

[34] P. Flicek, I. Ahmed, M. R. Amode et al., "Ensembl 2013," *Nucleic Acids Research*, vol. 41, no. 1, pp. D48–D55, 2013.

[35] A. E. Kel, E. Gossling, I. Reuter, E. Cheremushkin, O. V. Kel-Margoulis, and E. Wingender, "MATCH: a tool for searching transcription factor binding sites in DNA sequences," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3576–3579, 2003.

[36] W.-C. Tsai, S.-D. Hsu, C.-S. Hsu et al., "MicroRNA-122 plays a critical role in liver homeostasis and hepatocarcinogenesis," *The Journal of Clinical Investigation*, vol. 122, no. 8, pp. 2884–2897, 2012.

[37] Z. Y. Li, Y. Xi, W. N. Zhu et al., "Positive regulation of hepatic miR-122 expression by HNF4α," *Journal of Hepatology*, vol. 55, no. 3, pp. 602–611, 2011.

[38] C. Coulouarn, V. M. Factor, J. B. Andersen, M. E. Durkin, and S. S. Thorgeirsson, "Loss of miR-122 expression in liver cancer correlates with suppression of the hepatic phenotype and gain of metastatic properties," *Oncogene*, vol. 28, no. 40, pp. 3526–3536, 2009.

[39] H. de The, A. Marchio, P. Tiollais, and A. Dejean, "A novel steroid thyroid hormone receptor-related gene inappropriately expressed in human hepatocellular carcinoma," *Nature*, vol. 330, no. 6149, pp. 667–670, 1987.

[40] Y. Wang, H. C. Toh, P. Chow et al., "MicroRNA-224 is up-regulated in hepatocellular carcinoma through epigenetic mechanisms," *The FASEB Journal*, vol. 26, no. 7, pp. 3032–3041, 2012.

[41] J. Qin, M. J. Li, P. Wang, M. Q. Zhang, and J. Wang, "ChIP-Array: combinatory analysis of ChIP-seq/chip and microarray gene expression data to discover direct/indirect targets of a transcription factor," *Nucleic Acids Research*, vol. 39, supplement 2, pp. W430–W436, 2011.

[42] M. Hu, J. Yu, J. M. Taylor, A. M. Chinnaiyan, and Z. S. Qin, "On the detection and refinement of transcription factor binding sites using ChIP-Seq data," *Nucleic Acids Research*, vol. 38, no. 7, pp. 2154–2167, 2010.

[43] T. Handstad, M. B. Rye, F. Drablos, and P. Sætrom, "A ChIP-Seq benchmark shows that sequence conservation mainly improves detection of strong transcription factor binding sites," *PLoS ONE*, vol. 6, no. 4, article e18430, 2011.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 736473, 16 pages
http://dx.doi.org/10.1155/2014/736473

Research Article

A *De Novo* Genome Assembly Algorithm for Repeats and Nonrepeats

Shuaibin Lian,[1] Qingyan Li,[1] Zhiming Dai,[1,2] Qian Xiang,[1] and Xianhua Dai[1]

[1] *School of Information Science and Technology, Sun Yat-Sen University, Guangzhou High Education Mega City, No. 132 Waihuan East Road, Panyu District, Guangzhou 510006, China*
[2] *SYSU-CMU Shunde International Joint Research Institute, Shunde 528300, China*

Correspondence should be addressed to Xianhua Dai; issdxh@mail.sysu.edu.cn

Received 19 December 2013; Revised 1 April 2014; Accepted 8 April 2014; Published 25 May 2014

Academic Editor: Li-Ching Wu

Background. Next generation sequencing platforms can generate shorter reads, deeper coverage, and higher throughput than those of the Sanger sequencing. These short reads may be assembled *de novo* before some specific genome analyses. Up to now, the performances of assembling repeats of these current assemblers are very poor. *Results.* To improve this problem, we proposed a new genome assembly algorithm, named SWA, which has four properties: (1) assembling repeats and nonrepeats; (2) adopting a new overlapping extension strategy to extend each seed; (3) adopting sliding window to filter out the sequencing bias; and (4) proposing a compensational mechanism for low coverage datasets. SWA was evaluated and validated in both simulations and real sequencing datasets. The accuracy of assembling repeats and estimating the copy numbers is up to 99% and 100%, respectively. Finally, the extensive comparisons with other eight leading assemblers show that SWA outperformed others in terms of completeness and correctness of assembling repeats and nonrepeats. *Conclusions.* This paper proposed a new *de novo* genome assembly method for resolving complex repeats. SWA not only can detect where repeats or nonrepeats are but also can assemble them completely from NGS data, especially for assembling repeats. This is the advantage over other assemblers.

1. Background

Over the past twenty years, genome sequencing technologies have made great progress in many aspects, such as speed, cost, coverage, and so forth. The automated Sanger sequencing is regarded as the first-generation genome sequencing technology which has the ability to read longer than 1000 base pair (1000–2000 bp). The latter sequencing technologies are referred to as the next-generation sequencing (NGS) technologies. Currently, the available commercial NGS platforms include GA, MiSeq, and HiSeq from Illumina [1], SOLiD and Ion Torrent from Life Technologies [2], RS system from Pacific Bioscience, and Heliscope from Helicos Biosciences [3–7]. Next generation sequencing machines can sequence the whole human genome in a few days, and this capability has inspired a flood of new projects that are aimed at sequencing large kinds of animals and plants [8, 9]. NGS can be characterized by highly parallel operation, higher yield, simpler operation, shorter reads, and much lower cost [10]. However, the NGS technologies all share a common intrinsic characteristic of providing very short read length (30~250 bp), which is substantially shorter than the Sanger sequencing reads.

These short reads may be assembled *de novo* before further genome analysis if the reference genome is not available. Currently, there are tens of genome assembly algorithms and software. Among them ABySS [11], ALLPATHS-LG [12], Bambus2 [13], CABOG [14], MSR-CA (http://www.genome.umd.edu/masurca.html), SGA [15], SOAPdenovo [16], and Velvet [17] are the typical ones. Each of them is able to run large and whole genome assembly using NGS short read data from mate-pair or paired-end information. In terms of repeats smaller than read length, these current assemblers performed well. But for the repeats longer than read length, most of them performed poorly in completeness and accuracy of assembling repeats from NGS data. It is shown that repetitive DNA comprises a significant fraction of the eukaryotic genomes, for example, ~20% of

Caenorhabditis elegans and *Caenorhabditis briggsae* genomes [18] and ~50% of the human genome [19] have been identified as repetitive DNA. Most of these repetitive DNA sequences have some important biomedical functions and are closely related to some complex disease [20, 21], such as cancer [22], neuropsychiatric disorders [23], and autism [24]. Thus, it is necessary to improve the genome assembly algorithms, especially in assembling repeats.

To this end, we proposed a new genome assembly algorithm aiming for assembling repeats and nonrepeats, named SWA (sliding window assembly), which can assemble repeats and nonrepeats completely and accurately. In SWA, sliding window function is used to filter out the sequencing bias caused by sequencing process and improve the confidence of separating repeats and nonrepeats. There are five typical properties.

(1) Assembling repeats and nonrepeats completely and accurately. SWA can assemble repeats and nonrepeats from NGS data directly in a parallel way which can reduce the memory usage and executive time. In addition, SWA cannot only detect where repeats or nonrepeats are but also can assemble them completely. Therefore, SWA provides an alternative solution to resolve long repeats in some extent.

(2) Adopting dynamic overlapping strategy to extend each seed. The so-called dynamic overlapping strategy is to compute the reads overlapped with seed in intervals composed of maximum overlap and minimum overlap. This strategy can search the optimal read for extension in dynamic interval and jump over the short repeats.

(3) Adopting sliding window to filter out the sequencing bias so as to improve the confidence of detecting boundary of repeats. NGS data is always full of sequencing bias and is highly uneven, which caused the difficulty of distinguishing where repeats or nonrepeats are. Sliding window technique is used to filter out the sequencing bias in the genome assembly process so as to increase the confidence of detecting boundary of repeats.

(4) Proposing a compensational mechanism for the loss caused by low coverage. Low coverage makes the statistical properties of read counts less significant. To improve the statistical significance, SWA proposed a compensational mechanism based on sliding window. This mechanism can improve the statistical significance of read counts under the condition of low coverage.

(5) Estimating copies of assembled ones as an auxiliary function.

The main contributions of our approach are as follows. (1) Assembling repeats and nonrepeats completely and accurately rather than only detecting where repeats or nonrepeats are. Complex repeats structures have very important biomedical functions. Consequently, the completeness and accuracy of assembling repeats are what SWA is mainly concerned about the completeness of assembled repeats and nonrepeats rather than the continuity of whole genome assembly. (2) Sliding window functions to filter out the sequencing bias are used in genome assembling process. Filtering noise by window function is very common in information processing but is rare in genome assembly process. SWA adopts sliding window to filter out NGS data bias and improve the statistical significance of read counts. In addition, a compensational mechanism based on sliding window was embedded in SWA. This mechanism can improve the significance of read counts under the condition of low coverage.

The assessments were performed in simulated datasets and real NGS datasets which are all generated by Illumina sequencer. Simulated study is only used to validate the performances of SWA. Therefore, it has little meaning to compare with other assemblers in simulated datasets. Extensive comparisons were conducted with other eight famous assemblers, such as ABySS, ALLPATHS-LG, Bambus2, CABOG, MSR-CA, SGA, SOAPdenovo, and Velvet, in real NGS datasets. The results indicate that for whatever small genomes or large genomes, SWA outperformed other eight leading assemblers in the completeness of assembling repeats. SWA is freely available at http://222.200.182.71/swa/SWA.rar.

2. Results

2.1. SWA Algorithm. SWA runs in five key steps (Figures 1 and 2): preprocessing, unique processing (Figure 2(b)), hash index (Figure 3), seed selection (Figure 2(c)), and seed extension (Figures 2(d), 4, and 5).

Firstly, preprocessing is performed. SWA firstly filters out the raw reads that contain any "N", which is noninformative, or any low quality value region, which may contain errors and lead to false positive overlaps with other reads (Figure 2(a)). And then the parameters are set in advance (see parameters): the maximum overlapping length max, dynamic overlapping interval L_d, read length L_r, length of sliding window L_w, threshold of repetitive seeds H_p, threshold of nonrepetitive seeds L_p, threshold of repeats assembly T_2, threshold of nonrepeats assembly T_1, and sequencing depth after filtering by sliding window S_{fd}.

Thirdly, unique processing is performed (Figure 2(b)), which is used to find the unique reads and corresponding frequencies in raw datasets. This step is performed using *unique* function provided by MATLAB platform. For example, $C = \text{unique}(A)$ for the array A and returns the same values as in A but with no repetitions. The values of C are in sorted order. For the sequencing reads, this step is performed in a similar way. The reads that are exactly the same—that is identical reads—are collapsed into one unique read and the corresponding frequency is also recorded. For each raw read, the reverse-complement is also used. After unique processing, all unique reads are sorted in a table R, which is a variable to store these processed data. By unique processing, the amount of data is decreased, especially for the high coverage data, which can also reduce the computational requirement and accelerate the running time.

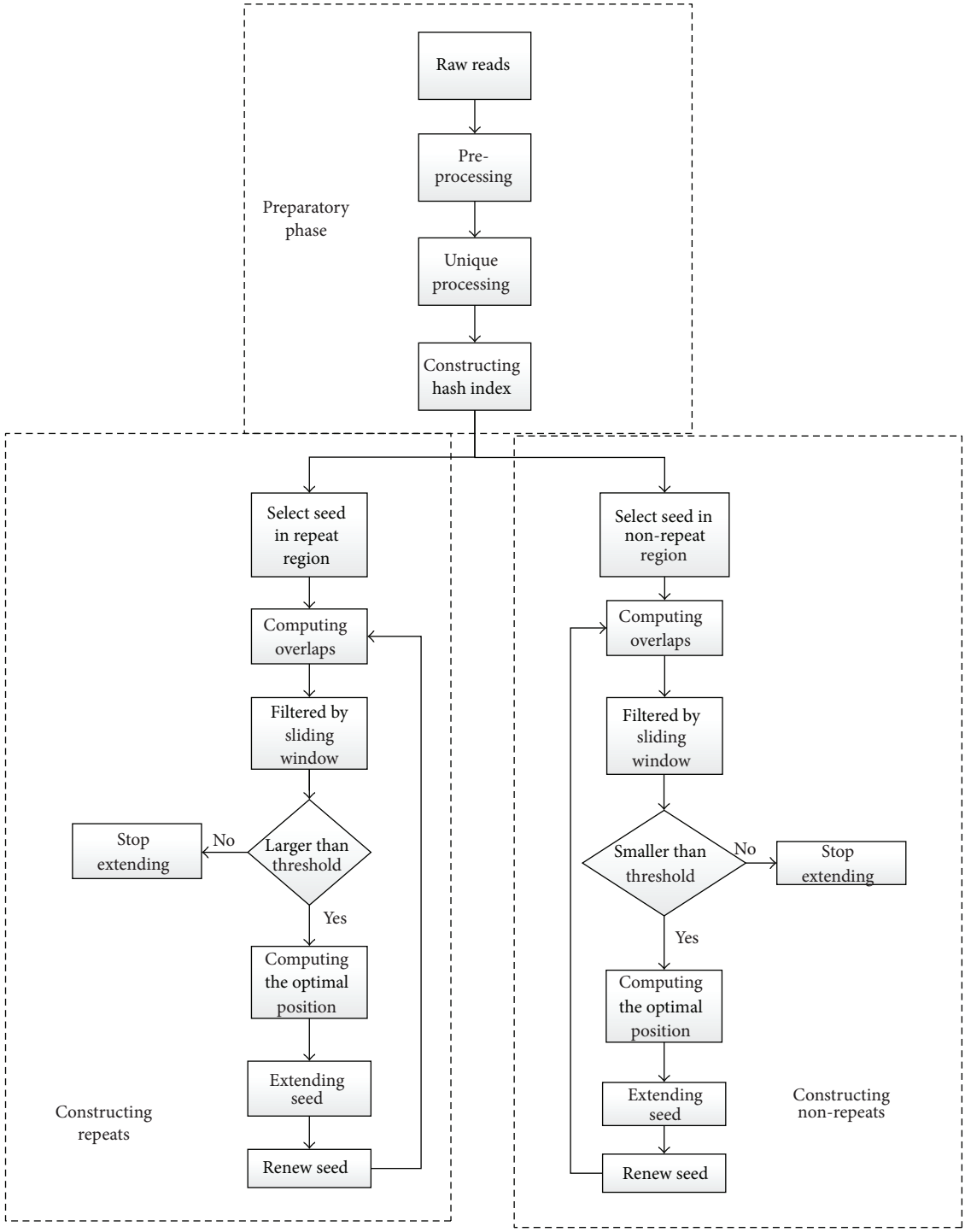


Figure 1: High-level diagram of the SWA assembly pipeline. The assembly has three main modules: preparatory stage, constructing repeats, and constructing nonrepeats. Preparatory stage includes data cleaning, unique processing, and hash index constructing. The stage of constructing repeats and constructing nonrepeats can be performed in parallel or in series. Figure 1 shows the parallel manner. The specific steps of constructing repeats and nonrepeats are detailed in Methods.

Thirdly, hash index is constructed (Figure 3). As it is time consuming to identify unique reads by directly comparing the whole raw reads one by one, SWA constructs the indirect hash index of unique reads which is able to index complex and nonsequential data in a way that facilitates rapid searching. The indirect hash index structures firstly transform the keywords to the quaternary integers and then index these integers rather than the strings directly. This is especially appropriate for DNA sequencing reads.

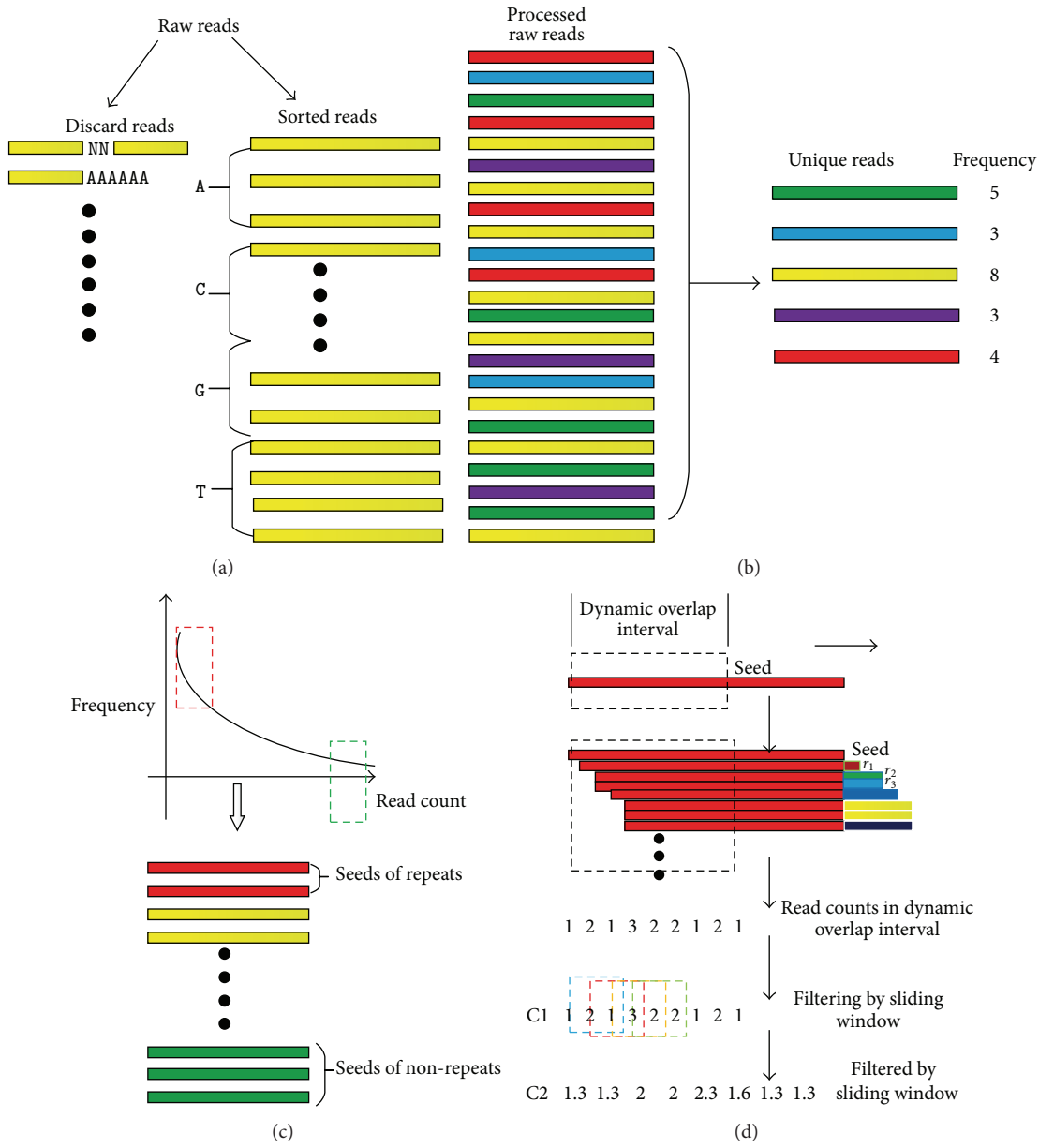


Figure 2: Some key steps of SWA. (a) Raw reads processing. Input reads containing any "N" or a low quality region are discarded and then sorted in alphabetical order. (b) Graphical illustration of unique process. The five different color lines represent the five unique reads in preprocessed raw reads. Each of them appears more than twice. By unique processing, the identical reads are collapsed into one unique and corresponding frequency. (c) Seed selection. The unique reads are ranked by read count (from high to low). Unique reads with read count larger than H_p are selected as seeds for repeat (the red dotted frame), while unique reads with read counts smaller than L_p are selected as seeds for nonrepeat extension (the blue dotted frame). (d) The graphical example of extending repeats using sliding window function in dynamic overlapping interval. The dotted box represents the dynamic overlapping interval L_d. After overlapping with seed in L_d, the overlapped read counts are recorded and then sliding window function is used to filter out the read bias in this interval continuously, as shown in C1. C2 is the corresponding results filtered by sliding window function, and then the mean value of this interval is recorded in variable M_n to detect the boundary of repeats (Figure 4) and nonrepeats (Figure 5). In this extension, SWA regards r_1 as the optimal extendable read. The extension of nonrepeats is performed in a similar way. The detailed extension and boundary detection are graphically shown in Figures 4 and 5.

Fourthly, seed selection is conducted. In SWA, each extension requires a unique read, called a seed, to initiate the extension. In an extension-based assembler, a good seed should not contain any sequencing errors and should not be selected from the boundary of repeats and nonrepeats. Consequently, data cleaning is necessary in the data preprocessing stage before seed selection, and then the seeds are selected in table R. In addition, theoretically, a read from repetitive region usually has a high read count because identical repeats from other loci are counted as well; a read

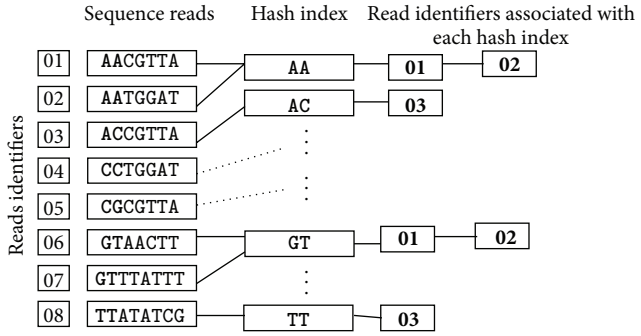


Figure 3: Schematic of a hash index of short sequences strategy. Sequence reads with associated read identifiers are shown, with the regions that will be used for seed selection in capital letters and matched seeds of two bases from AA to TT. Given read identifiers are associated with the seeds using a hash function (e.g., a unique integer representation of each seed). Once such hash table has been built for unique reads; the corresponding data can be scanned with the same hash function, resulting in a much smaller subset of reads to more exactly search the extendible reads.

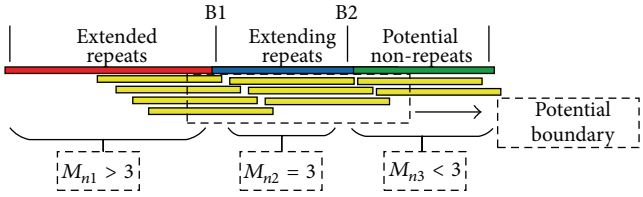


Figure 4: Schematic of extending repeats and boundary detection. The graphic illustration of extending repeats and boundary detection. Red line represents the extended repeats, blue line represents the extending repeats, and the green line represents the potential nonrepeats. The yellow lines represent the supporting reads overlapped with the extended contig. We assume that the sequencing depth $S_d = 2$, and let $T_2 = 3$. Therefore, in the process of extending repeats, the mean value M_n of dynamic overlapping interval filtered by sliding window as shown in Figure 2(d) should be larger than or equal to T_2. The dotted box represents the potential boundary of repeats. Consequently, if we set $M_n > T_2$, the extension will be stopped at B1 or the extension will be stopped at B2.

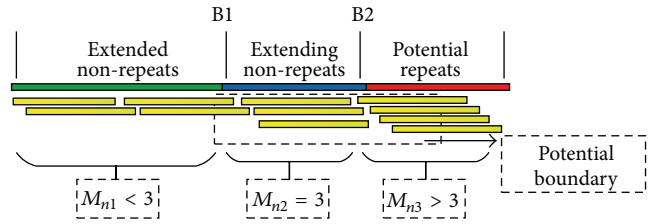


Figure 5: Schematic of extending nonrepeats and boundary detection. The graphic illustration of extending repeats and boundary detection. Green line represents the extended nonrepeats, blue line represents the extending nonrepeats, and the red line represents the potential repeats. The yellow lines represent the supporting reads overlapped with the extension. We assume that the sequencing depth $S_d = 2$, and let $T_1 = 3$. Therefore, in the process of extending nonrepeats, the mean value M_n of dynamic overlapping interval filtered by sliding window should be smaller than or equal to T_1. The dotted box represents the potential boundary of repeats. Consequently, if we set $M_n < T_1$, the extension will be stopped at B1 or the extension will be stopped at B2.

from nonrepetitive region always has a low read count. On the other hand, the read from the boundary always has the middle read count; these seeds always lead to misassembled and short contigs. Thus, the seeds for repeats are chosen with high read count in table R, larger than H_p, while the one for nonrepeats should be chosen with low read count in table R, smaller than L_p. And seeds with middle read count should be avoided. In order to avoid the risk of selecting seed with full errors, the lower limit of read count is also necessary. Furthermore, in seed selection stage, sequencing base quality value will be used to avoid the risk of picking seed with errors, especially for seed of nonrepeats. For the seeds with same read count, SWA selects the one with higher quality value, because higher base quality means lower errors. This strategy can avoid the risk of picking seed with errors to maximum extent.

Fifthly, seed extension (Figures 2(d), 4, and 5). We first define some terms. Let L_r be the length of each read and x and y be two unique reads in table R. We say that x and y overlap if the suffix t bases of x are identical to the prefix t bases of y, where $\min \le t \le \max$ (max and min are, resp., the allowed maximal and minimal numbers of overlapping bases) and the default value of them are $L_r/2 \le \min < \max = L_r - 1$. More explicitly, we say that the $3'$-end of x overlaps with the $5'$-end of y. In the circumstances of long reads, the max overlap max is not needed to use this default value, and the empirical value is $\min < \max < \min\{100, L_r - 1\}$. The dynamic overlapping interval L_d, that is the k-mer in SWA, and is defined as $L_d = k\text{-mer} = \max - \min$. Given a seed, SWA first extends it at the $3'$-end and then at the $5'$-end. A read is extendable for r_{seed} if its $5'$-end overlaps with the $3'$-end of r_{seed} and its $3'$-end overlaps with one or more unique reads. To extend a seed r_{seed} at the $3'$-end, SWA searches all unassembled unique reads in table R for extendable reads in the dynamic range of [1, k-mer], that is, from maximal overlapping to minimal overlapping. The read counts overlapping with seed in this interval are recorded (as shown in Figure 2(d)). Then sliding window function is used to filter out the read count bias of this interval, and then the mean value m_y filtered by sliding window is recorded in variable M_n which is used to detect the boundary of repeats. For the extension of repeats (Figure 4), the extension will continue if $M_n > T_2$, where T_2 is the threshold of repeats, else the extension will be stopped at this end. For the extension of nonrepeats (Figure 5), the extension will continue if $M_n < T_1$, where T_1 is the threshold of nonrepeats. Meanwhile, the mean value of variable M_n can be used to estimate the copy numbers.

In seed extending stage, dynamic overlapping interval L_d is used to search the optimal read to extend seed and which has three functions: (1) the optimal read can be searched in this interval for seed extension (Discussion); (2) the bias of the read counts in this interval can be filtered out by sliding window function so as to increase the confidence of detecting boundary (Figures 4 and 5); (3) copies of assembled contig can be estimated based on the combination of this interval.

Moreover, sliding window is a determinant factor of judging whether the extension of seed is up to the bound of repeats or nonrepeats. The threshold for repeats and nonrepeats is closely related to the sliding window and filtered times, which determines the accuracy and correctness of repeat or nonrepeat contig construction.

The concrete contents and abbreviations are described in detail in methods and parameters.

2.2. Compensational Mechanism. Another property of SWA is the mechanism of compensating the loss caused by low sequencing depth. In practice, the genome sequencing process is highly uneven among the whole genome and full sequencing bias. High coverage can decrease the influence of sequencing bias on the statistical significance of read counts. However, under the condition of low coverage, sequencing bias reduces statistical significance of read counts, which leads to difficulty of distinguishing boundary of repeats. In order to improve the statistical significance of read counts under low coverage, sequencing bias will be filtered out more completely. To this end, SWA proposed a compensational mechanism by the combination of filtered times and sliding window function. The so-called filtered times N_f is the parameter of SWA which means the times of sequencing data filtered by sliding window function. For example, if the read counts filtered by sliding window function is filtered by sliding window once more, that is filtered times $N_f = 2$. The main idea of the compensational mechanism is as follows: SWA can increase the length of dynamic overlapping interval L_d and the size of sliding window L_w and then coalesce several points into one point, which can be achieved by both increasing the size of sliding window and number of filtered times.

The relationship between sequencing depth filtered by sliding window, filtered times N_f, and average sequencing depth S_d is given as follows:

$$S_{\text{fd}} = S_d \times L_w^{N_f}. \quad (1)$$

After filtering by sliding window, read counts will be more flat. To run this compensational mechanism, we suggest that only the values, size of sliding window, and filtered times, may need to be adjusted. The empirical value of optimal sliding window should be set in the range k-mer/5 $\le L_w \le$ k-mer/3, where k-mer is the size of dynamic overlapping interval, and optimal filtered times N_f should be set $N_f = 2$.

Notably, this compensational mechanism is a double edge sword. On the one hand, it can decrease the sequencing bias and compensate the read loss of low coverage. On the other hand, it reduces the sensitiveness of detecting boundary of repeats and increases the computing complexity and executive time. In order to get the best assembly, the high coverage NGS data is also preferred.

2.3. Copy Number Estimation. The auxiliary function of SWA is to estimate the copy numbers of each assembled contig including repeats and nonrepeats directly from NGS data rather than aligning them back to reference genomes. Copy number estimation is also an important factor for the genomic function analysis related to CNVs. SWA estimates the copy number of the assembled contig when its extension is stopped at both ends. SWA output this finished contig and its corresponding copies simultaneously. Furthermore, the accuracy of estimated copy numbers is high up to 99% as shown in assessments in simulated datasets. Because in the process of extension of a seed, the mean value of each extension is recorded in variable M_n, which is used to estimate the copy number of corresponding items (Methods).

3. Assessments

3.1. Metrics. In order to evaluate the ability of SWA for assembling repeats and nonrepeats independently, we use the metrics including Types of Repeat Contigs (TRC), Copy Number of each Repeat Contig (CNRC), Number of Nonrepeat Contigs (NNC), Number of Contig (Number C), N50, N90, Mean Contig, Maximum Contig, CN-accuracy, Rep-accuracy, C-accuracy, E-size, and Genome coverage.

TRC is the types of repeats and CNRC is the copy numbers of each corresponding type of repeat. These two metrics are used to evaluate the correctness of assembled repeats. NNC is the number of nonrepeat contigs which is used to evaluate the correctness of assembled nonrepeats. Because N50 size might sometimes be a misleading statistic, we also computed another statistic, which we called E-size.

(i) CN-accuracy is the accuracy of estimated copy number of each repeat. And which is designed to evaluate the accuracy of estimating copy number of each repeat contig and defined as CN-accuracy $= 1 - (\sum N_{\text{error}}/N_{\text{total}})$, where N_{total} is the sum of the copy numbers of all types of repeats and N_{error} is the absolute value between estimated copy numbers and theoretical copy numbers. Therefore, the larger CN-accuracy is preferred. The larger the CN-accuracy is, the better the performances of SWA of estimating copy numbers of repeats will be.

(ii) Rep-accuracy (Rep-acc) is the accuracy of assembled repeat contigs, which is designed to evaluate the correctness of all types of assembled repeat contigs and defined as Rep-accuracy $= 1-(\sum |L_a - L_r| / \sum L_r)$, where L_r represents the length of real repeat and L_a represents the length of assembled repeat contig. $|L_a - L_r|$ is the error tolerance. The higher Rep-accuracy represents the better performances of SWA for assembling repeat contigs. So higher Rep-accuracy is preferred.

(iii) C-accuracy is the accuracy of the total contigs, which is designed to evaluate the accuracy of all assembled contigs totally and defined as C-accuracy $= 1 - (L_{\text{error}}/L_{\text{total}})$; L_{total} represents the total number of all contigs including repeats and nonrepeats and L_{error} is the sum of all error contigs.

(iv) The E-size is designed to answer the question: if you choose a location (a base) in the reference genome at random, what is the expected size of the contig or scaffold containing that location? This statistic is

Table 1: The performances of assembling different kinds of repeats.

Sequence (Containing)	Repeat		Contigs (kb)			Accuracy (%)			Genome coverage (%)
	TRC	CNRC	NNC	N50	Max	CN-accuracy	Rep-accuracy	C-accuracy	
Interspersed repeats	6	3, 6, 8, 5, 10, 8	73	9.1	26	100	99.9	100	100.9
Tandem repeats	6	6, 18, 9, 4, 20, 9	51	13.2	31.3	100	99.8	100	101
Compound repeats	12	Appendix	110	13.9	32.9	100	99.8	100	101

Three sequences with length $L = 500$ kb, 500 kb, and 1 Mb containing different types of repeats. Contigs of repeat and nonrepeat are generated independently by SWA with basic parameters: read length $L_r = 50$, filtered times = 1, sliding window size $L_w = 3$, and k-mer = 10. Contigs smaller than 200 are removed.

Table 2: The effect of depth to SWA.

Sequencing depth	Repeat		Contigs (kb)			Accuracy (%)			Genome coverage (%)
	TRC	CNRC	NNC	N50	Max	CN-accuracy	Rep-accuracy	C-accuracy	
6	5	3, 4, 2, 6, 5	21	19	48	100	99.9	100	100.4
4	5	3, 4, 6, 5, 2	25	19	48	100	99.9	100	100.4
2	9	Appendix	100	7.2	20.8	100	89.2	100	101.1
1	15	Appendix	323	2.1	11.8	100	86.7	100	101.9
0.5	14	Appendix	950	0.3	3.0	100	66.7	100	63.3
0.2	17	Appendix	54	0.2	1.0	80	56.7	100	8.3

Sequence length $L = 500$ kb containing five types of repeats. Sequencing depths are changing from 6 to 0.2. Contigs of repeat and nonrepeat are generated in an independent ways by SWA at different depths with basic parameters: read length $L_r = 60$, filtered times = 1, sliding window size $L_w = 3$, and k-mer = 10. Contigs smaller than 200 are removed.

one way to answer the related question: how many genes will be completely contained within assembled contigs or scaffolds, rather than split into multiple pieces? E-size is computed as $E = \sum_c (l_c)^2/G$, where l_c is the length of contig C and G is the genome length estimated by the sum of all contig lengths.

For evaluating correctness, the metrics, such as CN-accuracy, Rep-accuracy, and C-accuracy, are all computed by aligning the corresponding items back to the reference genome using program *swalign* from MATLAB platform. *Swalign* constructs local pairwise alignments between two sequences using Smith-Waterman algorithm [25].

Among these metrics, TRC, CNRC, NNC, CN-accuracy, and Rep-accuracy are specially designed for judging the correctness of assembling repeats and nonrepeats. Notably, the length of contigs is not what we are mainly concerned about due to following reasons.

(1) The primary goal of SWA is to resolve repeats by assembling repeats regions and nonrepeats regions separately. Therefore, the correctness of assembling repeats and nonrepeats is what SWA is firstly concerned about.

(2) In order to separate repeat or nonrepeat correctly, SWA must detect the boundary of them accurately and stop extending seed at the boundary automatically.

(3) The assembly with larger contig is always preferred. However, if the contig is assembled or constructed with errors, the larger the contig is, the worse the assembly will be. So accuracy is another important metric for the correct contig.

3.2. Assessments in Simulated Datasets. In this part, we validated the performances of SWA in three kinds of simulated datasets containing interspersed repeats, tandem repeats, and compound repeats, respectively. And then the effect of sequencing depth and read length to SWA was evaluated, respectively. The detailed results were shown in Tables 1, 2, and 3.

From Table 1, three kinds of simulated datasets containing interspersed repeats, tandem repeats, and compound repeats were used to validate the performances of SWA. The repetitive contents contained in these three sequences represented a wide range of repeats with different copies and lengths. The maximum of copy number and length of repetitive sequence are set up to 18 and 5 kb, respectively. The CN-accuracy, C-accuracy, and Rep-accuracy were almost up to 100% and 99.9%, respectively, which indicated that the estimated copy numbers of assembled repeats and the constructed contigs were all absolutely correct, and the error tolerance of constructed repeats was lower than 0.2%. The error rate of genome coverage was up to 1% low. All of these indicate that SWA not only can assemble different kinds of repeats and nonrepeats independently but also can estimate their copy numbers accurately.

From Table 2, we can clearly see that sequencing depth has a great influence on the performances of SWA. When depth = 6 or 4, the performances of SWA were so much better, metrics such as TRC, CNRC, CN-accuracy, and C-accuracy were absolutely correct and high. Rep-accuracy and N50 were a little down but still good; when depth dropped to 2 or 1, TRC and NNC were increasing, while N50, Max, and Rep-accuracy were decreasing; meanwhile the completeness of assembled repeats is getting worse. All of these indicated that the performances of SWA were degenerating; that is,

Table 3: The effect of read length to SWA.

Read length	Repeat		Contigs (kb)			Accuracy (%)			Genome coverage (%)
	TRC	CNRC	NNC	N50	Max	CN-accuracy	Rep-accuracy	C-accuracy	
36	10	Appendix	33	19	34	100	98.2	100	100.1
50	9	Appendix	23	19	48	100	99	100	100.3
101	7	3, 4, 6, 5, 2, 2, 2	21	19.1	48.1	100	99.2	100	100.8
150	6	3, 4, 2, 6, 5, 2	22	19.3	48.3	100	99.6	100	101

Sequence length $L = 500$ kb containing five types of repeats. Read length is changing from 36 to 150. Contigs of repeat and nonrepeat are generated in an independent ways by SWA at different levels with basic parameters: sequencing depth $S_d = 4$, filtered times = 1, sliding window size $L_w = 3$, and k-mer = 10. Contigs smaller than 200 are removed.

Table 4: The effect of sliding window to SWA.

Window size	Repeat		Contigs (kb)			Accuracy (%)			Genome coverage (%)
	TRC	CNRC	NNC	N50	Max	CN-accuracy	Rep-accuracy	C-accuracy	
3	11	Appendix	1016	0.4	3.0	100	76.1	100	81.2
7	10	Appendix	688	1	4.0	100	98.7	100	101
11	15	Appendix	612	1.1	4.0	100	89.6	100	101

Sequence length $L = 500$ kb containing four kinds of repeats. Size of sliding window varies from 3 to 11. Contigs of repeat and nonrepeat are generated in an independent way by SWA at different levels with basic parameters: sequencing depth $S_d = 0.5$, filtered times = 1, read length $L_r = 60$, and k-mer = 20. Contigs smaller than 200 are removed.

long repeats and nonrepeats were assembled into several short fragments. CNRC of corresponding assembled repeat contig was shown in Appendix. But CN-accuracy and C-accuracy were still up to 100% which indicated that the copy number estimation of each assembled repeat contig was still right. When depth fell to 0.5, the metrics except CN-accuracy and C-accuracy were almost not accurate. Particularly, the Rep-accuracy was only 66.7%, which indicated that the completeness of repeats was destroyed. Notably, when depth dropped to 0.2, almost all metrics were getting bad. TRC and NNC were far from the real value. CN-accuracy and Rep-accuracy were so low that almost half repeats were not assembled and their copy numbers were estimated with large errors. N50 and Max were so small which indicated that all long repeats and nonrepeats were broken into smaller fragments. The assembled repeats and nonrepeats were far from completeness. But C-accuracy was more robust than other metrics, which indicated that although these contigs were so small they were at least correct. All of these indicate that sequencing depth affects the performances of SWA greatly. Some extent of high coverage depth is necessary for SWA to generate best assembly.

From Table 3, we can clearly see that read length has little influence on the performances of SWA. When read length varied from 36 to 150, almost all metrics were good and had little change except TRC and NNC. TRC and NNC were decreasing, which indicated that long repeats and nonrepeats were assembled more completely. Therefore, N50 increased a little with the increase of read length. What is more, the Rep-accuracy and genome coverage were increasing a little with the increasing of read length, which indicated the completeness of assembled repeats and assembly redundant were increasing simultaneously.

From Tables 1, 2, and 3, the property of assembling repeats and nonrepeats independently and separately was validated. The effect of sequencing depth and read length to the performances of SWA was also evaluated, respectively. From these results we can safely come to a conclusion that SWA can assemble repeats and nonrepeats independently and correctly; meanwhile sequencing depth has a greater influence on SWA than read length. In high coverage depth, the total performances of SWA are perfectly good. But in a low coverage depth situation, the performances of SWA are a little down. In practice, the higher coverage generated increases the higher sequencing cost. So a compensational mechanism of low sequencing depth is described in the following section.

In the following section, we evaluated the effect of sliding window and filtered times to SWA in low sequencing depth situation, respectively. The detailed results were shown in Tables 4 and 5.

Table 4 indicated that the performances of SWA were not so much good in low sequencing depth and sliding window can improve the performances of some metrics, such as TRC, NNC, and Rep-accuracy. When sliding window size L_w varied from 3 to 11, the performances of SWA were getting from good to bad, then to bad generally. Particularly for the metric of Rep-accuracy, which was getting up to 98.7% from 76.1% and then getting down to 89.6%. So the appropriate sliding window can improve the performances of SWA and compensate the read loss of low coverage depth. The choice of optimal sliding window is very important for the effect of compensational mechanism and is closely related to read length, sequencing depth, and dynamic overlapping interval.

It was clearly shown in Table 5 that the appropriate increase of filtered times could also improve the accuracy

Table 5: The effect of filtered times to SWA.

Filtered times	Repeat		Contigs (kb)			Accuracy (%)			Genome coverage (%)
	TRC	CNRC	NNC	N50	Max	CN-accuracy	Rep-accuracy	C-accuracy	
1	11	Appendix	801	0.3	2.3	100	66.8	100	53.4
2	12	Appendix	418	1.8	7.6	100	82.1	100	101.9
3	13	Appendix	1025	0.3	2.3	100	68	100	77.4

Sequence length $L = 500$ kb containing five kinds of repeats. Contigs of repeat and nonrepeat are generated in an independent way by SWA at different levels with basic parameters: sequencing depth $S_d = 0.5$, size of sliding window $L_w = 3$, read length $L_r = 60$, and k-mer = 15. Contigs smaller than 200 are removed.

of assembling repeats. The effect of sliding window is to filter out the bias caused by sequencing process; therefore the increase of filtered times can improve the effect of filtering bias. It is clear that the Rep-accuracy was up to 82.1% after being filtered twice from 66.8% but got worse to 68% after three times filtering. So too much filtered times can lead to misassembled contigs across the boundary of repeats and nonrepeats as shown in Table 5. Therefore, for real NGS data, some compromise is necessary for choosing optimal filtered times.

3.3. Assessments in Reference Datasets. In this section, we validated the performances of SWA in reference genome datasets of three species in Materials. We analysed their repeat structures by whole genome scan using RepeatScout [26]. The repeats structures including lengths and copies were detailed in the Appendix and the link *detected repeats and their copies in Table 6* in supporting data. The results of SWA are presented in Table 6.

Table 6 shows the assembly statistics of three species by SWA. All contigs were corrected and verified by aligning them back to reference genome, so the C-accuracy was 100%. For chrIV-S.c datasets, SWA generated 32 repeats and 315 nonrepeats; the copy numbers of assembled repeats are presented in the Appendix. In our whole genome scanning, 41 repeats longer than 200 were identified. By aligning these assembled repeats back to the reference, 4 tandem repeats were assembled together and the left were all correct. So the CN-accuracy is about 88%, but C-accuracy and genome coverage are almost up to 100%. For *E. coli*, SWA generated 50 repeats and 259 nonrepeats; the copy numbers of assembled repeats are presented in the Appendix. By whole genome scanning, 57 repeats were identified. So CN-accuracy is about 88%, but genome coverage is 99.7%. For chrIII-C.e datasets, SWA generates 198 repeats and 5471 nonrepeats. In our whole genome scanning, 339 repeats longer than 200 were identified. By aligning these assembled repeats back to reference, 103 short tandem repeats were assembled together. So the accuracy is about 89%. But the genome coverage is a little lower.

3.4. Assessments in NGS Datasets. In order to assess the performances of SWA more comprehensively, we performed comparisons with other eight leading genome assemblers presented in GAGE [27], such as ABySS, ALLPATHS-LG, Bambus2, CABOG, MSR-CA, SGA, SOAPdenovo, and Velvet. We used the assembly evaluation script provided by GAGE to assess various assembly metrics. Briefly, the GAGE script aligns contigs back to the reference genome and calculates the corrected N50 length by breaking contigs at misassembled sites. Tables 7, 8, and 9 show the assembly metrics for SWA and eight others in three species including *S. aureus*, *R. sphaeroides*, and human chromosome 14. We did not run these assemblers on the whole human genomes due to the following reasons: (1) some of the assemblers in our comparison would take many weeks to assemble the complete genome and others would fail entirely; and (2) high computing platform is not available. The statistics for these eight assemblers were taken from GAGE study. In order to compare the ability of assembling repeats fairly, all contigs are aligned back to repeats using *swalign* function [25].

Table 7 shows the assembly statistics for *S. aureus* dataset by nine assemblers. For *S. a*, SWA performs best in assembling repeats and nonrepeats. By aligning them back to repeats, SWA generated 30 repeats with total size 20.7 kb. SGA generated 18 repeats with total size 19.4 kb. Velvet generated 15 repeats with total size 13.8 kb. In terms of the completeness of types of repeats, SWA achieved the best assembly. The Rep-acc of SWA and SGA and Velvet are 82% and 83.8%, respectively. Therefore, for the accuracy of assembling repeats, SWA and SGA are the top two assemblers. Other assemblers had poor performances in the accuracy of assembling repeats and nonrepeats. Particularly, Allpath-LG only generates 3 repeats with size 2.6 kb and Rep-acc 4.3%. Because Allpath-LG achieved the longest corrected N50 length (66.2 kb), long contig can cross the boundary of repeat and lead to indistinguishable contig. So the better the continuity of assembler is the worse the completeness of assembling repeats and nonrepeats will be. For the continuity of assembly, SWA was read loss to other eight assemblers. However, for the assembly size and genome coverage, SWA was also the best one with assembly size 2,939 kb and genome coverage 100.1%, which is most approximate to the real genome size. So in terms of completeness of assembly, SWA achieved the best assembly.

Table 8 shows the assembly statistics for *R. sphaeroides* dataset by nine assemblers. By aligning them back to repeats, SWA generated 13 repeats with total size 7.8 kb. AByss, SGA, and SOAPdenovo generated 13 repeats with total size 8.3 kb, 9 repeats with total size 3.9 kb, and 9 repeats with total size 3.7 kb, respectively. The Rep-acc of them is 44.5%, 44.5%, 64.3%, and 84.5%, respectively. Therefore, in terms of accuracy of assembling repeats, SOAPdenovo outperformed others. But for the completeness of size of repeats, SWA and AByss are the top two assemblers. Other

Table 6: The results of SWA in reference genome datasets.

Species	Repeat		Contigs (kb)			Accuracy (%)			Genome coverage (%)
	TRC	CNRC	NNC	N50	Max	CN-accuracy	Rep-accuracy	C-accuracy	
chrIV-S.c	32	Appendix	315	9.4	32.7	88	93	100	100
E. coli	50	Appendix	259	46.5	190.5	88	99.8	100	99.7
chrIII-C.e	198	Appendix	5471	4.6	26.7	89	92	100	97.7

Contigs of repeat and nonrepeat are generated in an independent way by SWA with basic parameters: sequencing depth $S_d = 2$, read length $L_r = 60$, filtered times = 1, and sliding window = 3. Contigs smaller than 200 are removed.

Table 7: Assemblies of *S. aureus* (genome size 2,903,081).

Assemblers	Repeat				Contigs (kb)					*E*-size	Assembly size (kb)	Genome coverage (%)
	TRC	NNC	Rep-size	Rep-acc	Num.C	N90	Mean	N50	Max			
SWA	30	1834	20.7 kb	82%	1864	0.8	1.6	2.5	14.4	3115	2939	100.1
ABySS	7	293	6.8 kb	30%	302	7.0	12	24.8	125	31403	3647	125.6
Allpaths-LG	3	57	2.6 kb	4.3%	60	31	47	66.2	234	90078	2869	98.8
Bambus2	6	103	7.5 kb	35%	109	11	15	16.7	158	19610	2833	97.6
MSR-CA	7	87	4.4 kb	15%	94	21	30	48.2	139	50381	2862	98.5
SGA	18	1232	19.4 kb	83.8%	1252	1.0	2.2	4.0	16.8	4712	2833	97.6
SOAPdenovo	9	98	7.5 kb	38.6%	107	35.5	27	62.7	518.7	68002	2909	100.2
Velvet	15	147	13.8 kb	39.3%	165	11.4	17.6	41.5	169	48511	2847	98

N50, N90, and mean values are based on the same genome size. The contigs are all corrected and those smaller than 200 were removed.

Table 8: Assemblies of *Rhodobacter sphaeroides* (genome size 4,603,060).

Assemblers	Repeat				Contigs (kb)					*E*-size	Assembly size (kb)	Genome coverage (%)
	TRC	NNC	Rep-size	Rep-acc	Num.C	N90	Mean	N50	Max			
SWA	13	1774	7.8 kb	44.5%	1787	1.3	2.6	4.2	29.4	5812	4600.3	99.94
ABySS	13	1910	8.3 kb	44.5%	1915	1.1	2.5	4.2	54.7	6877	4969.5	108
Allpaths-LG	2	202	1.4 kb	6.7%	204	11.5	22.5	34.4	106	35973	4587.8	99.6
Bambus2	2	175	1.3 kb	9.2%	177	6.1	8.5	12.8	279	16281	4371.6	94.9
CABOG	1	312	0.3 kb	8.9%	322	6.5	13	17.9	88.5	21539	4238	92
MSR-CA	7	388	4.1 kb	25.4%	395	5.7	11.3	19	83.7	21579	4465	97
SGA	9	3053	3.9 kb	64.3%	3067	0.66	1.4	2.9	29.5	4067	4502.7	97
SOAPdenovo	9	195	3.7 kb	84.5%	204	7.8	11.8	14.3	376	18553	4596	99.8
Velvet	5	578	2.2 kb	24.5%	583	4.6	7.7	14.5	60.7	16711	4503	97.8

N50, N90, and mean values are based on the same genome size. The contigs are all corrected and those smaller than 200 were removed.

Table 9: Assemblies of human chromosome 14 (genome size 88,289,540).

Assemblers	Repeat				Contigs (kb)					*E*-size	Assembly size (kb)	Genome coverage (%)
	TRC	NNC	Rep-size	Rep-acc	Num.C	N90	Mean	N50	Max			
SWA	198	26824	129.3 kb	88.3%	27021	1.6	3.3	6.7	52	8737	87936	99.6
ABySS	143	51647	121.4 kb	30.1%	51924	0.7	1.7	2.0	30	3134	73341	83
Allpaths-LG	88	4441	156.7 kb	47.3%	4529	9.7	18.7	21.0	240	27157	84435	95.6
Bambus2	154	13437	293 kb	66.7%	13592	2.4	3.1	4.3	261	6345	68243	77.3
CABOG	70	3291	255 kb	39.9%	3361	13.7	21.6	23.7	296	30689	86232	97.6
MSR-CA	190	29901	526 kb	85.3%	30103	1.2	2.7	4.3	53.9	5927	83291	94.3
SGA	187	56278	283 kb	50.4%	56939	0.6	1.4	2.7	30	3737	82375	93.3
SOAPdenovo	188	21552	476 kb	73.3%	22689	1.8	4.2	7.4	141	9801	92603	104.9
Velvet	192	45294	339 kb	56.9%	45564	0.7	1.6	2.1	22.5	3049	74740	84.6

N50, N90, and mean values are based on the same genome size. The contigs are all corrected and those smaller than 200 were removed.

five assemblers performed worse in assembling repeats and nonrepeats. Particularly, Allpath-LG, Bambus2, and CABOG only generated less than three repeats with length less than 1.4 kb, because in terms of continuity, these three assemblers, Allpath-LG, Bambus2, and CABOG, are the top ones. But long contigs can lead to indistinguishable contigs. However, in terms of assembly size and genome coverage, SWA also achieved the best assembly with genome size 4,600 kb and genome coverage 99.9%, which is most approximate to the real genome size. So in terms of completeness of assembly, SWA performed best among these nine assemblers.

Table 9 shows the assembly statistics for human chromosome 14 dataset by nine assemblers. For *H. s* 14, SWA performs best in assembling repeats and nonrepeats. By aligning them back to repeats, there are four top assemblers in terms of assembling repeats, such as SWA, MSR-CA, SOAPdenovo, and Velvet. There are 198 repeats with total size 129.3 kb, 190 repeats with total size 526 kb, 188 repeats with total size 476 kb, and 192 repeats with total size 339 kb generated by SWA, MSR-CA, SOAPdenovo, and Velvet, respectively. The Rep-acc of corresponding items is 88.3%, 85.3%, 73.3%, and 56.9%, respectively. For the accuracy of assembling repeats, SWA and MSR-CA achieved the best results. In terms of completeness of types of repeats, SWA achieved the best results. However, in terms of continuity, Allpath-LG, CABOG, and SOAPdenovo outperformed SWA. But for the genome size and genome coverage, SWA achieved the best results with assembled size 8,7936 kb and genome coverage 99.6%. So in terms of completeness of assembly, SWA outperformed other assemblers.

One can safely come to a conclusion from Tables 7, 8, and 9 that SWA performed best in assembling repeats and nonrepeats in three NGS datasets. In terms of assembling repeats and completeness of repeats, SWA is the top one among these nine assemblers. One may argue that the contiguity of SWA is not better than others; the metrics such as N50, N90, and E-size are smaller than some of the other assemblers. This is because SWA is specially designed for assembling repeats and nonrepeats. Therefore, SWA stops extending contigs automatically when the boundary of repeats is detected. Meanwhile, the continuity of assembly and the completeness of repeats and nonrepeats are the pair of contradiction. On the other hand, the better completeness of repeats and nonrepeats requiers that contigs must be stopped at the boundary of repeats. Therefore, the continuity of assembly will be down.

4. Discussions

4.1. Sequencing Strategies for SWA. The uniformity of the sequencing process is very important for SWA, because assembling repeats and nonrepeats independently of SWA is based on the combination of coverage depth and sliding window. The effect of sliding window is to filter out the bias caused by sequencing process, because sequencing bias makes the frequency of repeat and nonrepeat more ambiguous to determine, so large sequencing bias may result in short contigs or misassembly. For the appropriately uniformed sequencing data, SWA cannot only assemble repeats and nonrepeats independently but can also estimate their copy numbers correctly. In this situation, contigs of repeats and nonrepeats can be easily grouped into scaffolds by SWA using only the short insert paired-end information, while other current assemblers all need the good combination of several mate-pair libraries with different insert lengths, which increase the cost of sequencing and complexity of technologies. Therefore SWA provides a simple sequencing strategy for NGS technologies; that is, long-distant library is not necessary. So similar to the strategies recommended by ALLPATHS-LG [19], we recommend that for the Illumina technology one should use an overlapping paired-end library with a suitable insert size to generate PE raw reads for contig assembly and there is no need for several mate-pair libraries with different insert lengths to generate long-distant jumping reads for scaffolds. The average genome coverage is at least 100× or higher. For generating overlapping paired-end reads, we provide a simple formula to calculate the insert size for constructing a paired-end library: insert Size = (read length (l) + max error tolerance (m)) × 2 − max overlap length (n). For example, if the read length is $l = 150$ bp, the max overlap length $n = 100$ bp and the max error tolerance $m = 50$; then the recommended insert size is 300 bp.

4.2. Seed Selection. In an extension-based assembler, a good seed should not contain any sequencing errors and should not be selected from the boundary of repeats and nonrepeats. A read from repeat region usually has a high read count because identical repeats from other loci are counted as well. On the other hand, a read from nonrepeat region always has a low read count. However, the read from the boundary always has the middle count under the condition of uniformed sequencing process. These seeds are hard to determine whether they belong to repeat region or nonrepeat region and always lead to misassembly or short contigs. Thus, the seeds for repeat region are chosen with high read count, while the one for nonrepeat region should be chosen with low read count, and seeds with middle read count should be avoided.

4.3. Parallel Operation. Obviously, parallel operation can save the executive time and reduce the memory use. SWA can assemble repeats and nonrepeats independently by using two different computers without any communication. This property can shorten the executive time almost a half and reduce the memory use in some extent. Furthermore, the raw data also can be classified into two parts, repeats and nonrepets, according to read count. This strategy can reduce the memory usage largely.

4.4. Optimal Sliding Window. The sliding window plays an important role in contig construction in SWA. By filtering out sequencing bias, SWA can distinguish repeats and nonrepeats from NGS data easily so as to assemble repeats and nonrepeats independently. The mean value of read count in sliding window determines whether the extension should continue or not. So too small sliding window cannot filter out the

bias efficiently but has high sensitiveness of detecting changes of read count. A too large sliding window can filter out the bias efficiently but decreases the sensitiveness of detecting repeats and leads to misassembly. The optimal sliding window should have both the property of filtering bias efficiently and detecting repeats sensitively. So the compromise is necessary in practice.

4.5. Optimal Read. In the stage of seed extension, the optimal read is needed in order to extend the seed in dynamic overlapping interval.

In SWA, the optimal read was identified using the following strategy: the one overlapped most bases with seed in dynamic overlapping interval was taken as the optimal read. In theory, longer overlapping with seed means higher accuracy of assembly. But this strategy has low speed, because the seed only extends one or two bases at one extension. Of course, the other strategy for choosing optimal read in dynamic overlapping interval also can be adopted, such as the optimal read can be identified as the one with read counts nearest to the theoretically sequencing depth. This strategy has a higher speed, but the correctness of extension will be low compared with the first strategy. In practice, the strategy of identifying optimal read can be chosen by users.

4.6. Comparisons. SWA is specially designed for assembling repeats and nonrepeats, respectively. What we are mainly concerned with is the correctness and accuracy of assembling repeats and estimating their copy numbers rather than the length of assembled contigs, so SWA stops extending contig automatically when detecting the boundary of repeat and nonrepeat. Duo to this, the validations and evaluations are performed rather than comparisons with other assemblers in simulations and reference datasets. In real NGS datasets, the comparisons were performed comprehensively with other eight leading assemblers. But the accuracy and completeness of repeats are what we are firstly concerned with.

5. Conclusions

In this paper, we developed a *de novo* genome assembly algorithm named SWA, which can assemble repeats and nonrepeats independently. The most important features of SWA are (1) assembling repeats and nonrepeats completely and accurately; (2) adopting sliding window function to filter out sequencing bias in genome assembly process; (3) compensating the loss of low coverage; and (4) estimating the copies of each assembled contigs. Consequently, in this study, we have validated the performances of SWA and compared them with other leading assemblers in three real NGS datasets. For comparisons in real NGS datasets, the metrics such as TRC, NNC, and Rep-size are used to evaluate the completeness of assembled repeats and nonrepeats; the metrics such as Rep-accuracy, C-accuracy, and CN-accuracy are used to evaluate the accuracy of assembled repeats and nonrepeats, while the N50, N90, and maximum contig are used to evaluate the continuity of whole genome assembly. Results indicated that SWA outperformed other leading assemblers in the completeness and correctness of assembling repeats, but the continuity was not better than some of the others. It is natural, because SWA is not specially designed for whole genome assembly and continuity is not what SWA is firstly concerned with.

In general, without long insert-size libraries, repeats that extend beyond the paired-end insert sizes will be difficult to resolve and assemble. Although, some of the compared assemblers can assemble long repeats with simple structure, the completeness and accuracy are not good. It is natural that the continuity is what a whole genome assembler is mainly concerned with. However, SWA is not a whole genome assembly. So bridging two unique sequences around a repeat is not allowed by SWA in order to ensure the completeness of separating repeats and nonrepeats. Even though SWA was not aiming for the whole genome assembly, SWA also provided another solution to resolve long repeats without the help of long insert-size libraries by assembling from nonrepeats completely. In theory, for the whole genome assemblers, if repeats and nonrepeats are assembled correctly and completely, their copies are estimated correctly. Scaffolds can be grouped easily by using short-insert paired-end information rather than the good combination of several libraries. In practice, repeat characteristics in different genomes can vary extensively and depths of sequencing can be highly uneven along the genome, so the expected theoretical *de novo* assembly results from different genomes will also vary.

6. Methods

6.1. The Detailed Outlines

6.1.1. The Steps of Extending Repeats (Figure 4)

(1) Selecting a seed in repeat regions with high frequency larger than H_p in table R.

(2) Computing read counts overlapped with seeds in dynamic overlapping intervals.

(3) Filtering the overlapped read counts by sliding window and then computing the mean value of this interval and recording in M_n.

(4) Judging whether the extension of seed is out of the bound of repeat or not. If $M_n > T_2$, the extension will continue or else stop extension at this end.

(5) Extending seed using the optimal read in the dynamic overlapping interval. The optimal read in SWA is the unique read with longest overlapped bases.

(6) Continue Step 2–Step 5 until this seed is stopped at both $3'$-end and $5'$-end.

(7) If repetitive seed sets are not empty, go to Step 1 and repeat these steps, else repetitive contigs construction are finished.

6.1.2. The Steps of Extending Nonrepeats (Figure 5)

(1) Selecting a seed in nonrepeat regions with low frequency smaller than L_p in table R.

(2) Computing read counts overlapped with seeds in dynamic overlapping intervals.

(3) Filtering the overlapped read counts by sliding window and then computing the mean value of this interval and recording in M_n.

(4) Judging whether the extension of seed is up to the boundary of repeat or not. If $M_n < T_1$, the extension will continue, or else stop extension at this end.

(5) Extending seed using the optimal read in the dynamic overlapping interval. The optimal read in SWA is the unique read with longest overlapped bases.

(6) Continue Step 2–Step 5 until the extension is stopped at both $3'$-end and $5'$-end.

(7) If nonrepetitive seed sets are not empty, go to Step 1 and repeat these steps, else nonrepetitive contigs construction are finished.

6.2. Sliding Window. Coverage bias is inevitable in genome sequencing process and is usually caused by whole genome amplification (WGA) [28]. Three primary forms of WGA have been developed: multiple displacement amplification (MDA) [29], primer extension preamplification (PEP) [30], and degenerate oligonucleotide primed PCR (DOP) [31]. Furthermore, coverage bias also can be amplified by data cleaning and error correction stage. The existence of coverage bias increases the nonuniformity of read depth for detecting copy numbers of repeats, which can lead to extending contigs crossing the bound of repeats and incorrect estimation of copy numbers of repeat contigs. How to eliminate or decrease these noises is a very important step in constructing contigs of repeat or nonrepeat regions. So the sliding window is used to filter out the noise caused by coverage bias so as to improve the performances of distinguishing repeats from nonrepeats and estimating the copy number of each repeat contig.

In order to decrease the coverage bias, we use rectangular window function to smooth coverage bias. Rectangular window function is defined as follows:

$$w(n) = \begin{cases} 1, & 0 \le n \le L_w - 1 \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

here L_w is the length of sliding window. So

$$w(i) = \frac{1}{L_w} \sum_{j=i-(L_w/2)}^{i+(L_w/2)} x_j, \quad i = 1, 2, \ldots, L_d, \tag{3}$$

where x_j represents read counts by overlapping $L_r - j$ bases.

In theory, the longer the L_w is, the better the sliding effect will be. However, L_w cannot be too large or too small which is closely bounded to read length L_r and length of dynamic overlapping interval L_d. In order to guarantee the correctness of SWA, we set $L_d \ge (3/4)L_r, 2 \le L_w \le L_d/2$. Of course, this is an empirical value.

6.3. The Effect of Window Function. By mathematical derivation, we can clearly see that the variances smoothed by window function are smaller than the original ones. Letting $X_n = \{x_1, x_2, \ldots, x_n\}$ be the n preprocessed data points and $E(X_n) = (1/n)\sum_{i=1}^{n} x_i$ be the mean of X_n, the variance of X_n is as follows: $D(X_n) = (1/n)\sum_{i=1}^{n}(x_i - E(X_n))^2$. Letting sliding window function $y = w(x), Y_n = \{y_1, y_2, \ldots, y_n\}$ be corresponding smoothed data points, $y_i = w(x_i) = (1/L_w)\sum_{j=i-Lw/2}^{i+Lw/2} x_i$, the mean value of Y_n is $E(Y_n) = (1/n)\sum_{i=1}^{n} y_i$. It is clear that $E(X_n) = E(Y_n)$; $D(Y_n) = (1/n)\sum_{i=1}^{n}(y_i - E(Y_n))^2$. One can easily prove that $D(Y_n) \le (1/L_w)D(X_n)$.

6.4. Hash Index. Overlap computing is the most time consuming stage for all against all. In order to speed up SWA, we use hash index to store the location of each keyword and read rather than the keyword itself in hash index. The details are as follows. For each read, the N continuous bases from $3'$-end and $5'$-end are selected and then mapped into two variables, forward and backward, respectively. And then, map $A \to 0, C \to 1, G \to 2, T \to 3$. So a string consisting of N continuous bases is transferred into quaternary integers; the quaternary integers then are transferred into decimal integer. So each keyword is mapped to a unique location in hash index which stores the identification of unique processed reads. We define the hash function as

$$H(S[i, i + L_k]) = Q_i Q_{i+1}, \ldots, Q_{i+L_k}(\text{quaternary}), \tag{4}$$

where

$$Q_i = \begin{cases} 0, & s[i] = \text{“}A\text{”} \\ 1, & s[i] = \text{“}C\text{”} \\ 2, & s[i] = \text{“}G\text{”} \\ 3, & s[i] = \text{“}T\text{”}; \end{cases} \tag{5}$$

L_k is the length of keywords in hash function.

6.5. Parameters of Kernel SWA Program. The kernel program of SWA has eight parameters: the maximum overlapping length max, dynamic overlapping interval L_d, read length L_r, length of sliding window L_w, threshold of repetitive seeds H_p, threshold of nonrepetitive seeds L_p, threshold of repeats assembly T_2, threshold of nonrepeats assembly T_1, and sequencing depth after filtering by sliding window S_{fd}. On the basis of the extensive comparisons of three species as shown in the paper, we suggest that the only values of maximum overlapping length, length of dynamic overlapping interval, and read length may not be adjusted. We recommend the following: $L_r = 101$, max $= 100$, $L_w = 9$, and k-mer can be set in the range [35, 50]. These four parameters: H_p, L_p, T_1, and T_2 are closely related to sequencing depth S_d and length of sliding window L_w. For example, if sequencing depth is $S_d = 2$, length of sliding window $L_w = 3$ and filtered times $N_f = 1$, and then H_p should be higher than double sequencing depth; that is, $H_p = 5$. L_p should be lower than sequencing depth, so let $L_p = 1$. T_1 should be a little higher than the filtered sequencing depth $S_{\text{fd}} = S_d \times L_w^{N_f} = 2 \times 3 = 6$, so let $T_1 = 8$. T_2 should be a little lower than

the $2 \times S_{\text{fd}} = 2 \times 6 = 12$, so let $T_2 = 10$, where sequencing depth is $S_d = N_r/L$, coverage is $C = (N_r/L) \times L_r = S_d \times L_r$, and N_r is the number of reads.

Parameters have a great influence on the performances of SWA. Therefore, in practice, fine tuning is necessary for special genome with intrinsic complex repeat structure or different backgrounds of sequencing bias. We suggest that the parameters needed to fine tune are T_1 and T_2. The abbreviations are presented in the Abbreviation Section.

6.6. Thresholds. The threshold for repeats and nonrepeats are based on the size of confidence intervals and significance testing, which are closely related to coverage depth, size of sliding window, and filtered times.

Let $X = \{x_1, x_2, \ldots, x_n\}$ be the overlap numbers in dynamic overlapping stage and the mean of X is $m_x = (1/n) \sum x_i$. Let $Y = \{y_1, y_2, \ldots, y_n\}$ be the read counts filtered by sliding window, so the mean of Y is $m_y = (1/n) \sum y_i \approx m_x \times L_w$. According to the law of large number in the probability and statistic theory, we can easily get $Y \sim N(m_y, \sigma^2)$. So if the average sequencing depth is S_d, the average read counts filtered by sliding window is $S_{\text{fd}} = S_d \times L_w$. Thus, the read counts of the nonrepeat region filtered by sliding window should be S_{fd} if sequencing bias is free, and the corresponding items of repeat region with two copies should be $2S_{\text{fd}}$. Clearly $S_{\text{fd}} = m_y$ if sequencing bias is free. Let $\delta = F_d/2$; random variable m_y is the mean read counts filtered by sliding window.

If $P\{m_y \leq F_d + \delta_1\} \geq 1 - \alpha$, so the confidence upper limit T_1 of nonrepeat region at the confidence level $1 - \alpha$ is $T_1 = F_d + \delta_1$ which is the threshold of nonrepeats. If $P\{m_y \geq 2F_d - \delta_2\} \geq 1 - \alpha$, the confidence lower limit T_2 of nonrepeat region at the confidence level $1 - \alpha$ is $T_2 = 2F_d - \delta_2$ which is threshold of repeats, where $0 \leq \{\delta_1, \delta_2\} \leq \delta$, and $\{\delta_1, \delta_2\}$ can be used to control the type-I error and type-II error since the statistical tests of overlapping intervals of windows are not independent. The construction of repeat contig and nonrepeat contig is generated separately.

6.7. Estimating Copy Numbers. The copy number of each repeat contig is estimated by using significant testing methods. After finishing contig construction, the variable M_n has stored the whole filtered read counts and its mean value is computed. The copy number of corresponding items is estimated by rounding mean$(M_n)/S_{\text{fd}}$ to the nearest integer.

7. Materials

In this study, three kinds of datasets are used to validate the performances of SWA and compare with other assemblers. They are real simulated datasets, reference datasets, and real NGS datasets.

For real simulated datasets, the model sequences, A, B, and C are randomly sampled from {A, T, C, and G} with different repetitive contents. These three sequences contain tandem repeats, interspersed repeats, and compound repeats, respectively (as shown in Figure 6). These repetitive contents represent a wide range of length and copies.

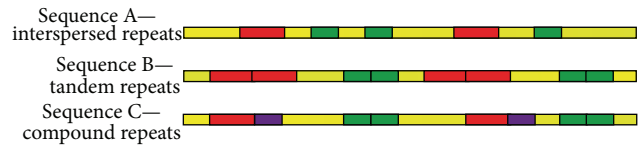


Figure 6: The graphic explaining the real simulated datasets. The yellow lines represent the model of random sequences. The red lines, blue lines, and purple lines represent different contents of repeats. *Sequence A* represents the interspersed repeats that is different repeats do not link each other closely. *Sequence B* represents tandem repeats that is same repeats link each other in the cascade manner. *Sequence C* contains the compound repeats, which is the combination of *sequence A* and *sequence B*. The detailed generation process is presented in supplementary materials.

The detailed information is presented in supplementary table (see Table S3 in Supplementary Material available online at http://dx.doi.org/10.1155/2014/736473). And the generation process is also presented in supplementary materials. Then, the paired-end NGS reads are randomly sampled from the fragments with normal distribution N (300, 30).

For reference genome datasets, we download the reference genome of *S. cerevisiae*, *C. elegans* from UCSC (http://hgdownload.soe.ucsc.edu/downloads.html) and *E. coli k12* (GenBank: U00096.3). For *S. cerevisiae* and *C. elegans*, we only randomly take chromosome IV and chromosome III, respectively. The sizes of chrIV-S.c, *E. coli*, and chrIII-C.e are 1,531,933 bp, 5,132,068 bp, and 13,783,700 bp, respectively. Their repeats structures can be easily analyzed by RepeatScout [26], which is a very effective and sensitive *de novo* repeats identification method for large genomes and is freely available at http://bix.ucsd.edu/repeatscout/. The repeats structures including lengths and copies are detailed in the Supplementary Materials Appendix and are freely available at http://222.200.182.71/swa/Table6.rar.

For real NGS datasets, two bacterial genomes (*Staphylococcus aureus* and *Rhodobacter sphaeroides*, genome sizes of 2.9 and 4.6 Mb, resp.) and *human chromosome 14* (genome size of 88.3 Mb) were downloaded from http://gage.cbcb.umd.edu/data/. In the GAGE study [27], all reads were error-corrected before assembly by ABySS, ALLPATHS-LG, Bambus2, Celera Assembler with the Best Overlap Graph (CABOG), Maryland Super-Reads Celera Assembler (MSR-CA), SGA, SOAPdenovo, and Velvet. For a fair comparison, we also obtained these corrected datasets for using in GAGE. These three species have perfect reference genomes. Therefore, their real repeats structures can be easily detected by RepeatScout [26]. For *S. a*, *R. s*, and *H. s* 14, there are 52 repeats, 21 repeats, and 259 repeats detected by RepeatScout [26], respectively, with length longer than 100 bp; their total sizes are 15.8 kb, 3.6 kb, and 146.5 kb, respectively.

7.1. Implementation. Table 10 presents the detailed memory usage and CPU times of SWA in three real NGS datasets. Different stages have different requirements of memory. The

Table 10: Memory usage and CPU times of SWA.

Species	Memory usage	CPU times
S. aureus	2.5 GB	59.5 minutes
Rhodobacter sphaeroides	3.4 GB	96.3 minutes
Human chromosome 14	22.6 GB	56.2 hours

memory usage presented in Table 10 is the maximum memory. The CPU times refer to the run time of main procedure except for the preprocessing stage. Because a different assembler has different hardware requirements; therefore the direct comparisons are not reasonable to some extent. However, Table 10 gives users a rough guidance for hardware requirement and run time. Those of others have been accessed clearly in GAGE study and are freely available at http://genome.cshlp.org/content/early/2012/01/12/gr.131383.111/suppl/DC1.

SWA is implemented in MATLAB computing environment. Programming language: m language. Operation systems: Windows, Linux. Computing platform: 3.5 GHz eight Intel Celeron CPU with 32 GB RAM and 64-bit operational system.

8. Availability of Supporting Data

The supporting data including NGS data and assembling results are freely available at http://222.200.182.71/swa/Results.rar.

The detected repeats and their copies of three species used in Table 6 can be freely found at http://222.200.182.71/swa/Table6.rar.

The detected repeats and their copies of three species used in Table 7, 8, 9 can be freely found at http://222.200.182.71/swa/Tables789.rar.

Abbreviations

SWA:	Sliding window assembling
TRC:	The types of repetitive contigs
CNRC:	Copy number of each repeat contig
NNC:	Number of nonrepeat contigs
Rep-acc:	The accuracy of assembled repeat contigs
C-accuracy:	The accuracy of the total contigs
CN-accuracy:	The accuracy of estimated copy number of each repeat
k-mer:	Maximum overlap minus the minimum overlap
N_f:	The number of filtered times by sliding window
L_w:	The length of sliding window
L_r:	The length of reads
L_d:	The length of dynamic overlapping interval
H_p:	The threshold of repetitive seeds
L_p:	The threshold of nonrepetitive seeds
L_d:	The length of sequencing reads
S_d:	The sequencing depth
T_1:	Threshold for nonrepeats extension
T_2:	Threshold for repeats extension.

Conflict of Interests

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

Authors' Contribution

The algorithm was developed by Shuaibin Lian and Xianhua Dai. The experiments were performed by Shuaibin Lian and Qingyan Li. The paper was prepared by Shuaibin Lian and Qingyan Li. Qingyan Li, Zhiming Dai, and Qian Xiang participated in the analysis and discussion. All authors read and approved the final paper.

Acknowledgment

This paper is supported by the National Nature Science Fund of China, Grant no. 61174163.

References

[1] D. R. Bentley, "Whole-genome re-sequencing," *Current Opinion in Genetics and Development*, vol. 16, no. 6, pp. 545–552, 2006.

[2] T. D. Harris, P. R. Buzby, H. Babcock et al., "Single-molecule DNA sequencing of a viral genome," *Science*, vol. 320, no. 5872, pp. 106–109, 2008.

[3] M. L. Metzker, "Sequencing technologies the next generation," *Nature Reviews Genetics*, vol. 11, no. 1, pp. 31–46, 2010.

[4] E. R. Mardis, "The impact of next-generation sequencing technology on genetics," *Trends in Genetics*, vol. 24, no. 3, pp. 133–141, 2008.

[5] O. Morozova and M. A. Marra, "Applications of next-generation sequencing technologies in functional genomics," *Genomics*, vol. 92, no. 5, pp. 255–264, 2008.

[6] R. L. Strausberg, S. Levy, and Y.-H. Rogers, "Emerging DNA sequencing technologies for human genomic medicine," *Drug Discovery Today*, vol. 13, no. 13-14, pp. 569–577, 2008.

[7] E. Pettersson, J. Lundeberg, and A. Ahmadian, "Generations of sequencing technologies," *Genomics*, vol. 93, no. 2, pp. 105–111, 2009.

[8] The 1000 Genomes Project Consurtium, "A map of human genome variation from population-scale sequencing," *Nature*, vol. 467, pp. 1061–1073, 2010.

[9] Genome 10K Community of Scientists, "Genome 10K: a proposal to obtain whole-genome sequence for 10, 000 vertebrate species," *Journal of Heredity*, vol. 100, pp. 659–674, 2009.

[10] T. J. Treangen and S. L. Salzberg, "Repetitive DNA and next-generation sequencing: computational challenges and solutions," *Nature Reviews Genetics*, vol. 13, no. 1, pp. 36–46, 2012.

[11] J. T. Simpson, K. Wong, S. D. Jackman, J. E. Schein, S. J. M. Jones, and I. Birol, "ABySS: a parallel assembler for short read sequence data," *Genome Research*, vol. 19, no. 6, pp. 1117–1123, 2009.

[12] S. Gnerre, I. MacCallum, D. Przybylski et al., "High-quality draft assemblies of mammalian genomes from massively parallel sequence data," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 4, pp. 1513–1518, 2011.

[13] S. Koren, T. J. Treangen, and M. Pop, "Bambus 2: scaffolding metagenomes," *Bioinformatics*, vol. 27, no. 21, Article ID btr520, pp. 2964–2971, 2011.

[14] J. R. Miller, A. L. Delcher, S. Koren et al., "Aggressive assembly of pyrosequencing reads with mates," *Bioinformatics*, vol. 24, no. 24, pp. 2818–2824, 2008.

[15] J. T. Simpson and R. Durbin, "Efficient de novo assembly of large genomes using compressed data structures," *Genome Research*, vol. 22, no. 3, pp. 549–556, 2012.

[16] R. Li, H. Zhu, J. Ruan et al., "De novo assembly of human genomes with massively parallel short read sequencing," *Genome Research*, vol. 20, no. 2, pp. 265–272, 2010.

[17] D. R. Zerbino and E. Birney, "Velvet: algorithms for de novo short read assembly using de Bruijn graphs," *Genome Research*, vol. 18, no. 5, pp. 821–829, 2008.

[18] L. D. Stein, Z. Bao, D. Blasiar et al., "The genome sequence of Caenorhabditis briggsae: a platform for comparative genomics," *PLoS Biology*, vol. 1, no. 2, article E45, 2003.

[19] International Human Genome Consortium, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, pp. 860–921, 2001.

[20] A. J. Iafrate, L. Feuk, M. N. Rivera et al., "Detection of large-scale variation in the human genome," *Nature Genetics*, vol. 36, no. 9, pp. 949–951, 2004.

[21] L. Feuk, A. R. Carson, and S. W. Scherer, "Structural variation in the human genome," *Nature Reviews Genetics*, vol. 7, no. 2, pp. 85–97, 2006.

[22] D. G. Albertson and D. Pinkel, "Genomic microarrays in human genetic disease and cancer," *Human Molecular Genetics*, vol. 12, no. 2, pp. R145–R152, 2003.

[23] C. R. Marshall, A. Noor, J. B. Vincent et al., "Structural variation of chromosomes in autism spectrum disorder," *The American Journal of Human Genetics*, vol. 82, no. 2, pp. 477–488, 2008.

[24] J. Sebat, B. Lakshmi, D. Malhotra et al., "Strong association of de novo copy number mutations with autism," *Science*, vol. 316, no. 5823, pp. 445–449, 2007.

[25] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *Journal of Molecular Biology*, vol. 147, no. 1, pp. 195–197, 1981.

[26] A. L. Price, N. C. Jones, and P. A. Pevzner, "De novo identification of repeat families in large genomes," *Bioinformatics*, vol. 21, no. 1, pp. i351–i358, 2005.

[27] S. L. Salzberg, A. M. Phillippy, A. Zimin et al., "GAGE: a critical evaluation of genome assemblies and assembly algorithms," *Genome Research*, vol. 22, no. 3, pp. 557–567, 2012.

[28] R. Pinard, A. de Winter, G. J. Sarkis et al., "Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing," *BMC Genomics*, vol. 7, article 216, 2006.

[29] J. M. Lage, J. H. Leamon, T. Pejovic et al., "Whole genome analysis of genetic alterations in small DNA samples using hyperbranched strand displacement amplification and array-CGH," *Genome Research*, vol. 13, no. 2, pp. 294–307, 2003.

[30] L. Zhang, X. Cui, K. Schmitt, R. Hubert, W. Navidi, and N. Arnheim, "Whole genome amplification from a single cell: implications for genetic analysis," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 89, no. 13, pp. 5847–5851, 1992.

[31] H. Telenius, N. P. Carter, C. E. Bebb, M. Nordenskjold, B. A. J. Ponder, and A. Tunnacliffe, "Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer," *Genomics*, vol. 13, no. 3, pp. 718–725, 1992.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 416543, 9 pages
http://dx.doi.org/10.1155/2014/416543

Research Article

Bioinformatic Prediction of WSSV-Host Protein-Protein Interaction

Zheng Sun,[1] Shihao Li,[1,2] Fuhua Li,[1,2] and Jianhai Xiang[1]

[1] *Key Laboratory of Experimental Marine Biology, Institute of Oceanology, Chinese Academy of Sciences, 7 Nanhai Road, Qingdao 266071, China*
[2] *National & Local Joint Engineering Laboratory of Ecological Mariculture, 7 Nanhai Road, Qingdao 266071, China*

Correspondence should be addressed to Fuhua Li; fhli@qdio.ac.cn and Jianhai Xiang; jhxiang@qdio.ac.cn

Received 5 February 2014; Revised 22 April 2014; Accepted 6 May 2014; Published 19 May 2014

Academic Editor: Wen-Chi Chang

WSSV is one of the most dangerous pathogens in shrimp aquaculture. However, the molecular mechanism of how WSSV interacts with shrimp is still not very clear. In the present study, bioinformatic approaches were used to predict interactions between proteins from WSSV and shrimp. The genome data of WSSV (NC_003225.1) and the constructed transcriptome data of *F. chinensis* were used to screen potentially interacting proteins by searching in protein interaction databases, including STRING, Reactome, and DIP. Forty-four pairs of proteins were suggested to have interactions between WSSV and the shrimp. Gene ontology analysis revealed that 6 pairs of these interacting proteins were classified into "extracellular region" or "receptor complex" GO-terms. KEGG pathway analysis showed that they were involved in the "ECM-receptor interaction pathway." In the 6 pairs of interacting proteins, an envelope protein called "collagen-like protein" (WSSV-CLP) encoded by an early virus gene "wsv001" in WSSV interacted with 6 deduced proteins from the shrimp, including three integrin alpha (ITGA), two integrin beta (ITGB), and one syndecan (SDC). Sequence analysis on WSSV-CLP, ITGA, ITGB, and SDC revealed that they possessed the sequence features for protein-protein interactions. This study might provide new insights into the interaction mechanisms between WSSV and shrimp.

1. Introduction

WSSV is one of the most dangerous pathogens that are destructive to penaeid shrimp, which results in up to 100% mortality in commercial shrimp farms [1]. In order to find out feasible approaches dealing with the virus, more and more studies have been carried out in crustaceans in last decade. The transcriptional profile of WSSV genes in shrimp was detected by DNA microarray and some early genes were discovered [2]. Many host genes and proteins responding to WSSV infection were also identified through large scale approaches [3–7]. From these studies, a lot of host genes and proteins were found upregulated or downregulated after WSSV infection. However, evidence on the direct interaction between WSSV and the host proteins is still urgent for understanding the pathogenesis of WSSV in shrimp.

Previous studies have noticed the importance of genes and proteins involved in WSSV/shrimp interaction. The beta-integrin, a cell surface molecule, was found to be a possible cellular receptor for WSSV infection by interacting with WSSV envelope protein VP187 [8]. Neutralization analysis with antibodies revealed that the WSSV envelope proteins VP68, VP281, and VP466 played roles in WSSV infection in shrimp [9]. The activity of the immediate-early gene *ie1* of WSSV could be upregulated by shrimp NF-κB through binding to the promoter of *ie1* gene [10]. Although these data provide us some useful information about WSSV infection mechanism, it is still not very clear about the molecular mechanism of WSSV infection. At present, the whole genome of two different WSSV isolates has been sequenced, one of which is about 293 kb encoding 184 open reading frames (ORFs) [11] and another one is about 305 kb containing 181 ORFs [12]. Meanwhile, high-throughput data on the Chinese shrimp transcriptome has also been published, which contains 64,188,426 Illumina reads and isolates 46,676 unigene sequences [7]. Under this condition, bioinformatic analysis will provide a highly effective approach for identifying genes and proteins involved in WSSV/shrimp interaction based on the public protein-protein interaction (PPI) databases.

The most widely used PPI databases mainly include the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING), the Database of Interacting Proteins (DIP), and Reactome. STRING is a database of known and predicted protein interactions based on the sources derived from the genomic context, high-throughput experiments, coexpression, and previous knowledge [13]. DIP is a database that records experimentally proved PPIs and provides scientific community with a comprehensive and integrated tool for browsing and efficiently extracting information about protein interactions and interaction networks in biological processes [14]. Another database, Reactome, is a manually curated and peer-reviewed pathway database [15], providing pathway related PPI information. These databases are useful resources for analyzing PPIs.

In the present study, the PPIs between WSSV and shrimp were predicted through searching the databases of STRING, DIP, and Reactome by using the data of WSSV genome sequences downloaded from the GenBank and the transcriptome data of the Chinese shrimp *Fenneropenaeus chinensis* sequenced by our lab [7]. Forty-four pairs of PPIs between WSSV and the shrimp were totally predicted. Further analysis on PPIs between the WSSV envelope proteins and the shrimp membrane proteins was carried out and the WSSV collagen like protein (WSSV-CLP) was predicted interacting with integrin and syndecan protein of the shrimp.

2. Materials and Methods

2.1. Data Preparation. WSSV genome data (accession number: NC_003225) was downloaded from the GenBank (http://www.ncbi.nlm.nih.gov/genome) and called "WSVG" in the present study. The transcriptome data of the Chinese shrimp *Fenneropenaeus chinensis* was sequenced by our lab [7] and called "FCT" in this study. Three PPI databases were localized using related data downloaded from websites. The information of downloaded files and databases was listed in Table 1.

2.2. Bioinformatic Analyses

2.2.1. Screening of WSSV/Shrimp Interaction Proteins. Before screening, the BLAST program was downloaded for localization from the NCBI website (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/). The procedure of screening of WSSV/shrimp interaction proteins consisted of two steps. The first step was searching for similar sequences between WSVG data (or FCT data) and three PPI databases, including DIP (http://dip.doe-mbi.ucla.edu) database, Reactome (http://www.reactome.org/ReactomeGWT/entrypoint.html) database, and STRING (http://string-db.org/) database using the localized BLASTx program, respectively. The WSVG data and the FCT data were used as query sequences and the PPI databases were used as references (*E* value cutoff: $<10^{-5}$). The output data were designated as WSVG_STRING_SBJCT, FCT_STRING_SBJCT, WSVG_DIP_SBJCT, FCT_DIP_SBJCT, WSVG_Reactome_SBJCT, and FCT_Reactome_SBJCT, respectively. The second step was to predict the potential interacting proteins using similar sequences data generated in the first step. The interacting proteins predicted by STRING were identified after comparison of the WSVG_STRING_SBJCT and FCT_STRING_SBJCT in the STRING interaction relation table. Similarly, DIP and Reactome databases were used to predict the interacting proteins following the above described procedures. In addition, interactions between predicted WSSV proteins and whole WSSV proteins and predicted shrimp proteins and all shrimp proteins were analyzed using the above-mentioned methods.

2.2.2. Analysis of Predicted Interacting Proteins. Interacting proteins predicted in Section 2.2.1 were used for gene ontology (GO) analysis by Blast2GO (http://www.blast2go.org/). Subsequent pathway analysis was carried out using the online software KAAS-KEGG (http://www.genome.jp/tools/kaas/). Alignment analyses of the interested genes or proteins were performed by online software ClustalX (http://www.ebi.ac.uk/Tools/msa/clustalw2/), and functional domains and protein binding sites of interested proteins were analyzed by Conserved Domains Database (CDD) search program (http://www.ncbi.nlm.nih.gov/cdd). O-linked glycosylation sites in SDC protein were predicted using the online software NetOGlyc 3.1 (http://www.cbs.dtu.dk/services/NetOGlyc/).

3. Results

3.1. Prediction of WSSV/Shrimp Interacting Proteins. Forty-four pairs of WSSV/shrimp interacting proteins, including 38 pairs STRING predicted PPIs, 32 pairs DIP predicted PPIs, and 39 pairs Reactome predicted PPIs, were totally predicted after screening of the three PPI databases using WSVG and FCT data (Table 2). Seven deduced proteins encoded by WSSV genes, including wsv001 (collagen like protein), wsv026 (hypothetical protein), wsv067 (thymidylate synthetase), wsv112 (dUTP diphosphatase), wsv128 (hypothetical protein), wsv172 (Ribonucleoside-diphosphate reductase large chain), and wsv188 (ribonucleoside-diphosphate reductase small chain), probably interacted with 32 proteins encoded by the shrimp genes.

Possible interactions between predicted proteins and their endogenous proteins were also predicted. For predicted WSSV proteins, five out of seven predicted members showed interactions with other WSSV proteins. For predicted shrimp proteins, 31 out of 32 members showed interactions with other shrimp proteins (see Table S1 in Supplementary Material available online at http://dx.doi.org/10.1155/2014/416543).

3.2. Potential Interaction between WSSV Envelope Proteins and Shrimp Membrane Proteins. GO analysis (based on cellular components, level 2) showed that 5 WSSV proteins and 23 shrimp proteins had GO annotation (Table S2). Among them, a WSSV envelope protein "collagen-like protein" (WSSV-CLP, NP_477523.1) encoded by wsv001 belonged to the GO-term of "extracellular region." Six proteins from the shrimp, including three integrin alpha proteins (ITGAs) (s_12679, s_18988, and s_3390), two integrin beta proteins (ITGBs) (s_1537 and s_16763), and one syndecan protein (SDC)

Table 1: The downloaded information used for database construction in the present study.

Database	Download address	Definition	Data type	Name used in the present study
Reactome	http://www.uniprot.org/downloads	UniProtKB/TrEMBL	Reference sequence (Fasta)	Reactome gene database
	http://www.reactome.org/download/all_interactions.html	Drosophila melanogaster protein-protein interaction pairs	Interaction relation table	Reactome interaction relation table
DIP	http://dip.doe-mbi.ucla.edu/dip/Download.cgi?SM=4	fasta20120218.seq	Reference sequence (Fasta)	DIP gene database
	http://dip.doe-mbi.ucla.edu/dip/Download.cgi?SM=3	dip20120818.txt	Interaction relation table	DIP interaction relation table
STRING9.0	http://string-db.org/newstring_cgi/show_download_page.pl?UserId=DNt0ic8Ndem8&sessionId=heGmVXVt8w_a	protein.sequences.v9.0.fa.gz	Reference sequence (Fasta)	STRING gene database
		protein.actions.v9.0.txt.gz	Interaction relation table	STRING interaction relation table

Table 2: Predicted interacting proteins between WSSV and the Chinese shrimp.

WSSV_ID	Annotation	Interacting protein in FCT		Database		
		FCT_ID	Annotation	STRING	DIP	Reactome
wsv001	Collagen like protein	s_18988	Integrin alpha 4	/	/	√
		s_12679	Integrin alpha 5			
		s_3390	Integrin alpha 8			
		s_1537	Integrin beta 1			
		s_16763	Integrin beta 6			
		s_2496	Syndecan			
wsv026	hypothetical protein	s_5219	Putative methylase/helicase	√	/	√
wsv067	thymidylate synthetase	s_3110	N/A	√	√	√
		s_13286	N/A			
		s_2426	Glycine dehydrogenase			
		s_5707	Inosine triphosphatase			
		s_2204	Proliferating cell nuclear antigen			
		s_3879	Ribonucleotide reductase 1			
		s_1099	Thymidine kinase 1			
wsv112	dUTP diphosphatase	s_16017	Nucleoside diphosphate kinase	√	√	√
		s_20130	Thymidylate kinase			
		s_5707	Inosine triphosphatase			
		s_60	Nucleoside diphosphate kinase			
		s_3879	Ribonucleotide reductase 1			
		s_1099	Thymidine kinase 1			
wsv128	hypothetical protein	s_15702	Beta-transducin repeat containing protein	√	/	/
		s_3485	Hypothetical protein			
		s_5154	Hypothetical protein			
		s_6901	Hypothetical protein			
		s_13825	Hypothetical protein			
wsv172	Ribonucleoside-diphosphate reductase large chain	s_4560	Adenylate kinase	√	√	√
		s_6839	Minichromosome maintenance complex component 4			
		s_60	Nucleoside diphosphate kinase			
		s_18721	Cell division control protein 45 homolog			
		s_18944	Probable thymidylate synthase			
		s_4876	Guanylate kinase			
		s_2904	Probable uridylate kinase			
		s_1744	Ribonucleoside-diphosphate reductase small chain			
		s_20130	Thymidylate kinase			
wsv188	Ribonucleoside-diphosphate reductase small chain	s_5545	Adenylate kinase 3	√	√	√
		s_20130	Thymidylate kinase			
		s_16017	Nucleoside diphosphate kinase			
		s_2904	Probable uridylate kinase			
		s_386	Adenylate kinase isoenzyme			
		s_4876	Guanylate kinase			
		s_20711	Conserved hypothetical protein (*Aedes aegypti*)			
		s_60	Nucleoside diphosphate kinase			
		s_4560	Adenylate kinase			
		s_3879	Ribonucleotide reductase 1			

Note. "/" means absence of prediction in relevant database, while "√" means presence of prediction in relevant database.

Table 3: Endogenous proteins interacting with WSSV-CLP, ITGA, ITGB, and SDC.

Predicted gene ID	Predicted gene annotation	Gene ID of endogenous interacting proteins	Annotation of endogenous interacting proteins
s_12679	Integrin alpha 5	s_17653	fms-related tyrosine kinase 4
		s_25665	Fibronectin 1
		s_1537	Integrin beta 1
		s_1294	Talin 1
s_18988	Integrin alpha 4	s_21444	Insulin-like receptor
		s_1537	Integrin beta 1
		s_5849	Integrin alpha 1
		s_2170	Myospheroid
		s_22423	Ultraspiracle
s_3390	Integrin alpha 8	s_30884	Ecdysone receptor
		s_21444	Insulin-like receptor
		s_1537	Integrin beta 1
		s_5849	Integrin alpha 1
		s_2170	Myospheroid
s_16763	Integrin beta 6	s_23243	Fibronectin 1
		s_5849	Integrin alpha 1
		s_6902	Integrin alpha 4
		s_26063	Integrin alpha 5
s_1537	Integrin beta 1	s_8242	Integrin-linked kinase
		s_5849	Integrin alpha 1
		s_7013	Lysosomal-associated membrane protein 1
		s_11802	Lysosomal-associated membrane protein 2
		s_26063	Integrin alpha 5
s_2496	Syndecan	s_24189	CG3194
		s_3573	CG9298
		s_18063	Calcium/calmodulin-dependent protein kinase
		s_18648	Dally-like
		s_7954	Kon-tiki
wsv001	Collagen triple helix repeat protein	wsv188	Ribonucleotide reductase small subunit
		wsv172	Ribonucleotide reductase large subunit RR1

(s_2496) showed interactions with WSSV-CLP (Table 2). The six proteins were classified into the "extracellular region" or "receptor complex" GO-terms (Table S2). The subsequent KEGG analysis (Table S2) revealed that the interactions between WSSV-CLP and ITGA/ITGB/SDC were involved in the "ECM-receptor interaction" and "focal adhesion" pathways (Figure 1).

Further analysis on endogenous proteins interacting with WSSV-CLP or ITGA/ITGB/SDC was carried out. As shown in Table 3, WSSV-CLP could interact with two other WSSV proteins, including ribonucleotide reductase small subunit (wsv188) and ribonucleotide reductase large subunit RR1 (wsv172). In shrimp, 20 endogenous proteins were found interacting with ITGA, ITGB, or SDC. The endogenous proteins interacting with ITGA mainly included fms-related tyrosine kinase 4, fibronectin 1, integrin beta 1, talin 1, insulin-like receptor, integrin alpha 1, myospheroid, ultraspiracle, and ecdysone receptor (Table 3). Endogenous proteins interacting with ITGB contained fibronectin 1, integrin alpha 1, integrin alpha 4, integrin alpha 5, integrin-linked kinase, lysosomal-associated membrane protein 1, and Lysosomal-associated membrane protein 2 (Table 3). Endogenous proteins interacting with SDC were annotated as calcium/calmodulin-dependent protein kinase, dally-like, kon-tiki, and *Drosophila* transcripts CG3194 and CG9298 (Table 3).

3.3. Analysis on the Amino Acid Sequence of WSSV-CLP. The WSSV-CLP, encoding by the published WSSV collagen like

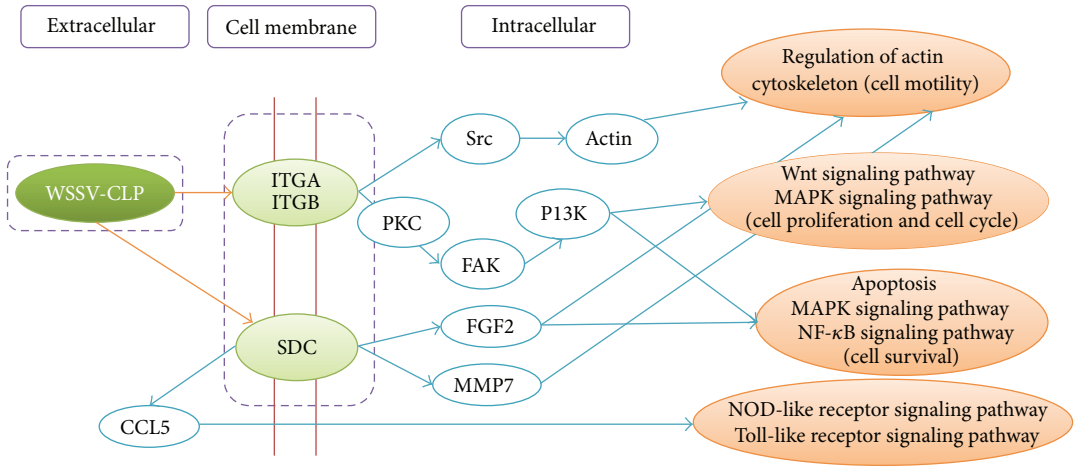


Figure 1: A schematic diagram was drawn to describe the predicted protein-protein interactions between WSSV-CLP and membrane proteins from the shrimp and the intracellular signaling pathways induced by the interactions. The diagram was drawn based on the KEGG pathway map04510 (focal adhesion) and map04512 (ECM-receptor interaction). WSSV-CLP: WSSV collagen like protein; ITGA: integrin alpha; ITGB: integrin beta; SDC: syndecan; CCL5: chemokine ligand 5; PKC: protein kinase C; FAK: focal adhesion kinase; PI3K: phosphoinositide 3-kinase; FGF2: fibroblast growth factor 2; MMP7: matrix metalloproteinase-7.

protein gene, was predicted possessing three collagen triple helix repeat domains. In the amino acid sequence of WSSV-CLP, the Gly-X-X (GXX, where X represents any amino acid) motifs were widely distributed from 161 aa to 1327 aa, where the GXXGER motif appeared for 21 times and the GXXGEN motif appeared for eight times (data not shown).

3.4. Sequence Analyses on ITGAs and ITGBs. Three ITGA homologs and two ITGB homologs were identified as WSSV-CLP interaction proteins. They were designated here as ITGA_4 (accession number: KC715736), ITGA_5 (accession number: KC715737), ITGA_8 (accession number: KC715738), ITGB_1 (accession number: KC715739), and ITGB_6 (accession number: KC715740) according to the BlastX annotation, respectively. Sequence analysis revealed that ITGA_5, ITGA_8, and ITGB_1 contained complete open reading frame (ORF), while ITGA_4 and ITGB_6 had partial ORF (see online submitted sequences). Alignment of above sequences by ClustalX showed that they shared poor similarity (data not shown). CDD analysis revealed that all the three ITGAs contained FG-GAP repeats in the N terminus (Figure 2(a)). In addition, ITGA_5 and ITGA_8 also possessed the integrin alpha 2 domain. ITGB_1 contained four conserved domains, including an integrin beta domain, an integrin beta tail domain (tail, pfam07965), an integrin beta cytoplasmic domain (cyt, pfam08725), and a conserved domain of eukaryotic metallothioneins family (Euk2, pfam12809) Metallothi_Euk2 (Figure 2(b)). In the Integrin beta domain, there is a βA domain (amino acids number 141–180), which is a member of the type A domain superfamily containing a prototype molecule called von Willebrand factor (vWF). The βA domain contained the conserved ligand-binding sites of DLSNS, DDK, and FGSFVD (Figure 2(b)).

3.5. Sequence Analysis on SDC. The SDC (accession number: KC733458) isolated from the FCT data contained a complete ORF and the deduced amino acid sequence had a signal peptide and five conserved domains, including ectodomain, transmembrane domain, C1 domain, C2 domain, and V domain (Figure 3), which showed the typical characteristics of syndecans. The C1 domain showed a conserved sequence feature of RMK(R)KKDEGSY, and the C2 domain had a conserved sequence feature of EF(I)YA. It showed high similarities in transmembrane domain, C1 domain and C2 domain among shrimp syndecan (FcSDC), and syndecan 1 from mammals (HsSDC1). However, the ectodomain and V domain showed low conservation among them. One serine residue and seven threonine residues were predicted in the C terminus of the ectodomain with the potential to be *O*-linked glycosylated. Three Ser-Gly (SG) sites, including YGSGD, EGSGH, and EGSGT, located in the N terminus of the ectodomain of FcSDC. In the ectodomain of HsSDC1, potential *O*-linked glycosylation sites mainly located in the T/S-rich region and four SG sites also existed (Figure 3).

4. Discussions

Exploitation of protein-protein interaction information with bioinformatic approaches provides an effective way to analyze high-throughput experimental data and has been widely applied in distinct organisms [16, 17]. With the genome data of WSSV [11, 12] and abundant transcriptome data from shrimp, it becomes possible to identify WSSV/shrimp interacting proteins with developed bioinformatic techniques. In the present study, we screened all possible WSSV/shrimp interacting proteins and a total of seven putative proteins encoded by WSSV were predicted interacting with deduced proteins from shrimp. Although only the WSSV-CLP was focused for further analysis, the other six putative proteins from WSSV also provided important information. Four of them were annotated as thymidylate synthetase, dUTP diphosphatase, ribonucleoside-diphosphate reductase

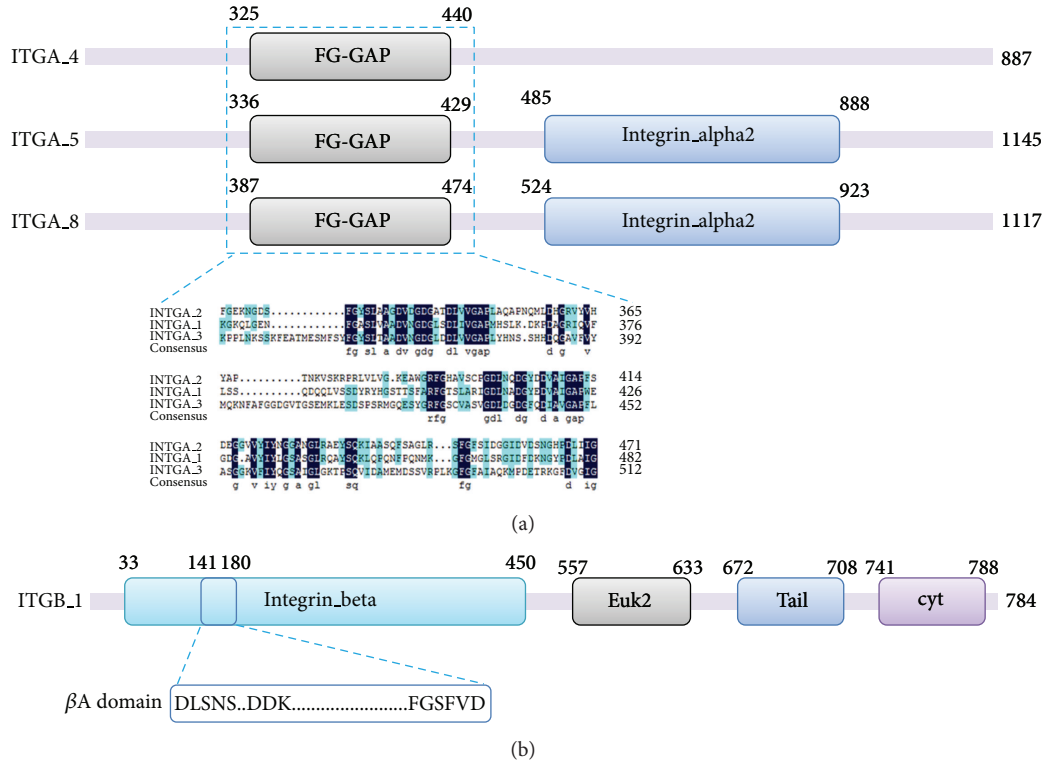


FIGURE 2: Schematic description of domains in ITGA (a) and ITGB (b). The FG-GAP domain, integrin alpha 2 domain, integrin beta domain, βA domain, EuK2 domain, tail (integrin beta tail) domain, and cyt (integrin beta cytoplasmic) domain located in ITGA and ITGB were shown with their beginning residue sites and stopping residue sites. The detailed sequence information of FG-GAP domain in ITGA_4, ITGA_5, and ITGA_8 were listed and aligned. The common feature of βA domain in ITGB_1 integrin beta domain was shown in a green box.

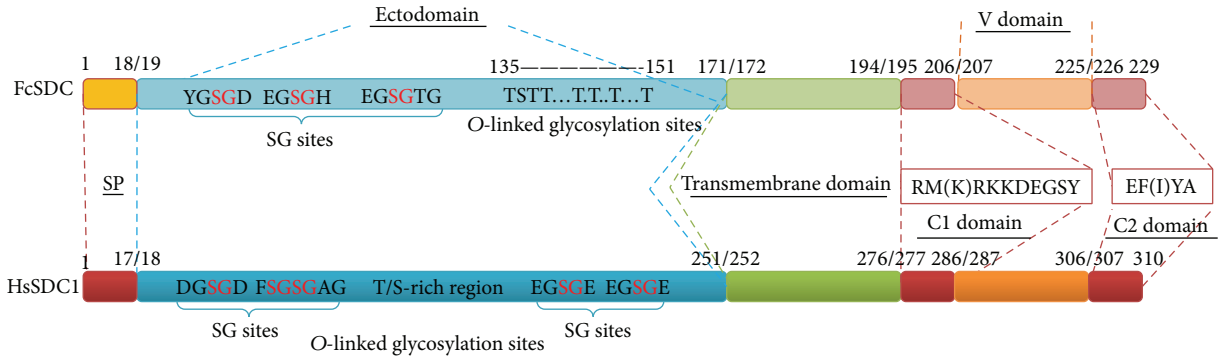


FIGURE 3: A compared schematic structure of syndecan from shrimp (FcSDC) and syndecan-1 (accession number: NP_001006947) from *Homo sapiens* (HsSDC1). The signal peptide (SP), syndecan conserved domains (transmembrane domain, C1 domain and C2 domain), and syndecan-type specific domains (ectodomain and V domain) were displayed with the number of their boundary amino acid residues. The potent glycosylation sites, including SG sites and predicted *O*-linked glycosylation sites, were shown. The sequence information of C1 domain and C2 domain was also listed.

large chain, and ribonucleoside-diphosphate reductase small chain, respectively. Thymidylate synthase [18] and dUTP diphosphatase [19] could generate direct or indirect substrates for DNA synthesis and reduce the risk of DNA repair. Ribonucleoside-diphosphate reductase is also an important enzyme for DNA replication [20–22]. As WSSV original proteins, these enzymes might play the key roles during WSSV genome replication. Identification of host proteins interacting with these enzymes provided new clues to develop disease control techniques.

The invasion route that WSSV infects the host cells is the point that we are interested in. After GO classification and pathway analysis, the previously reported WSSV envelope protein WSSV-CLP was predicted interacting with ITGA, ITGB, and SDC from the shrimp. Endogenous proteins which might interact with WSSV-CLP, ITGA, ITGB, and SDC were

also screened, which provided new insights into how WSSV interact with host cells. In the present study, WSSV-CLP and its interacting proteins in shrimp were further analyzed. The WSSV-CLP was first reported after whole genome sequencing [12] and then described as an early viral gene encoding an envelope protein [23]. However, there is still no report revealing WSSV-CLP interaction proteins from the hosts. In mammalian, monoclonal antibodies of $\alpha 2\beta 1$ integrin could block the collagen-induced morphogenesis of human mammary epithelial cells [24]. The influence might be caused by interaction between $\alpha 2\beta 1$ integrin and specific sites of collagen because the $\alpha 2\beta 1$ integrin was a cell-surface receptor for collagen [25] and the synthetic collagen-mimetic peptides, GXXGER and GXXGEN, showed binding affinities with $\alpha 2\beta 1$ integrin [26, 27]. The collagen binding site of $\alpha 2\beta 1$ integrin was localized to the $\alpha 2$ von Willebrand factor type A (vWFA) domain [28]. The reported β-integrin in shrimp could also bind WSSV and the recombinant extracellular region of the β-integrin could partially block WSSV infection to shrimp [29]. This binding was deemed as an interaction of β-integrin with RGD motif presented in WSSV proteins [29] and previous study revealed that an RGD motif containing protein in WSSV, VP187 encoded by wsv209, was a potent ligand of β-integrin in shrimp [8]. In the present study, the WSSV-CLP was predicted interacting with ITGA and ITGB from the shrimp. Sequence analysis revealed that WSSV-CLP contained 21 GXXGER motifs and 8 GXXGEN motifs in the collagen triple helix repeat domain. These motifs might be the binding sites of WSSV-CLP with integrin according to the previous studies in mammalians. The FG-GAP repeats found in the N terminus of ITGA chains have been shown to be important for ligands binding [30], such as collagens, fibronectins, fibrinogen, and laminins [31]. A βA domain existed in the integrin beta domain of ITGB_1. It was a RGD binding site that contained the conserved ligand-binding sites of DLSNS, DDK, and FGSFVD and could be recognized by VP187 protein of WSSV [8]. These data might indicate that ITGA_5 and ITGB_1 are possible receptors of WSSV in shrimp. The function of ITGAs and ITGBs in WSSV infection route is worthy to be investigated further.

Syndecan was another predicted protein in shrimp which could interact with WSSV-CLP. Synergistic interaction between syndecan and integrin in cell adhesion [32] and cell spreading [33] was already reported. Furthermore, syndecan-1 in mammals could support the integrin $\alpha 2\beta 1$-mediated adhesion to collagen [34]. The SDC isolated from shrimp displayed a similar sequence feature with syndecan-1 from *Homo sapiens*. The SG sites (YGSGD, EGSGH, and EGSGT) presented in the N terminus of SDC ectodomain shared great sequence and location similarities with the SG sites in syndecan-1 from *Homo sapiens*, which possessed the SG sites such as FSGSGTG, DGSGD, EGSGE, and ETSGE responsible for HS or CS chains formation [35]. In addition, the predicted *O*-linked glycosylation sites located in the C terminus of SDC ectodomain also provided possible glycosylation sites to generate HS or CS chains. These characteristics were probably responsible for WSSV-CLP binding during infection. In the present study, both integrin and SDC were predicted having interaction with WSSV-CLP. Integrin and SDC are regarded as membrane receptors. Interaction of WSSV-CLP with integrin or SDC might lead to the regulation of actin for cell motility or initiate the intracellular signaling pathways, such as MAPK, NF-kappa B signaling pathway, which are responsive to WSSV challenge in shrimp [35–38].

In conclusion, the present study identified the WSSV/shrimp interacting proteins by bioinformatic analysis on the high-throughput gene data. The predicted interactions between WSSV-CLP and integrins and between WSSV-CLP and SDC, which might be either independent interaction or synergistic interaction, provided possible invasion approaches during WSSV infection to host cells. Moreover, these interactions could also lead to intracellular signaling pathways initiated by integrins or SDC as described in Figure 1. Further experimental confirmation is necessary for the prediction results in the future.

Conflict of Interests

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

Authors' Contribution

Zheng Sun and Shihao Li contributed equally to this work.

Acknowledgments

This work was financially supported by the Major State Basic Research Development Program (2012CB114403), National High Tech Research and Development Program (2012AA10A404, 2012AA092205), National Natural Science Foundation Program (31072203), China Agriculture Research System-47 (CARS-47), and Special Fund for Agro-scientific Research in the Public Interest (201103034).

References

[1] H. Liu, K. Söderhäll, and P. Jiravanichpaisal, "Antiviral immunity in crustaceans," *Fish and Shellfish Immunology*, vol. 27, no. 2, pp. 79–88, 2009.

[2] Y. Lan, X. Xu, F. Yang, and X. Zhang, "Transcriptional profile of shrimp white spot syndrome virus (WSSV) genes with DNA microarray," *Archives of Virology*, vol. 151, no. 9, pp. 1723–1733, 2006.

[3] H. Liu, R. Chen, Q. Zhang, H. Peng, and K. Wang, "Differential gene expression profile from haematopoietic tissue stem cells of red claw crayfish, *Cherax quadricarinatus*, in response to WSSV infection," *Developmental and Comparative Immunology*, vol. 35, no. 7, pp. 716–724, 2011.

[4] B. Wang, F. Li, B. Dong, X. Zhang, C. Zhang, and J. Xiang, "Discovery of the genes in response to white spot syndrome virus (WSSV) infection in *Fenneropenaeus chinensis* through cDNA microarray," *Marine Biotechnology*, vol. 8, no. 5, pp. 491–500, 2006.

[5] H. Wang, H. Wang, J. Leu, G. Kou, A. H.-J. Wang, and C. Lo, "Protein expression profiling of the shrimp cellular response to white spot syndrome virus infection," *Developmental and Comparative Immunology*, vol. 31, no. 7, pp. 672–686, 2007.

[6] Z. Zhao, Z. Yin, S. Weng et al., "Profiling of differentially expressed genes in hepatopancreas of white spot syndrome virus-resistant shrimp (*Litopenaeus vannamei*) by suppression subtractive hybridisation," *Fish and Shellfish Immunology*, vol. 22, no. 5, pp. 520–534, 2007.

[7] S. H. Li, X. J. Zhang, Z. Sun, F. H. Li, and J. H. Xiang, "Transcriptome analysis on Chinese shrimp *Fenneropenaeus chinensis* during WSSV acute infection," *Plos ONE*, vol. 8, Article ID e58627, 2013.

[8] D. Li, M. Zhang, H. Yang, Y. Zhu, and X. Xu, "β-integrin mediates WSSV infection," *Virology*, vol. 368, no. 1, pp. 122–132, 2007.

[9] W. Wu, L. Wang, and X. Zhang, "Identification of white spot syndrome virus (WSSV) envelope proteins involved in shrimp infection," *Virology*, vol. 332, no. 2, pp. 578–583, 2005.

[10] X. Huang, L. Zhao, H. Zhang et al., "Shrimp NF-κB binds to the immediate-early gene ie1 promoter of white spot syndrome virus and upregulates its activity," *Virology*, vol. 406, no. 2, pp. 176–180, 2010.

[11] M. C. W. van Hulten, J. Witteveldt, S. Peters et al., "The white spot syndrome virus DNA genome sequence," *Virology*, vol. 286, no. 1, pp. 7–22, 2001.

[12] F. Yang, J. He, X. Lin et al., "Complete genome sequence of the shrimp white spot bacilliform virus," *Journal of Virology*, vol. 75, no. 23, pp. 11811–11820, 2001.

[13] 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, no. 1, pp. D561–D568, 2011.

[14] I. Xenarios, Ł. Salwínski, X. J. Duan, P. Higney, S. Kim, and D. Eisenberg, "DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions," *Nucleic Acids Research*, vol. 30, no. 1, pp. 303–305, 2002.

[15] G. Joshi-Tope, M. Gillespie, I. Vastrik et al., "Reactome: a knowledgebase of biological pathways," *Nucleic Acids Research*, vol. 33, pp. D428–D432, 2005.

[16] K. A. Pattin and J. H. Moore, "Role for protein-protein interaction databases in human genetics," *Expert Review of Proteomics*, vol. 6, no. 6, pp. 647–659, 2009.

[17] N. Remmerie, T. de Vijlder, K. Laukens et al., "Next generation functional proteomics in non-model plants: a survey on techniques and applications for the analysis of protein complexes and post-translational modifications," *Phytochemistry*, vol. 72, no. 10, pp. 1192–1218, 2011.

[18] S. Kaneda, J. Nalbantoglu, K. Takeishi et al., "Structural and functional analysis of the human thymidylate synthase gene," *The Journal of Biological Chemistry*, vol. 265, no. 33, pp. 20277–20284, 1990.

[19] B. G. Vértessy and J. Tóth, "Keeping uracil out of DNA: physiological role, structure and catalytic mechanism of dUTPases," *Accounts of Chemical Research*, vol. 42, no. 1, pp. 97–106, 2009.

[20] D. Filpula and J. A. Fuchs, "Regulation of the synthesis of ribonucleoside diphosphate reductase in *Escherichia coli*: specific activity of the enzyme in relationship to perturbations of DNA replication," *Journal of Bacteriology*, vol. 135, no. 2, pp. 429–435, 1978.

[21] E. C. Guzmán, J. L. Caballero, and A. Jiménez-Sánchez, "Ribonucleoside diphosphate reductase is a component of the replication hyperstructure in *Escherichia coli*," *Molecular Microbiology*, vol. 43, no. 2, pp. 487–495, 2002.

[22] M. A. Sánchez-Romero, F. Molina, and A. Jiménez-Sánchez, "Organization of ribonucleoside diphosphate reductase during multifork chromosome replication in *Escherichia coli*," *Microbiology*, vol. 157, no. 8, pp. 2220–2225, 2011.

[23] Q. Li, Y. Chen, and F. Yang, "Identification of a collagen-like protein gene from white spot syndrome virus," *Archives of Virology*, vol. 149, no. 2, pp. 215–223, 2004.

[24] N. Berdichevsky, C. Gilbert, M. Shearer, and J. Taylor-Papadimitriou, "Collagen-induced rapid morphogenesis of human mammary epithelial cells: the role of the $\alpha 2\beta 1$ integrin," *Journal of Cell Science*, vol. 102, no. 3, pp. 437–446, 1992.

[25] M. M. Zutter and S. A. Santoro, "Widespread histologic distribution of the $\alpha 2\beta 1$ integrin cell-surface collagen receptor," *American Journal of Pathology*, vol. 137, no. 1, pp. 113–120, 1990.

[26] N. Raynal, S. W. Hamaia, P. R.-M. Siljander et al., "Use of synthetic peptides to locate novel integrin α (2)β(1)-binding motifs in human collagen III," *The Journal of Biological Chemistry*, vol. 281, no. 7, pp. 3821–3831, 2006.

[27] I. C. A. Munnix, K. Gilio, P. R. M. Siljander et al., "Collagen-mimetic peptides mediate flow-dependent thrombus formation by high- or low-affinity binding of integrin $\alpha 2\beta 1$ and glycoprotein VI," *Journal of Thrombosis and Haemostasis*, vol. 6, no. 12, pp. 2132–2142, 2008.

[28] A. Aquilina, M. Korda, J. M. Bergelson, M. J. Humphries, R. W. Farndale, and D. Tuckwell, "A novel gain-of-function mutation of the integrin $\alpha 2$ VWFA domain," *European Journal of Biochemistry*, vol. 269, no. 4, pp. 1136–1144, 2002.

[29] X. Q. Tang, X. L. Wang, and W. B. Zhan, "An integrin beta subunit of Chinese shrimp *Fenneropenaeus chinensis* involved in WSSV infection," *Aquaculture*, vol. 368, pp. 1–9, 2012.

[30] T. A. Springer, "Folding of the N-terminal, ligand-binding region of integrin alpha-subunits into a beta-propeller domain," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 94, pp. 65–72, 1997.

[31] E. F. Plow, T. A. Haas, L. Zhang, J. Loftus, and J. W. Smith, "Ligand binding to integrins," *The Journal of Biological Chemistry*, vol. 275, no. 29, pp. 21785–21788, 2000.

[32] M. R. Morgan, M. J. Humphries, and M. D. Bass, "Synergistic control of cell adhesion by integrins and syndecans," *Nature Reviews Molecular Cell Biology*, vol. 8, no. 12, pp. 957–969, 2007.

[33] D. M. Beauvais and A. C. Rapraeger, "Syndecan-1-mediated cell spreading requires signaling by $\alpha v\beta 3$ integrins in human breast carcinoma cells," *Experimental Cell Research*, vol. 286, no. 2, pp. 219–232, 2003.

[34] K. Vuoriluoto, J. Jokinen, K. Kallio, M. Salmivirta, J. Heino, and J. Ivaska, "Syndecan-1 supports integrin $\alpha 2\beta 1$-mediated adhesion to collagen," *Experimental Cell Research*, vol. 314, no. 18, pp. 3369–3381, 2008.

[35] R. Kokenyesi and M. Bernfield, "Core protein structure and sequence determine the site and presence of heparan sulfate and chondroitin sulfate on syndecan-1," *The Journal of Biological Chemistry*, vol. 269, no. 16, pp. 12304–12309, 1994.

[36] F. Li, D. Wang, S. Li et al., "A Dorsal homolog (FcDorsal) in the Chinese shrimp *Fenneropenaeus chinensis* is responsive to both bacteria and WSSV challenge," *Developmental and Comparative Immunology*, vol. 34, no. 8, pp. 874–883, 2010.

[37] P. Wang, Z. Gu, D. Wan et al., "The shrimp NF-κB pathway is activated by white spot syndrome virus (WSSV) 449 to facilitate the expression of WSSV069 (ie1), WSSV303 and WSSV371," *PLoS ONE*, vol. 6, no. 9, Article ID e24773, 2011.

[38] H. Shi, X. F. Yan, L. W. Ruan, and X. Xu, "A novel JNK from *Litopenaeus vannamei* involved in white spot syndrome virus infection," *Developmental and Comparative Immunology*, vol. 37, pp. 421–428, 2012.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 393280, 7 pages
http://dx.doi.org/10.1155/2014/393280

Research Article

High-Dimensional Additive Hazards Regression for Oral Squamous Cell Carcinoma Using Microarray Data: A Comparative Study

Omid Hamidi,[1] Lily Tapak,[2] Aarefeh Jafarzadeh Kohneloo,[3] and Majid Sadeghifar[4]

[1] *Department of Science, Hamadan University of Technology, Hamedan 65155, Iran*
[2] *Department of Biostatistics and Epidemiology, School of Public Health, Hamadan University of Medical Sciences, Hamadan 6517838695, Iran*
[3] *Department of Epidemiology & Biostatistics, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran*
[4] *Department of Statistics, Faculty of Science, Bu-Ali Sina University, Hamadan 6517838695, Iran*

Correspondence should be addressed to Lily Tapak; l.tapak@umsha.ac.ir

Received 27 February 2014; Revised 22 April 2014; Accepted 6 May 2014; Published 19 May 2014

Academic Editor: Li-Ching Wu

Microarray technology results in high-dimensional and low-sample size data sets. Therefore, fitting sparse models is substantial because only a small number of influential genes can reliably be identified. A number of variable selection approaches have been proposed for high-dimensional time-to-event data based on Cox proportional hazards where censoring is present. The present study applied three sparse variable selection techniques of Lasso, smoothly clipped absolute deviation and the smooth integration of counting, and absolute deviation for gene expression survival time data using the additive risk model which is adopted when the absolute effects of multiple predictors on the hazard function are of interest. The performances of used techniques were evaluated by time dependent ROC curve and bootstrap .632+ prediction error curves. The selected genes by all methods were highly significant ($P < 0.001$). The Lasso showed maximum median of area under ROC curve over time (0.95) and smoothly clipped absolute deviation showed the lowest prediction error (0.105). It was observed that the selected genes by all methods improved the prediction of purely clinical model indicating the valuable information containing in the microarray features. So it was concluded that used approaches can satisfactorily predict survival based on selected gene expression measurements.

1. Introduction

Head and neck squamous cell carcinoma (HNSCC) is the sixth most common cancer worldwide that has variation across subsites (oral cavity, oropharynx, larynx, or hypopharynx) in many characteristics including age, sex, ethnicity, histologic grade, treatment modality, and prognosis [1–3]. HNSCC as the most common smoking related cancer after lung cancer is one of the most aggressive malignancies in human population [1, 3] and causes about 438,000 smoking-attributable mortalities in the world each year. The most common anatomic site of HNSCC counting for approximately 50% of all HNSCC is oral squamous cell carcinoma (OSCC) which contains two most commonly diagnosed oral premalignant lesions (OPL) including Leukoplakia and erythroplakia [3].

Although tremendous efforts were committed to early detection, prevention, and treatment during the last decades, HNSCC prognosis remains very poor with the rising incidence in developed countries and younger population [1]. Even for HNSCC diagnosed at early stages, surgery (current standard care) is a debilitating, substantially morbid procedure that leads to severe impairments in quality of life for the patients [3]. Therefore, developing new approaches including diagnosis of the disease before the cancerous stage and preventing development of invasive cancers such as HNSCC

is very substantial. The HNSCC is a multistep process and genetic factors play an important role in its etiology [1]. Generally, it is well accepted that malignant tumors, on the molecular level, are disease of genes [1].

The rapid development of biotechniques during the past decade has brought in a wealth of biomedical data including DNA microarrays which can be used to measure the expression of thousands of genes in a sample of cells or to identify hundreds of thousands of single nucleotide polymorphisms for an individual at the same time [4]. This kind of information leads to high-dimensional and low-sample size (HDLSS) data sets (i.e., $p \gg n$ where p and n are the number of covariates and patients) which pose tremendous challenges to effective statistical inference especially for the time-to-event due to the presence of censoring and the use of much more complicated models. In this case, a number of the outcome variables are typically censored. Besides, establishing link between high-dimensional biomedical data and patient's survivals would be informative in the presence of clinical data.

In this regard, a fundamental objective is to identify a small set of genes from a huge number of features whose expression levels are significantly correlated with a given clinical outcome such as cancer. Identified genes are then often used to create a predictive model for conducting prediction of outcome for new patients [5].

Recently, several variable selection techniques based on the maximization of a penalized likelihood have been proposed for HDLSS time-to-event data under the Cox proportional hazards model. Some of the most common penalization methods are the least absolute shrinkage and selection operator (lasso) [6], smoothly clipped absolute deviation (SCAD) [7], Dantzig selector [8], LARS [9], and the smooth integration of counting and absolute deviation (SICA) penalty [10]. These techniques can yield sparse models and hence perform simultaneous variable selection and estimation. To select ideal penalty functions for model selection Fan and Li [7] advocated penalty functions giving rise to estimators with three desired properties of sparsity, unbiasedness, and continuity. This class of penalty functions has been studied by Lv and Fan [11] and Fan and Lv [12] in (generalized) linear models contexts. The reasonable performance of penalization techniques such as Lasso, SCAD, and SICA was also investigated by Lin and Lv [4].

Despite the extensive study of the proportional hazards model for survival data with high-dimensional covariates, a few authors have utilized variable selection methods based on additive risk model which may provide more insights beyond the proportional hazards model analysis [4, 13]. Additive models assume that the covariate effects under consideration contribute additively to the conditional hazard [13].

The aim of the present study was to compare three sparse variable selection methods of Lasso, SICA, and SCAD for high-dimensional low-sample size data using an additive hazard approach to predict survival time in patients with OSCC and to determine the influential genes on survival time.

2. Methods

2.1. Data Source. The proposed techniques are illustrated with publicly available microarray data from patients with OPL [3]. The dataset consists of gene expression measurements for 29096 genes and survival outcomes to develop oral cancer on 86 patients. These 86 patients were selected from the 162 patients who were involved in a chemoprevention trial. The number of 35 patients who developed oral cancer during the study period was selected for gene expression profiling. Besides, 51 samples (ad hoc choice) from patients who did not develop OSCC were randomly selected among 106 patients. The median follow-up time was 5 years for this sample. The data analyzed in this study is available from the NIH Gene Expression Omnibus database at www.ncbi.nlm.nih.gov/geo under the accession number GSE26549. Potentially important clinical covariates were age, histology at baseline, deltaNp63, and podoplanin expression at baseline.

2.2. Variable Selection via Additive Hazards Model. Regularization techniques are particularly useful for variable selection in high-dimensional setting where the number of variables is much greater than the sample size and have gained increasing popularity. These estimation techniques pose a penalty term on the coefficients in objective function and shrink the estimates of the coefficients towards zero relative to the maximum likelihood estimates. The goal of this shrinkage is to prevent overfitting which arises due to either collinearity of the covariates or high-dimensionality [14]. When the outcome is survival time, the objective function is usually written based on hazards at the failure time and posing penalties on the coefficients can yield sparse models and hence perform simultaneous variable selection and estimation.

Consider a sample of n independent observations. Let T_i $(i = 1, \ldots, n)$ be the time to an event like death from cancer for the ith subject and conditionally independent of the censoring time C_i, given the p-dimensional possibly time-dependent covariate vector $Z_i = (Z_{i1}, Z_{i2}, \ldots, Z_{ip})^T$. Let $X_i = \min(T_i, C_i)$ and $\Delta_i = I(T_i \leq C_i)$ for right censored data, where $I(\cdot)$ is the indicator function. Then the observed data consists of (X_i, Δ_i, Z_i). The hazard function of a failure time T given a p-vector of possibly time-dependent covariates Z based on Lin and Ying additive hazards model is as follows:

$$\lambda(t; Z) = \lambda_0(t) + \beta_0^T Z, \tag{1}$$

where λ_0 is an unknown baseline hazard function which is common to all subjects and β_0 is a p-vector of regression coefficients [4, 15]. The pseudo score linear in $\beta \in \mathbb{R}^p$ function for Lin and Ying [16] model can be defined as follows:

$$\mathrm{U}(\beta) = \frac{1}{n} \sum_{i=1}^{n} \int_0^{\tau} \left\{ Z_i - \overline{Z} \right\} \left\{ dN_i(t) - Y_i(t) Z_i dt \right\}, \tag{2}$$

where $N_i(t) = I(X_i \leq t, \Delta_i \varepsilon_i = k)$ is the observed-failure counting process, $Y_i(t) = I(X_i \leq t)$ is at-risk indicator,

$\overline{Z} = \sum_{j=1}^{n} Y_j(t)Z_j / \sum_{j=1}^{n} Y_j(t)$, and τ is the maximum follow-up time. Performing some algebraic manipulation results in

$$U(\beta) = b - V\beta, \tag{3}$$

where $b = (1/n)\sum_{i=1}^{n} \int_0^{\tau} \{Z_i - \overline{Z}\} dN_i(t)$ and $V = (1/n)\sum_{i=1}^{n} \int_0^{\tau} Y_i(t)\{(Z_i - \overline{Z})(Z_i - \overline{Z})^T\} dt$. Because V is positive semidefinite, integrating $-U(\beta)$ with respect to β results in the least squares type loss function as

$$L(\beta) = \frac{1}{2}\beta^T V\beta - b^T\beta. \tag{4}$$

Then, the penalized estimator $\widehat{\beta}$ is a solution to the regularization problem

$$\widehat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \left\{ Q(\beta) \equiv L(\beta) + \sum_{j=1}^{p} p_\lambda\left(\left|\beta_j\right|\right) \right\}, \tag{5}$$

where $L(\beta)$ is the likelihood of beta for additive model, $p_\lambda(\theta)$, $\theta \geq 0$ is a penalty function based on the regularization parameter $\lambda \geq 0$ and is often rewritten as $p_\lambda(\cdot) = \lambda\rho(\cdot)$ [4].

The present study considered three commonly used sparse penalty functions which correspond to Lasso, SCAD, and SICA. In this regard, the Lasso uses the L_1-penalty; that is, $\rho(\theta) = \theta$. The SCAD penalty is given by the derivative $\rho_\lambda'(\theta) = I(\theta \leq \lambda) + ((a\lambda - \theta)_+/(a-1)\lambda)I(\theta > \lambda)$ with some $a > 2$ as a shape parameter and SICA penalty takes the form $\rho(\theta) = (a+1)\theta/(a+\theta)$ and $a > 0$ is a shape parameter. Estimation of $\widehat{\beta}$ is then accomplished via the *coordinate descent algorithm* [4].

Selecting the optimal regularization parameter λ is conducted through 10-fold cross-validation based on the following cross-validation score function:

$$CV(\lambda) = \frac{1}{10}\sum_{m=1}^{10} L^{(m)}\left(\widehat{\beta}^{(-m)}(\lambda)\right), \tag{6}$$

where $L^{(m)}(\cdot)$ is the least squares type loss function computed from the mth part of the data, and $\widehat{\beta}^{(-m)}(\lambda)$ is the estimate from the data with the mth part removed [4].

The SCAD and SICA have one additional tuning parameter a. For the SCAD penalty, $a = 3.7$ was suggested by Fan and Li [7] from a Bayesian perspective. Selecting a for the SICA penalty requires a little more caution, because small values of a that often needs to yield a superior theoretical performance may sometimes lead to the computational instability [4]. This study used the method proposed by Lin and Lv [4] to determine a in SICA.

2.3. Properties of Used Penalties. In spite of choosing the penalty function, the performance of the regularized estimators depends on various factors, such as the dimensionality of the model, the correlation among the variables, and the choice of the regularization parameter. Also, penalization techniques may suffer from some drawbacks.

Although the Lasso method enjoys the advantage of computational simplicity, it suffers from several shortcomings. All coefficients are shrunken toward zero by Lasso. So, the large elements of coefficients tend to be underestimated. The SCAD penalty is proposed to eliminate the bias caused by the Lasso and does not have this drawback. Also, it has been shown to enjoy the oracle property. It means that the resulting estimator performs asymptotically as well as the oracle estimator which knew the true sparse model in advance [4]. SCAD does not require the irrepresentable condition to consistency. In spite of nonconvexity of SCAD, algorithms are available to compute its solution [7, 17]. On the other hand Lv and Fan showed that the regularity conditions needed for the Lasso can be substantially relaxed by using concave penalties due to nonasymptotic weak oracle property [11]. The SICA family proposed in their study has the noticeable feature that it can be used to define a sequence of regularization problems with varying theoretical performance and computational complexity [4].

2.4. Predictive Performance. Assessment of the performance of three variable selection methods was performed by using time-dependent receiver operator characteristic (ROC) curves [18] and bootstrap .632+ prediction error curves [19] which were used to evaluate the performance of the models. The prediction error was obtained as squared difference between the true state (0 for being still under risk and 1 if an event of cancer occurred) at time t and predicted survival probability. Lower prediction errors suggest better performance.

2.5. Software. Analyses were performed by using the R software programming (http://www.r-project.org) based on a publically available R package which has been provided by Lin and Lv [4] (http://www-scf.usc.edu/~linwei/software.html). In addition, evaluating the predictive performance of used methods was utilized using "pec," "peperr," and "survAUC" R packages. Besides, to select the most effective genes, "timeROC" package was utilized.

3. Results

Three variable selection techniques of Lasso, SCAD, and SICA were implemented on microarray gene expression data of 86 leukoplakia samples of OPL patients based on additive hazards model. The frequency of occurrences of the genes along with means of coefficients and standard errors over 100 replicates, obtained from three techniques, was shown in Table 1. As shown in Table 1, the number of selected genes by Lasso was greater than SCAD and SICA. The SICA selected only a small set of genes due to the sparseness. In addition, there are seven common genes (7905589, 7908407, 8106919, 8126931, 8161169, 8174970, and 8180388) among three variable selection techniques.

In order to assess predictive performance, the median area under ROC curve over time (AUC(t)) was computed and plotted based on 69 (~80%) training and 17 (~20%) test sets for each method. Results are shown in Figure 1.

Table 1: Influential genes on OSCC patients' survival based on additive hazards model using Lasso, SCAD and SICA. Values are frequency of occurrences of the genes, means of coefficients (standard errors) over 100 replicates.

Probeset ID	Lasso		SCAD		SICA	
	Frequency	Coefficient (SE)	Frequency	Coefficient (SE)	Frequency	Coefficient (SE)
7897663	100	0.141 (0.006)				
7905589	100	−0.122 (0.032)	97	−0.182 (0.086)	88	−0.324 (0.132)
7908407	100	−0.639 (0.054)	98	−1.239 (0.645)	92	−1.367 (0.460)
7916489	99	−0.012 (0.009)	56	−0.022 (0.021)		
7918825	99	0.112 (0.041)				
7919157	81	0.081 (0.049)				
7922793	99	−0.069 (0.035)	48	−0.012 (0.017)		
7925161					1	−0.002 (0.017)
7946565			44	0.009 (0.010)		
7964627	91	0.102 (0.056)			12	0.030 (0.084)
7965467	99	−0.100 (0.050)	56	−0.034 (0.051)		
7971191	91	0.037 (0.019)	30	0.003 (0.007)		
7978754	100	−0.136 (0.015)	50	−0.005 (0.008)		
7981968	100	−0.086 (0.021)				
7982129	99	−0.012 (0.005)	56	−0.002 (0.003)		
8002247	91	0.100 (0.052)			2	0.002 (0.017)
8018097	100	−0.140 (0.008)	98	−0.060 (0.018)		
8020844	36	−0.011 (0.019)	94	−0.024 (0.012)		
8035398	75	−0.018 (0.012)				
8035829	64	0.018 (0.014)				
8040338	91	−0.016 (0.007)				
8044733	91	−0.039 (0.019)				
8047690			56	0.023 (0.028)		
8048595	91	−0.055 (0.030)	30	−0.005 (0.013)		
8065392	91	−0.168 (0.086)			1	−0.003 (0.029)
8076511	49	0.011 (0.011)				
8075691			26	−0.034 (0.058)		
8093764			56	−0.018 (0.020)		
8095441			56	−0.006 (0.006)		
8103368	100	−0.011 (0.003)	56	−0.009 (0.013)		
8106814	75	0.040 (0.029)	56	0.024 (0.033)		
8106919	81	0.070 (0.046)	97	−0.092 (0.028)	1	0.002 (0.020)
8109828	100	−0.128 (0.003)			47	−0.068 (0.082)
8110880			56	0.047 (0.041)		
8112916			68	0.005 (0.009)		
8120206	91	0.030 (0.015)				
8123338	75	−0.065 (0.045)			1	−0.002 (0.018)
8126931	100	0.157 (0.025)	84	0.022 (0.017)	2	0.002 (0.014)
8138531					90	0.873 (0.338)
8139808	99	−0.059 (0.016)				
8127993			30	−0.008 (0.014)		
8158952			84	0.900 (0.676)		
8161169	100	0.281 (0.029)	97	0.154 (0.041)	70	0.144 (0.175)
8174710	82	−0.023 (0.013)	24	−0.007 (0.012)		
8174970	100	0.360 (0.030)	94	0.060 (0.046)	35	0.101 (0.140)
8180388	100	−0.403 (0.039)	98	−0.139 (0.077)	12	−0.072 (0.199)

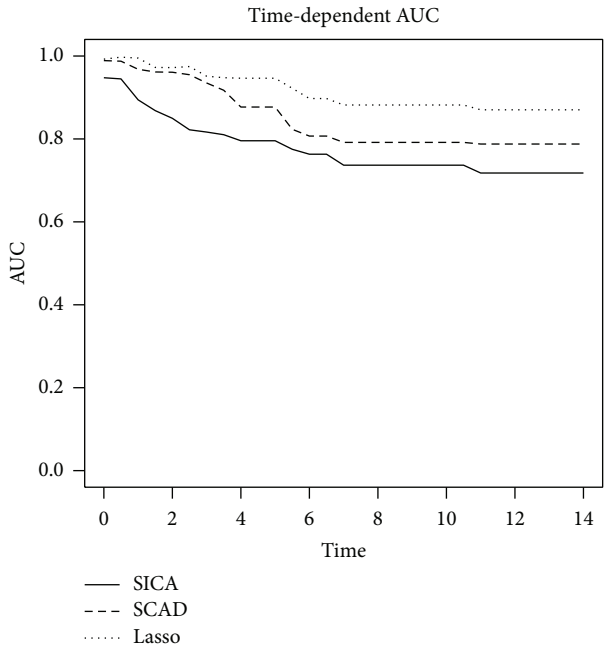


Figure 1: Comparison of predictive performance (area under the ROC curve, over time) for the OPL patients.

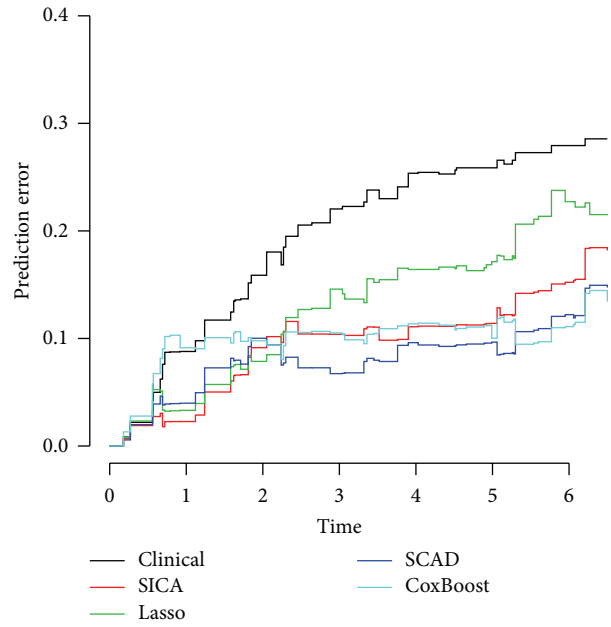


Figure 2: Model comparison using prediction error curves. Clinical model used age, histology, and podoplanin and deltaNp63 expression as predictors. The SICA, SCAD, and Lasso used selected microarray data as well as age, histology, and podoplanin and deltaNp63 expression as predictors.

For the follow-up period, the median AUC for Lasso, SCAD, and SICA was 0.95, 0.91, and 0.83, respectively. As can be seen from Figure 1, the predictive performance of the Lasso was superior to SCAD and SICA in this data analysis.

In addition, to evaluate prediction performance improvement by including selected genes over a purely clinical model, bootstrap .632+ prediction error curves were plotted based on $B = 100$ bootstrap samples drawn without replacement. Figure 2 shows the results. As shown, including selected microarray features in the models clearly improved prediction performance over the purely clinical model indicating the valuable information containing in the microarray features. Also, it can be seen that the performance of SCAD (bootstrap .632+ prediction error = 0.105) was slightly superior to Lasso (bootstrap .632+ prediction error = 0.124) and SICA (bootstrap .632+ prediction error = 0.132). In order to compare, the prediction error curve of the method utilized by Saintigny et al. [3] to select effective genes named component-wise likelihood-based boosting was also provided in Figure 2. As shown, based on prediction error curve, the performance of SCAD was better compared to component-wise likelihood-based boosting.

Finally, five most effective genes (8126931, 8120206, 7971191, 8161169, and 7897663) were selected using area under ROC curve by means of timeROC package. The median AUC of these genes was 0.825, 0.816, 0.787, 0.786, and 0.702, respectively. Genes 8126931, 8120206, and 7897663 were significant ($P < 0.001$, $P = 0.01$, and $P < 0.001$) by using additive model and the expression of all three genes decreased the survival time.

4. Discussion

Due to the low observations for microarray data, only a small number of influential genes can reliably be identified [20]. Therefore, fitting sparse models with only a few nonzero coefficients is interesting [20].

In this study, the performance of three sparse variable selection techniques was investigated based on an additive hazards model for survival data, in the presence of high-dimensional covariates. It was indicated that, based on AUC criterion, Lasso penalty resulted in higher capability of prediction than other methods. The vast majority of genes selected by Lasso approach had frequency greater than 90 percent indicating the stability of the selected gene in this approach. In a study conducted by Ma and Huang [13] the high-dimensional survival time modeled using the additive hazards model. Their results, based on two real data sets, showed appropriate performance of Lasso. Also, in a study conducted by Lin and Lv [4] performance of Lasso in an additive hazards model was reasonable.

Although Lasso performs shrinkage and variable selection simultaneously for better prediction and model interpretation, if there is a group of variables among which the pairwise correlations are very high, then the LASSO tends to arbitrarily select only one variable from the group [21]. This restriction might lead to some problems in the analysis of gene expression data where identification of an entire set of correlated genes may lead to an improved understanding of the biological pathway [5]. Superior performance of the elastic net penalty compared to Lasso was confirmed by Engler and Li based on a Cox proportional hazards model.

It is suggested to use the elastic net penalization approach to deal with high-dimensional low-sample size survival data in an additive manner.

Despite the lower number of selected genes, the AUC of the SCAD and SICA was also comparable to the Lasso. On the other hand, based on the prediction error criterion, the SCAD showed minimum value indicating better prediction. Lin and Lv [4] conducted a simulation study to investigate and to compare some sparse variable selection techniques. Their results showed that the performance of SCAD and SICA was good with slight superiority of SICA.

In addition, the selected genes exhibited some consistency across three used methods in terms of selecting common genes. However, variability was observed due to the low-sample size and ultrahigh-dimensionality as well as different penalty functions which were utilized by different techniques [4]. Reducing the dimensionality by some techniques including sure independence screening [22] is suggested before using the regularization methods to gain more stability. Fan and Lv [22] showed that the prediction performance of all methods tends to improve.

Based on the AUC, three genes were diagnosed as the most effective genes on OPL patients' survival. Accordingly, these genes can predict the survival time of the patients with OPL, if they are taken into account simultaneously in an additive hazards model. The expression of genes 8126931, 8120206, and 7897663 can decrease the survival time.

Saintigny et al. [3] conducted a study on the same dataset using a boosting approach which is also a sparse technique. They utilized Cox proportional hazards model and identified a small cluster of genes expression. However, due to the sparseness of utilized techniques, there is only a small overlap with the present study, with only one common microarray feature (8095441). Also the performance of boosting based on AUC (AUC = 0.95) was comparable with three methods used in the present study. However, based on prediction error curve the performance of SCAD was superior. It is suggested that simulation studies are conducted to further assessment of the performance of these techniques as well as different data sets.

One important restriction in the present study was that gene expression data were not identified for a large part of the dataset. Eliminating this part of dataset may result in selection bias in the present study's result. Besides, we could not assess the biological mechanisms and the effect of the selected genes on the pathogenesis or progress of OPL; therefore, it is suggested that the pathogenic effect of reported genes is evaluated in the future studies.

The present study introduced a new set of influential microarray features in predicting OPL patients' survival from a different perspective of the proportional hazards. According to the result of the present study, despite the small number of selected genes three methods of Lasso, SCAD, and SICA showed reasonable performance in additive manner and the selected genes improved prediction performance over a purely clinical model.

The additive risk models and used variable selection approaches provide a useful alternative to existing dimension reduction techniques based on Cox's model for survival data with high-dimensional covariates. The performance of the different models and dimension reduction techniques are data dependent with no method dominating the others [23]. Comprehensive simulation studies and data analysis will be required to draw more definitive conclusions.

5. Conclusion

The present study indicated that three feature selection approaches of the Lasso, SCAD, and SICA performed reasonably well in handling high-dimensional time-to-event data corresponding to OPL patients based on additive risk. However, the SICA selected smaller number of genes than the Lasso and SCAD. Furthermore, the selected genes by these methods improved the prediction of purely clinical model and can satisfactorily predict survival.

Conflict of Interests

The authors declare that there is no conflict of interests.

References

[1] Z. Magić, G. Supić, M. Branković-Magić, and N. Jovic, "DNA methylation in the pathogenesis of head and neck cancer," in *Methylation: From DNA, RNA and Histones to Diseases and Treatment*, InTech, 2013.

[2] L. G. T. Morris, A. G. Sikora, S. G. Patel, R. B. Hayes, and I. Ganly, "Second primary cancers after an index head and neck cancer: subsite-specific trends in the era of human papillomavirus—associated oropharyngeal cancer," *Journal of Clinical Oncology*, vol. 29, no. 6, pp. 739–746, 2011.

[3] P. Saintigny, L. Zhang, Y.-H. Fan et al., "Gene expression profiling predicts the development of oral cancer," *Cancer Prevention Research*, vol. 4, no. 2, pp. 218–229, 2011.

[4] W. Lin and J. Lv, "High-dimensional sparse additive hazards regression," *Journal of the American Statistical Association*, vol. 108, no. 501, pp. 247–264, 2013.

[5] D. Engler and Y. Li, "Survival analysis with high-dimensional covariates: an application in microarray studies," *Statistical Applications in Genetics and Molecular Biology*, vol. 8, no. 1, article 14, 2009.

[6] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society B: Statistical Methodology*, vol. 58, no. 1, pp. 267–288, 1996.

[7] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[8] A. Antoniadis, P. Fryzlewicz, and F. Letué, "The dantzig selector in Cox's proportional hazards model," *Scandinavian Journal of Statistics*, vol. 37, no. 4, pp. 531–552, 2010.

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

[10] J. Lv and Y. Fan, "A unified approach to model selection and sparse recovery using regularized least squares," *The Annals of Statistics*, vol. 37, no. 6 A, pp. 3498–3528, 2009.

[11] J. Lv and Y. Fan, "A unified approach to model selection and sparse recovery using regularized least squares," *The Annals of Statistics*, vol. 37, no. 6A, pp. 3498–3528, 2009.

[12] J. Fan and J. Lv, "Nonconcave penalized likelihood with NP-dimensionality," *IEEE Transactions on Information Theory*, vol. 57, no. 8, pp. 5467–5484, 2011.

[13] S. Ma and J. Huang, "Additive risk survival model with microarray data," *BMC Bioinformatics*, vol. 8, no. 1, article 192, 2007.

[14] J. J. Goeman, "L1 penalized estimation in the Cox proportional hazards model," *Biometrical Journal*, vol. 52, no. 1, pp. 70–84, 2010.

[15] X. Xie, H. D. Strickler, and X. Xue, "Additive hazard regression models: an application to the natural history of human papillomavirus," *Computational and Mathematical Methods in Medicine*, vol. 2013, Article ID 796270, 7 pages, 2013.

[16] D. Lin and Z. Ying, "Semiparametric analysis of the additive risk model," *Biometrika*, vol. 81, no. 1, pp. 61–71, 1994.

[17] H. Zou and R. Li, "One-step sparse estimates in nonconcave penalized likelihood models," *The Annals of Statistics*, vol. 36, no. 4, pp. 1509–1533, 2008.

[18] P. J. Heagerty, T. Lumley, and M. S. Pepe, "Time-dependent ROC curves for censored survival data and a diagnostic marker," *Biometrics*, vol. 56, no. 2, pp. 337–344, 2000.

[19] B. Efron and R. Tibshirani, "Improvements on cross-validation: the 632+ bootstrap method," *Journal of the American Statistical Association*, vol. 92, no. 438, pp. 548–560, 1997.

[20] H. Binder, A. Allignol, M. Schumacher, and J. Beyersmann, "Boosting for high-dimensional time-to-event data with competing risks," *Bioinformatics*, vol. 25, no. 7, pp. 890–896, 2009.

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

[22] J. Fan and J. Lv, "Sure independence screening for ultrahigh dimensional feature space," *Journal of the Royal Statistical Society B: Statistical Methodology*, vol. 70, no. 5, pp. 849–911, 2008.

[23] S. Ma, M. R. Kosorok, and J. P. Fine, "Additive risk models for survival data with high-dimensional covariates," *Biometrics*, vol. 62, no. 1, pp. 202–210, 2006.