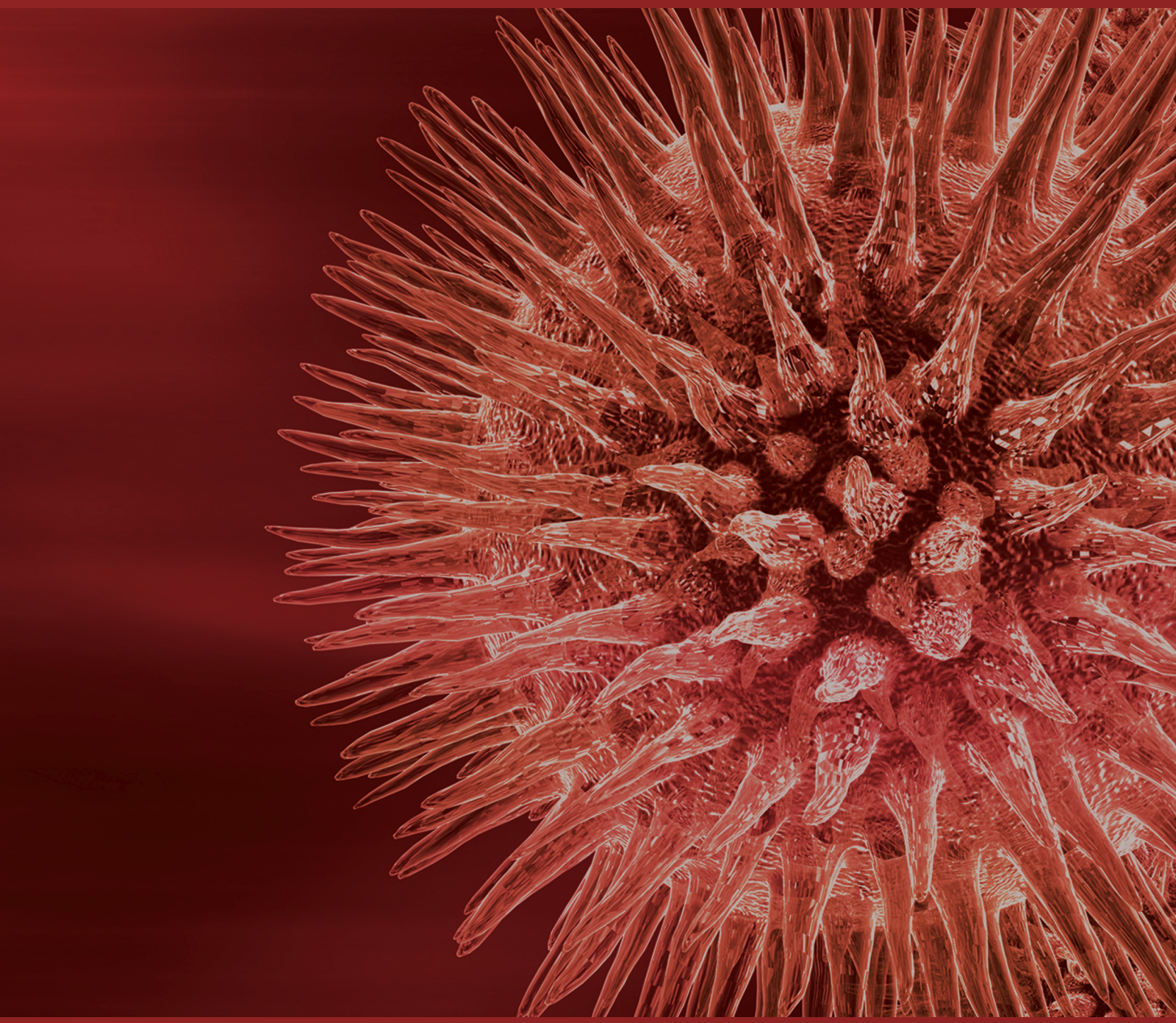
BioMed Research International

# Integrative Genomics and Computational Systems Medicine

Guest Editors: Zhongming Zhao, Bing Zhang, Yufei Huang, Hua Xu, and Jason E. Mcdermott



# Integrative Genomics and Computational Systems Medicine

BioMed Research International

# Integrative Genomics and Computational Systems Medicine

Guest Editors: Zhongming Zhao, Bing Zhang, Yufei Huang, Hua Xu, and Jason E. Mcdermott

# Contents

**Integrative Genomics and Computational Systems Medicine**, Jason E. McDermott,Yufei Huang, Bing Zhang, Hua Xu, Zhongming Zhao
Volume 2014, Article ID 945253, 3 pages

**Development of Dual Inhibitors against Alzheimer's Disease Using Fragment-Based QSAR and Molecular Docking**, Manisha Goyal, Jaspreet Kaur Dhanjal, Sukriti Goyal, Chetna Tyagi, Rabia Hamid, and Abhinav Grover
Volume 2014, Article ID 979606, 12 pages

**MultiRankSeq: Multiperspective Approach for RNAseq Differential Expression Analysis and Quality Control**, Yan Guo, Shilin Zhao, Fei Ye, Quanhu Sheng, and Yu Shyr
Volume 2014, Article ID 248090, 8 pages

**A Diverse Stochastic Search Algorithm for Combination Therapeutics**, Mehmet Umut Caglar and Ranadip Pal
Volume 2014, Article ID 873436, 9 pages

**Evaluating Word Representation Features in Biomedical Named Entity Recognition Tasks**, Buzhou Tang, Hongxin Cao, Xiaolong Wang, Qingcai Chen, and Hua Xu
Volume 2014, Article ID 240403, 6 pages

**Integrative Analysis of miRNA-mRNA and miRNA-miRNA Interactions**, Li Guo, Yang Zhao, Sheng Yang, Hui Zhang, and Feng Chen
Volume 2014, Article ID 907420, 8 pages

**Network-Assisted Prediction of Potential Drugs for Addiction**, Jingchun Sun, Liang-Chin Huang, Hua Xu, and Zhongming Zhao
Volume 2014, Article ID 258784, 9 pages

**Computational Analysis of Transcriptional Circuitries in Human Embryonic Stem Cells Reveals Multiple and Independent Networks**, Xiaosheng Wang and Chittibabu Guda
Volume 2014, Article ID 725780, 10 pages

**HGF Accelerates Wound Healing by Promoting the Dedifferentiation of Epidermal Cells through $\beta_1$-Integrin/ILK Pathway**, Jin-Feng Li, Hai-Feng Duan, Chu-Tse Wu, Da-Jin Zhang, Youping Deng, Hong-Lei Yin, Bing Han, Hui-Cui Gong, Hong-Wei Wang, and Yun-Liang Wang
Volume 2013, Article ID 470418, 9 pages

**Multiple Biomarker Panels for Early Detection of Breast Cancer in Peripheral Blood**, Fan Zhang, Youping Deng, and Renee Drabier
Volume 2013, Article ID 781618, 7 pages

**Expression Sensitivity Analysis of Human Disease Related Genes**, Liang-Xiao Ma, Ya-Jun Wang, Jing-Fang Wang, Xuan Li, and Pei Hao
Volume 2013, Article ID 637424, 8 pages

**DeGNServer: Deciphering Genome-Scale Gene Networks through High Performance Reverse Engineering Analysis**, Jun Li, Hairong Wei, and Patrick Xuechun Zhao
Volume 2013, Article ID 856325, 10 pages

**Novel Natural Structure Corrector of ApoE4 for Checking Alzheimer's Disease: Benefits from High Throughput Screening and Molecular Dynamics Simulations**, Manisha Goyal, Sonam Grover, Jaspreet Kaur Dhanjal, Sukriti Goyal, Chetna Tyagi, Sajeev Chacko, and Abhinav Grover
Volume 2013, Article ID 620793, 8 pages

**QPLOT: A Quality Assessment Tool for Next Generation Sequencing Data**, Bingshan Li, Xiaowei Zhan, Mary-Kate Wing, Paul Anderson, Hyun Min Kang, and Goncalo R. Abecasis
Volume 2013, Article ID 865181, 4 pages

**Comparative Study of Exome Copy Number Variation Estimation Tools Using Array Comparative Genomic Hybridization as Control**, Yan Guo, Quanghu Sheng, David C. Samuels, Brian Lehmann, Joshua A. Bauer, Jennifer Pietenpol, and Yu Shyr
Volume 2013, Article ID 915636, 7 pages

**New aQTL SNPs for the CYP2D6 Identified by a Novel Mediation Analysis of Genome-Wide SNP Arrays, Gene Expression Arrays, and CYP2D6 Activity**, Guanglong Jiang, Arindom Chakraborty, Zhiping Wang, Malaz Boustani, Yunlong Liu, Todd Skaar, and Lang Li
Volume 2013, Article ID 493019, 7 pages

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 945253, 3 pages
http://dx.doi.org/10.1155/2014/945253

# *Editorial*
# Integrative Genomics and Computational Systems Medicine

**Jason E. McDermott,[1] Yufei Huang,[2] Bing Zhang,[3,4,5] Hua Xu,[6] and Zhongming Zhao[3,4,5]**

[1] *Computational Biology and Bioinformatics Group, Pacific Northwest National Laboratory, Richland, WA 99352, USA*
[2] *Department of Electrical and Computer Engineering, The University of Texas at San Antonio, San Antonio, TX 78249, USA*
[3] *Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, TN 37203, USA*
[4] *Center for Quantitative Sciences, Vanderbilt University Medical Center, Nashville, TN 37232, USA*
[5] *Department of Cancer Biology, Vanderbilt University School of Medicine, Nashville, TN 37232, USA*
[6] *School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA*

Correspondence should be addressed to Zhongming Zhao; zhongming.zhao@vanderbilt.edu

Received 14 May 2014; Accepted 14 May 2014; Published 15 June 2014



The exponential growth in generation of large amounts of genomic data from biological samples has driven the emerging field of systems medicine. This field is promising because it improves our understanding of disease processes at the systems level. However, the field is still in its young stage. There exists a great need for novel computational methods and approaches to effectively utilize and integrate various -omics data.

Systems medicine has been growing rapidly in part due to the emerging technologies to gather high-volume measurements from biological samples. One of the first such technologies, the mRNA microarray, is being replaced by next generation sequencing (NGS), which provides a much higher resolution (digital measurement) of genetic information (e.g., at the mRNA transcript level). Array and NGS-based methods to characterize genetic variation (single nucleotide polymorphisms, short insertions and deletions, copy number variation, and structural variants), DNA methylation changes, microRNAs (miRNAs) differential expression, and other types of biological information have dramatically expanded the generation of biological data. Other sources of data from mass spectrometry-based proteomics and metabolomics to high-throughput determination of protein-protein interactions and regulatory relationships provide further information for a systems-level understanding of disease. Finally, collection of clinical data and electronic medical records (EMRs) has made modern biomedical research possible on the full scale of data integration, that is, an integration scenario of using genomic, transcriptomic, proteomic, metabolomic, and phenotypic data.

The rationale of discovery-based high-throughput investigation of disease is that there are molecular signatures (composed of genes, transcripts, proteins, and small molecules) that can be identified for better diagnosis, prognosis, and/or treatment of disease. However, challenges arise in the analysis of high-throughput data because of the large number of possible variables raising the very real potential for false-positive predictions and overfitting of data, as well as many other potential problems (e.g., data quality, missingness, lack of power, etc.). To ameliorate these problems, computational approaches have been developed that utilize existing knowledge, such as overlaying high-throughput observations on regulatory or protein-protein interaction networks or canonical biological pathways.

For this special issue we solicited manuscripts in several different subject areas including data integration from multiple high-throughput sources, NGS data analysis and applications, personalized medicine and translational bioinformatics, modeling of pathways and networks, and data mining and pattern recognition in biomedical applications. We briefly describe the accepted papers in this special issue in the remainder of this editorial.

Two papers, "*MultiRankSeq: multi-perspective approach for RNAseq differential expression analysis and quality control*"

by Y. Guo et al. and "*QPLOT: a quality assessment tool for next generation sequencing data*" by B. Li et al., describe algorithms for analysis of NGS data. Y. Guo et al. introduce a novel tool, namely, MultiRankSeq, which combines the output of three independent programs to determine differential expression from RNAseq data to provide a single improved output. QPLOT is a tool for assessing the quality of NGS runs by providing both summary quality metrics and graphical representations of these metrics. In another paper, entitled "*Comparative study of exome copy number variation estimation tools using array comparative genomic hybridization as control*," Y. Guo et al. systematically compare four different tools for detecting copy number variations (CNVs) from whole exome sequencing (WES) against a standard array-based method for CNV evaluation.

In "*Computational analysis of transcriptional circuitries in human embryonic stem cells reveals multiple and independent networks*," X. Wang and C. Guda assess the role of core transcription factors in the pluripotency of embryonic stem (ES) cells. Their computational analyses identified several additional transcriptional regulatory networks that might be involved in this complex regulatory process, providing interesting hypotheses about mechanisms of fate determination in ES cells. The paper "*Network-assisted prediction of potential drugs for addiction*" by J. Sun et al. describes computational analyses of drug-target networks for addictive and nonaddictive drugs. The authors analyzed the topology of these networks and found that drugs with similar effects could cluster together and identified a set of nonaddictive drugs that might have therapeutic benefits for treatment of addiction. This paper was called out in the recent "Translational Bioinformatics Year-in-Review" in 2014 Joint Summits on Translational Science (http://www.amia.org/jointsummits2014). In "*DeGNServer: deciphering genome-scale gene networks through high performance reverse engineering analysis*," J. Li et al. describe their webserver to infer transcriptional regulatory networks from large-scale datasets. The server makes use of a computer cluster to run a number of network inference algorithms and return the results to the user very quickly, thus facilitating genome-scale network reconstruction.

Two papers from M. Goyal et al., "*Development of dual inhibitors against Alzheimer's disease using fragment-based QSAR and molecular docking*" and "*Novel natural structure corrector of ApoE4 for checking Alzheimer's disease: benefits from high throughput screening and molecular dynamics simulations*," deal with molecular docking simulations to determine small-molecule inhibitors targeting Alzheimer's disease. In the first paper, M. Goyal et al. used a fragment-based quantitative structure activity relationship (QSAR) analysis to identify lead compounds that might inhibit interaction of proteins that drive Alzheimer's disease pathogenesis. In the second paper, the authors describe large-scale docking simulations to screen for inhibitors of the conformational change of apolipoprotein E4 (ApoE4) that is thought to drive Alzheimer's pathogenesis. They further show the value of molecular dynamics simulations to screen candidates to eliminate molecules that do not have stable binding properties with targets. In "*HGF accelerates wound healing by promoting the dedifferentiation of epidermal cells through $\beta_1$-integrin/ILK pathway*," J.-F. Li et al. experimentally investigate the contribution of hepatocyte growth factor (HGF) to wound healing. They showed that treatment of diabetic mice promoted proliferation and migration of epithelial cells and that this effect could be blocked by silencing the $\beta_1$-integrin signaling pathway.

In the paper "*Integrative analysis of miRNA-mRNA and miRNA-miRNA interactions*," the authors first generated RNAseq data for normal and tumor cell lines and then identified aberrantly expressed mRNAs and miRNAs. Groups of similarly expressed miRNAs and mRNAs were analyzed to highlight examples of flexible and selective regulatory networks underlying these interactions. In "*A diverse stochastic search algorithm for combination therapeutics*," M. U. Caglar and R. Pal show how the use of a stochastic search algorithm can be useful in identification of optimal combinations of drugs for therapy, the so-called drug cocktails. Their novel method greatly reduces the number of experimental steps needed to assess the optimal combination of drugs for a particular therapy. In "*Evaluating word representation features in biomedical named entity recognition tasks*" by B. Tang et al., the authors present a comparative analysis of three different methods for word representation in recognition of named entities from biomedical literature. Their findings indicate that a combination of the complementary approaches can improve results on benchmark recognition tasks.

In the paper by F. Zhang et al., "*Multiple biomarker panels for early detection of breast cancer in peripheral blood*," the authors describe the use of machine-learning approaches to identify a five-gene panel that can identify breast cancer from peripheral blood samples. In the paper by Jiang et al., "*New aQTL SNPs for the CYP2D6 identified by a novel mediation analysis of genome-wide SNP arrays, gene expression arrays, and CYP2D6 activity*," the authors develop a novel approach for the detection of transexpression quantitative trait loci (eQTLs) from genome-wide association studies by considering indirect effects introduced by a mediator gene. They apply their method to analyze indirect regulatory effects on the important liver enzyme, CYP2D6. Finally, in "*Expression sensitivity analysis of human disease related genes*," L.-X. Ma et al. examine the expression of genes implicated in a range of diseases. They report that genes that are robustly expressed under different perturbations are more likely to be associated with lethal diseases, whereas less robustly expressed genes are associated with nonlethal diseases.

## Acknowledgments

We would like to acknowledge the anonymous reviewers for their critical comments that helped to improve the quality of the papers in this special issue. We would like to acknowledge the organizers and committee members of the International Conference on Intelligent Biology and Medicine (ICIBM 2013, held on August 11–13, 2013) for their efforts to provide a forum to discuss integrative genomics and computational systems medicine, through which this special issue was made possible. We thank the National Science Foundation (NSF Grant IIS-1329380) and Vanderbilt Center for Quantitative

Sciences for financial support of ICIBM 2013. JEM was supported by the Clinical Proteomic Tumor Analysis Consortium [NIH/NCI CA160019] and the Signature Discovery Initiative, a component of the Laboratory Directed Research and Development Program at Pacific Northwest National Laboratory, a multiprogram national laboratory operated by Battelle for the U.S. Department of Energy under Contract DE-AC05-76RL01830.

*Jason E. McDermott*
*Yufei Huang*
*Bing Zhang*
*Hua Xu*
*Zhongming Zhao*

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 979606, 12 pages
http://dx.doi.org/10.1155/2014/979606

*Research Article*

# Development of Dual Inhibitors against Alzheimer's Disease Using Fragment-Based QSAR and Molecular Docking

**Manisha Goyal,[1] Jaspreet Kaur Dhanjal,[2] Sukriti Goyal,[1] Chetna Tyagi,[2] Rabia Hamid,[3] and Abhinav Grover[2]**

[1] *Apaji Institute of Mathematics & Applied Computer Technology, Banasthali University, Tonk, Rajasthan 304022, India*
[2] *School of Biotechnology, Jawaharlal Nehru University, New Delhi 110067, India*
[3] *Department of Biochemistry, University of Kashmir, Srinagar 190006, India*

Correspondence should be addressed to Abhinav Grover; abhinavgr@gmail.com

Received 17 December 2013; Revised 27 March 2014; Accepted 27 March 2014; Published 12 June 2014

Academic Editor: Jason E. Mcdermott



Alzheimer's (AD) is the leading cause of dementia among elderly people. Considering the complex heterogeneous etiology of AD, there is an urgent need to develop multitargeted drugs for its suppression. $\beta$-amyloid cleavage enzyme (BACE-1) and acetylcholinesterase (AChE), being important for AD progression, have been considered as promising drug targets. In this study, a robust and highly predictive group-based QSAR (GQSAR) model has been developed based on the descriptors calculated for the fragments of 20 1,4-dihydropyridine (DHP) derivatives. A large combinatorial library of DHP analogues was created, the activity of each compound was predicted, and the top compounds were analyzed using refined molecular docking. A detailed interaction analysis was carried out for the top two compounds (EDC and FDC) which showed significant binding affinity for BACE-1 and AChE. This study paves way for consideration of these lead molecules as prospective drugs for the effective dual inhibition of BACE-1 and AChE. The GQSAR model provides site-specific clues about the molecules where certain modifications can result in increased biological activity. This information could be of high value for design and development of multifunctional drugs for combating AD.

## 1. Introduction

Alzheimer's disease (AD) is an irreversible chronic brain disorder among elderly people [1–3]. AD is characterized by steady cognitive impairment, memory loss, and decline in language. It is one of the leading causes of death in the world. For instance, it was estimated that 5.2 million Americans of all ages were suffering from AD in 2013 making it the sixth leading cause of death in the United States (Alzheimer's association; http://www.alz.org/). The devastating pathological hallmarks of AD are extracellular accumulation of neurotoxic amyloid $\beta$ (A$\beta$) peptides [4], loss of the presynaptic markers of the cholinergic system in the brain, mitochondrial dysfunction, and formation of dense neurofibrillary tangles of hyperphosphorylated tau protein in the central nervous system [5–7].

Most of the U.S. Food and Drug Administration approved drugs are available for the symptomatic treatment of AD. Among these drugs, donepezil, tacrine, rivastigmine, and galantamine are based on cholinergic hypothesis [8–11]. Furthermore, memantine is an antagonist drug of N-methyl-D-aspartate receptor [12–14]. However, the observable toxic issues such as hepatotoxicity, vomiting, diarrhea, and nausea forced these drugs to move out from the pharmaceutical market [15]. Moreover, medicational pharmacokinetic effects of these drugs are just for marginally alleviating the symptoms and not to have interruption in neurodegenerative cascade which is the root pathophysiology of AD [16–18]. Considering the complex heterogeneous etiology of AD, modulation of one enzyme might not be sufficient enough for the effective treatment of AD. Therefore, the present day research in AD drug development is shifting towards identification and design of multitargeted novel molecules instead of single targeted molecules for the long term suppression of AD. For instance, Piazzi et al. report AChE inhibitor purposely

designed to bind at both the catalytic and the peripheral sites of the human enzyme [19].

Most of the experimental evidences suggest that deposition of amyloid plaques in the brain of Alzheimer's patients is the key factor of pathogenic cascade of the disease [16, 20]. A$\beta$, which is the core component of the amyloid plaques [15], is produced by subsequent cleavage of a large transmembrane protein—amyloid precursor protein (APP)—by two different proteolytic enzymes $\beta$- and $\gamma$-secretase [21]. The complete biochemical mechanism of proteolytic cleavage depends on the protein-protein interactions between APP and $\beta$-amyloid cleavage enzymes (BACE-1) [22]. Blocking the interface between these protein interactions has huge therapeutic potential for slowing down the long term progression of AD. It has been reported that acetylcholine esterase (AChE) also plays an important role in accumulation of A$\beta$ and acts as a promoter of A$\beta$ fibril production [23]. This activity of AChE is associated with its peripheral anionic site (PAS). Since BACE-1 plays a major role in the initiation of neuropathological cascade of plaque formation and AChE accelerates A$\beta$ deposition in brain, both of these enzymes hold considerable promise as therapeutic targets of AD. Thus, dual target directed strategy is more likely to show comprehensive obliteration of AD in synergistic manner. Multitarget drugs are more efficient as they prevent unwanted compensatory mechanisms, which might result in cellular redundancy, from developing [24].

Discovery of small molecules for targeting protein-protein interfaces beholds enormous challenges and is accounted by various factors, namely, shape of typical protein-protein interface and flexibility of proteins among others. To speed up the drug discovery process, various fast and accurate computational methods have been illustrated which assist the development of novel therapeutic drugs to interrupt the interaction between proteins [25, 26]. Usage of quantitative structure activity relationship- (QSAR) based approaches is worthwhile when knowledge of ligand molecules for a particular target is available. Group-based QSAR (GQSAR) is one of the most recent and effective ligand-based drug designing approaches which uses descriptors evaluated specifically for the substituent groups or fragments of the ligands. This approach identifies the specific sites where the groups need to be modified for designing optimized molecules with enhanced biological activity [27]. GQSAR model can be developed by applying statistical methods like partial least square (PLS), principle component regression, multiple regression, continuum regression, and k-Nearest Neighbour on a series of congeneric compounds in order to gain insights into the effects of descriptors on their biological activity [27, 28].

Herein, our attempts are focused on the discovery of novel small molecules that can compete to bind with one of the interacting proteins with higher binding affinity in order to disrupt the interactions between APP and BACE-1 and simultaneously are able to bind to the PAS site of AChE. Present study describes a detailed GQSAR analysis on 1,4-dihydropyridine (DHP) derivatives, reported as potential inhibitors of BACE-1 [4], in order to elucidate the structural features of the molecular fragments of these molecules that lay significant contribution towards their biological activity. GQSAR model was further used to develop a combinatorial library of novel molecules followed by their activity prediction. Mechanistic analysis of binding modes of these identified leads within the active site of both targets was performed using docking studies. Thus, our study delineates identification of novel leads having dual inhibiting effects due to binding to both, BACE-1 and the PAS of AChE.

Table 1: Unicolumn statistical parameters for the selected biological dataset.

| | Average | Max. | Min. | Std. dev. | Sum |
|---|---|---|---|---|---|
| Training set | 4.74 | 5.10 | 4.50 | 0.20 | 71.13 |
| Test set | 4.65 | 4.83 | 4.41 | 0.17 | 23.27 |

## 2. Materials and Methods

*2.1. Biological Dataset.* A biological data set of 20 compounds of DHP derivatives was chosen in the present study to carry out the GQSAR analysis. DHP were found to have strong inhibitory capability against BACE-1 [4]. The experimentally reported inhibitory activity [$IC_{50}$ ($\mu$M)] of all the 20 compounds was converted into $pIC_{50}$ [$-\log_{10} IC_{50}$], which was then subsequently used as response or dependent variable for GQSAR model building. The 2D structures of compounds were drawn using Marvin Sketch (v 5.12.1, ChemAxon) [21]. 2D chemical structures of DHP analogues along with their biological activities are presented in Table 1. Molecules were converted into 3D format and then energetically optimized using Vlife Engine module of Vlife Molecular Design Suite (Vlife MDS) [29]. The optimized molecules were generated using Merck Molecular force field, distance dependent function, and energy gradient of 0.01 kcal/mol.

*2.2. Fragmentation and Descriptor Calculation.* All molecules considered here had a common DHP scaffold and 4 substitution sites where different R-groups were attached. On the basis of different R-groups, each molecule was divided into 4 fragments or groups in order to perform GQSAR analysis. Optimized dataset of all molecules was considered for GQSAR analysis on the basis of common DHP template. A total of 705 physicochemical descriptors were calculated for various groups present at each substitution site using Vlife MDS. These included 2D descriptors such as element count, extended topological indices, Merck molecular force field atom type count, and electrotopological and alignment independent descriptors among others [30]. Independent variable calculation was further followed by removal of invariable columns containing constant values for more than 90% molecules, which finally resulted in 311 independent variables from the large pool of descriptors.

*2.3. Selection of Test Set and Training Set.* With an aim to develop a GQSAR model, the dataset was split into two optimal training and test sets using random selection method. The robustness of these sets was evaluated by generating unicolumn statistical parameters such as mean,

standard deviation, maximum, and minimum for both test and training sets. The dataset division satisfied the criteria of an appropriate model; namely, the maximum of the test set was less than the maximum of the training set and the minimum of the training set was greater than the minimum of the test set. This analysis validated the selected training and test sets.

*2.4. GQSAR Model Generation.* To select the optimal subset of variables (descriptors) that can significantly correlate with biological activity of molecules from the pool of descriptors, various variable selection methods such as step-wise search algorithm, genetic algorithm, and simulated annealing among others can be used. A number of statistical methods such as partial least square (PLS), multiple regression, and principle component regression can be used for model building. Herein, simulated annealing combined with PLS regression was used to generate the GQSAR model. Simulation of a physical process is known as simulated annealing, which involves heating the system to a high temperature and then gradually cooling it down to room temperature [31]. All the values of statistical parameters for simulated annealing were kept as default. The number of terms (number of descriptors) to be included in the final GQSAR model was kept as 3.

*2.5. Model Evaluation and Validation.* The developed GQSAR model was evaluated using two types of validation—internal and external validations. Internal (cross) validation was carried out using leave-one-out method [32]. Cross-validation coefficient $q^2$ was calculated as

$$q^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - y_{\text{mean}})^2}, \quad (1)$$

where $y_i$ and $\hat{y}_i$ are the actual and the predicted activity of the $i$th molecule in the training set, respectively, and $y_{\text{mean}}$ is the average activity of all molecules in the training set.

For external validation of the model, the $\text{pIC}_{50}$ values of the test set molecules were predicted and the pred_$r^2$ value that provides the statistical correlation between predicted and actual activities of the test set compounds was calculated as follows:

$$\text{pred\_}r^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - y_{\text{mean}})^2}, \quad (2)$$

where $y_i$ and $\hat{y}_i$ are the actual and the predicted activity of the $i$th molecule in the test set, respectively, and $y_{\text{mean}}$ is the average activity of all molecules in the training set.

All these statistical parameters were used to evaluate the quality of the model. Correlation coefficient ($r^2$) described the fitness of training set data whereas predictive correlation coefficient (pred-$r^2$) was used to evaluate the fitness of test set. Cross-validation coefficient ($q^2$) and $F$-test (Fischer's value) showed the statistical significance of the regression model and the standard errors (pred_$r^2$_se, $q^2$_se, and $r^2$_se) gave an idea of the quality and fitness of the model. Low standard error values indicated that the model is absolute and robust. The model is said to be robust and predictive if these statistical parameters satisfy the following conditions: $r^2 > 0.6$, pred_$r^2 > 0.5$, and $q^2 > 0.6$ [33, 34].

*2.6. Combinatorial Library Generation and Activity Prediction.* A combinatorial library was generated using Leadgrow module of Vlife MDS. For library generation a number of substitutions were made using various atoms and groups like alkyl, alkene, acids, aromatic rings, rings, carbonyl, cyanate, –O–$CH_3$, –O–$C_2H_5$, amide, benz, and hydrzo at all substitution sites (R1, R2, R3, and R4) of DHP template. The final GQSAR model generated was used for biological activity prediction of the compounds of the combinatorial library.

*2.7. Docking Studies.* The 3D structure of human BACE-1 (resolution: 1.70 Å) was obtained from PDB (PDB ID: 2B8L) [35]. The water molecules and all other heteroatoms were removed from the protein crystal structure. The protein was further prepared using Schrodinger's protein preparation wizard [36]. Conversion of all combinatorial structures to 3D form and further optimization were carried out using LigPrep module of the Schrodinger suite. All possible conformers for each molecule were generated using LigPrep. Docking studies were performed using Glide module of Schrodinger suite by creating a cubic grid (10 × 10 × 10 Å) around the active site residues of BACE-1 that are involved in cleavage of APP. The molecules of combinatorial library with high predicted activity were subjected to high throughput virtual screening (HTVS) protocol followed by Glide's extra precision (XP) docking protocol for futher docking refinement.

*2.8. Dual Inhibition Effect Studies.* Keeping in mind our aim to discover potent novel dual inhibitors of AChE and BACE-1, the above screened molecules were again subjected to docking at PAS site of AChE. This PAS site is involved in accumulation of A$\beta$ in the human brain. Crystal structure of human AChE (resolution: 2.0 Å) was obtained from PDB (PDB ID: 4M0E) [35]. Protein preparation and optimization was done using Schrodinger suite. Selected molecules having high XP scores were then checked for their drug-like properties using Lipinski filters. The two top scoring compounds showing dual inhibitory property were analyzed to observe the molecular mode of interaction between the target proteins and the ligands using ligplot program [37].

## 3. Results and Discussion

Here we have attempted to identify a novel GQSAR model depicting robust statistical correlation between structure and activity of DHP analogues which have been reported as potent suppressors of BACE-1. The adopted strategy initially identified a pool of 311 molecular descriptors to be used as independent variables. The $\text{pIC}_{50}$ value was used as the dependent variable. The dataset of 20 compounds was divided into two groups: test set including 5 molecules and training set including the rest of the molecules. The training set was

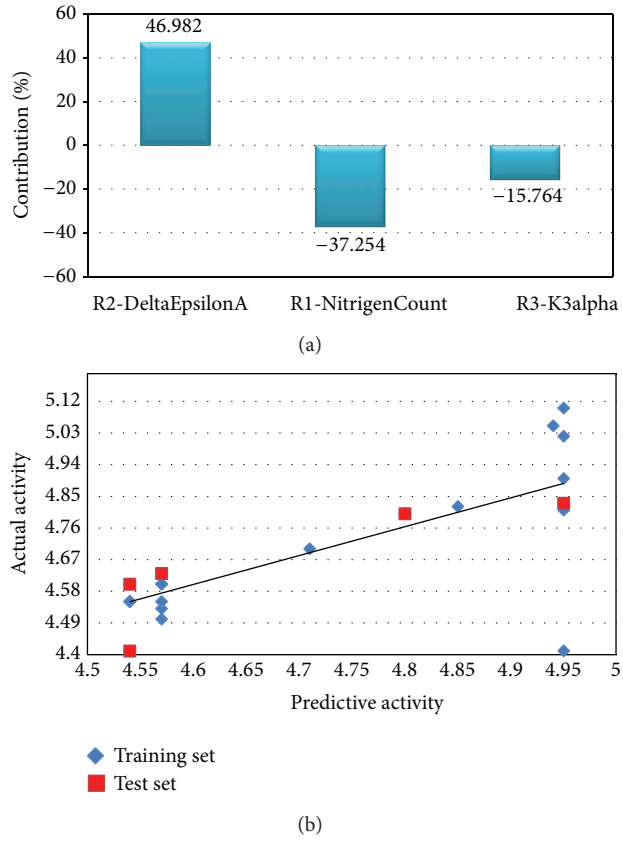



Figure 1: (a) The contribution of descriptors to the enhancement of biological activity of molecules. (b) Linear scatter plot depicting the distance of training and test data points from the regression line.

used for model building (Supplementary Table 1 available online at http://dx.doi.org/10.1155/2014/979606).

*3.1. Dataset Evaluation.* Before proceeding towards the next step, evaluation of the chosen test set is always a beneficial option to obtain a good predictive model. This was done by interpreting the unicolumn statistics mentioned in Table 1. Unicolumn statistics are stated in terms of min., max., average, std. dev. (standard deviation), and sum. The min. of test set should be equal or higher than the min. of training set and the max. of test set should be equal or lower than the max. of training set. Here, the dataset was found satisfying the required conditions, thus suggesting that the test set was interpolative. Along with these parameters, average and std. dev. determines the density distribution of both the test and the training sets. Interestingly, in this dataset, higher values of mean and std. dev. for training set indicated the presence of comparably high number of active molecules rather than the inactive ones and the presence of highly distributed activity of the molecules in the training set.

*3.2. Generated GQSAR Model.* The GQSAR model was generated using simulated annealing variable selection method in combination with PLS regression model building method. The statistical measurements of generated PLS regression model of GQSAR are summarized in Table 2. PLSR method predicts the correlation between the molecular fields and the inhibitory activity of the compounds [38]. It specifies the linear relationship between dependent variables ($pIC_{50}$) and the predictor variables (descriptors). Predicted activity of the dataset and the values of calculated descriptors for each molecule are mentioned in Supplementary Table 2. The reported GQSAR model can be stated in the form of a polynomial equation as follows:

$$\begin{aligned} pIC_{50} = {} & 3.48219\,(\text{R2-DeltaEpsilonA}) \\ & - 0.409885\,(\text{R1-NitrogensCount}) \\ & - 0.279723\,(\text{R3-k3alpha}) + 4.56912, \end{aligned} \quad (3)$$

where R1, R2, and R3 are the 2D descriptors along with their respective coefficient and the last numerical term in this equation is the regression constant. This equation explains that the descriptor DeltaEpsilonA shows positive contribution at substitution site R2 of DHP common moiety. However, the other two descriptors, NitrogensCount and K3alpha at R1 and R3 substitution sites, respectively, contribute negatively towards the biological activity of molecules. The contribution of these descriptors is illustrated in Figure 1. Below is the brief description of these molecular descriptors.

*R2-DeltaEpsilonA.* DeltaEpsilonA falls into the category of extended topochemical atom (ETA) indices which is an

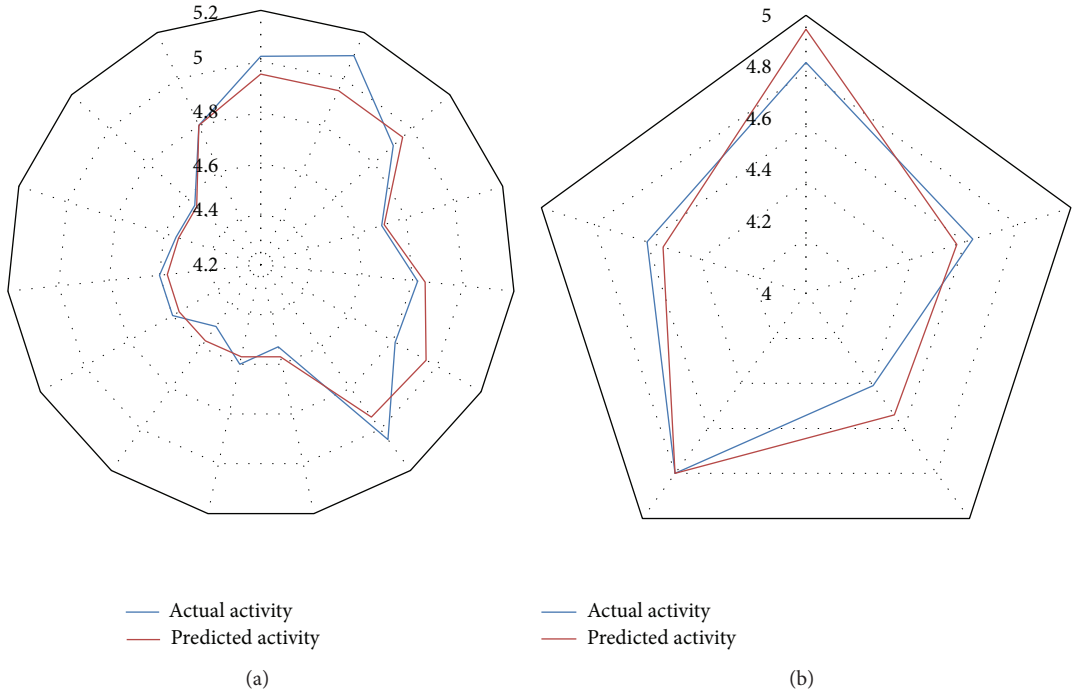



Figure 2: (a) Radar plot showing fitness of predicted and actual activity values of training set. (b) Radar plot exploring fitness of predicted and actual activity values of test set.

Table 2: Statistical parameters of generated GQSAR model.

| Statistical parameter | Value |
|---|---|
| $r^2$ | 0.85 |
| $q^2$ | 0.68 |
| $F$-test | 34.39 |
| $r^2$ se | 0.08 |
| $q^2$ se | 0.12 |
| pred_$r^2$ | 0.75 |
| pred_$r^2$se | 0.1 |
| $Z$score$R^2$ | 5.29 |
| Best rand $R^2$ | 0.52 |

extension of topochemically arrived unique parameters [39, 40]. Among the various basic parameters of ETA, DeltaEpsilonA is a measure of contribution of unsaturation and electronegative atom count [41] which is extensively applied for modelling various toxicity end-points in the quantitative domain of structure-activity relationships [42]. Here, it was observed that DeltaEpsilonA showed 46.98% contribution in activity enhancement of molecule when present at R2 site. Originally, R2 site was occupied by three different groups, namely, methylbenzylamine [NH-($\alpha$) methylBn], benzyl ester (OBn), and acetyl group.

*R1-NitrogensCount*. This physicochemical descriptor lies in the section of element count descriptors. As the name suggests, it indicates the number of nitrogen atoms present in a compound. This descriptor was observed to provide a 37.25% negative contribution at R1 substitution site which was originally engaged with different alkyl groups.

*R3-k3alpha*. The Kier and Hall Kappa molecular shape indices are intended to capture the overall aspects of molecular shape [43]. Third order Kappa Alpha (K3alpha) shape index is a subset of Kappa indices and the information encoded in it specifically refers to attributes of the shape of molecule. In present GQSAR model, K3alpha was found to have 15.74% negative participation at R3 substitution site for the enhancement of biological activity of molecules. This site was originally occupied by sulphonamide group, amide group, and ester group.

*3.3. GQSAR Model Validation.* The quality of the GQSAR model was judged on the basis of standard values of statistical parameters calculated during model generation. In this study, the convincing parametric values for GQSAR model were observed in terms of correlation coefficient $r^2$ (0.8514), predicted correlation coefficient pred_$r^2$ (0.7525), cross-correlation coefficient $q^2$ (0.6817), low standard error $r^2$_se (0.0847), $q^2$_se (0.1239), and pred_$r^2$se (0.0976) which implied that the model can be considered stable and accurate. Moreover, high values of other statistical parameters like $F$-test (34.3899) provided additional support that the model was significant and robust with minimum chance of failure. For better understanding of the relationship between the structural features of DHP derived molecules and their biological activity, two different graphical representations of predicted and actual activity values are shown in Figures 1(b) and 2. Two separate radar plots describe the fitness

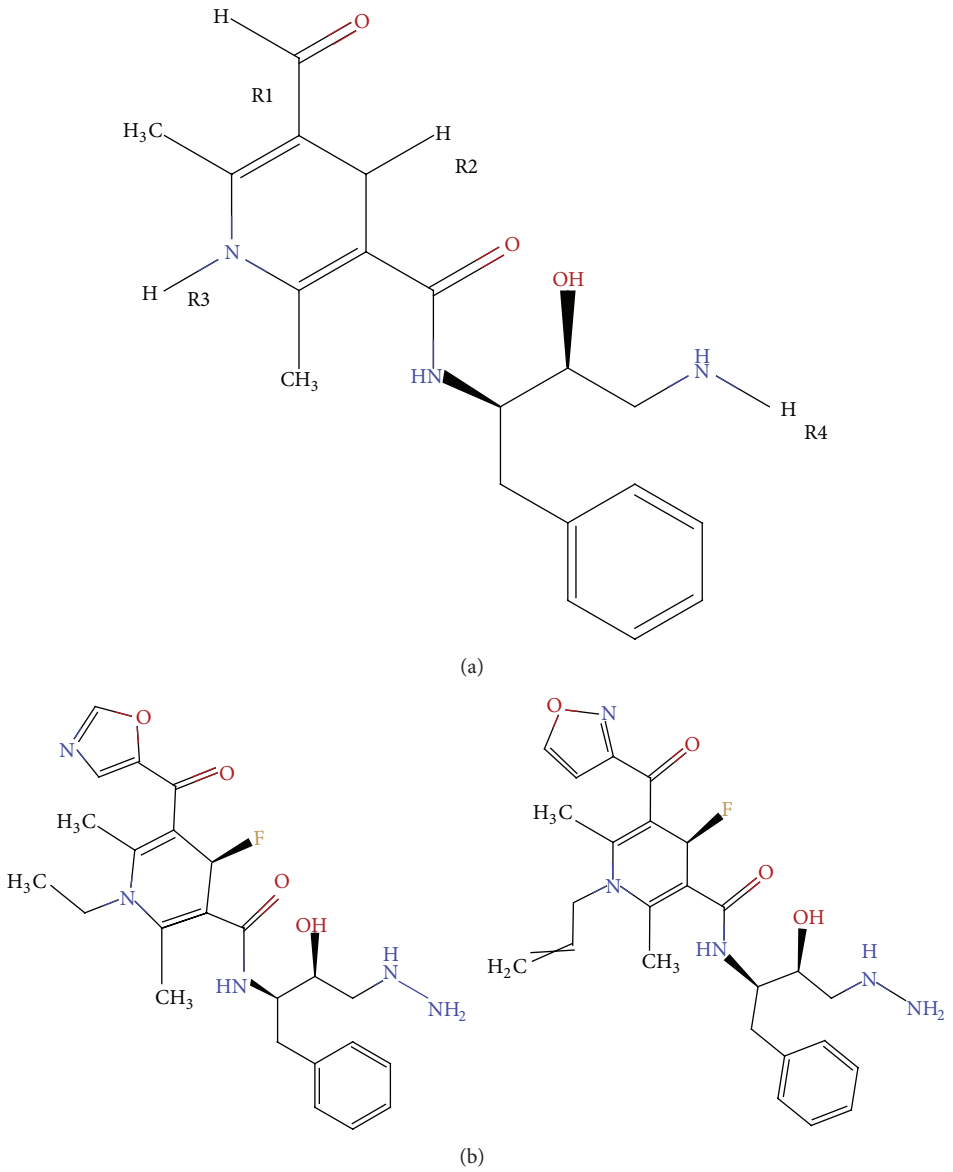



FIGURE 3: (a) 2D structure of common moiety of DHP derivatives. (b) 2D structures of selected molecules (EDC and FDC) possessing dual inhibitory property.

of predicted over actual values for training and test sets, respectively, and the linear scatter plot depicts the distance of training and test data points from the regression line which relatively gives an idea about the difference between actual and predicted activity values of both sets.

*3.4. Combinatorial Library Preparation and Activity Prediction.* The common moiety (Figure 3(a)) of DHP derivatives was taken into account for generation of the combinatorial library of novel compounds. This works by putting different chemical groups or atoms at four different substitution sites, namely, R1, R2, R3, and R4 of common template. At R1 site, different groups like alkyl, vinyl, and allyl acetate were added. At R2 site, alkyl, phenyl, pyrrole, benzopyrrole, thiophenone, oxazolyl, pyrimidinyl groups, and aromatic rings were placed. Number of different atoms as in S, N, H, He, Li, F, alkyl groups, and other groups such as –O–$CH_3$, –O–$C_2H_5$, amide, cyanide, cyanate, isocyanate, –C=N, –N=C, azo, and hydrazo were added at R3 site. R4 site was filled with atoms (O, N, F, Be.) and different cyclic rings. All possible combinations of different chemical groups at four substitution sites resulted in a large combinatorial library of 86,400 compounds. The complete library was than subjected to biological activity prediction using the generated GQSAR model. 3405 compounds possessing higher activity values (>5.0) were chosen for further binding analysis against AChE and BACE-1. Compound 4 was observed to have maximum activity (6.51) in which R1 site was occupied by 2-thiophene group; R2 site was found to have F, with ethyl group and N

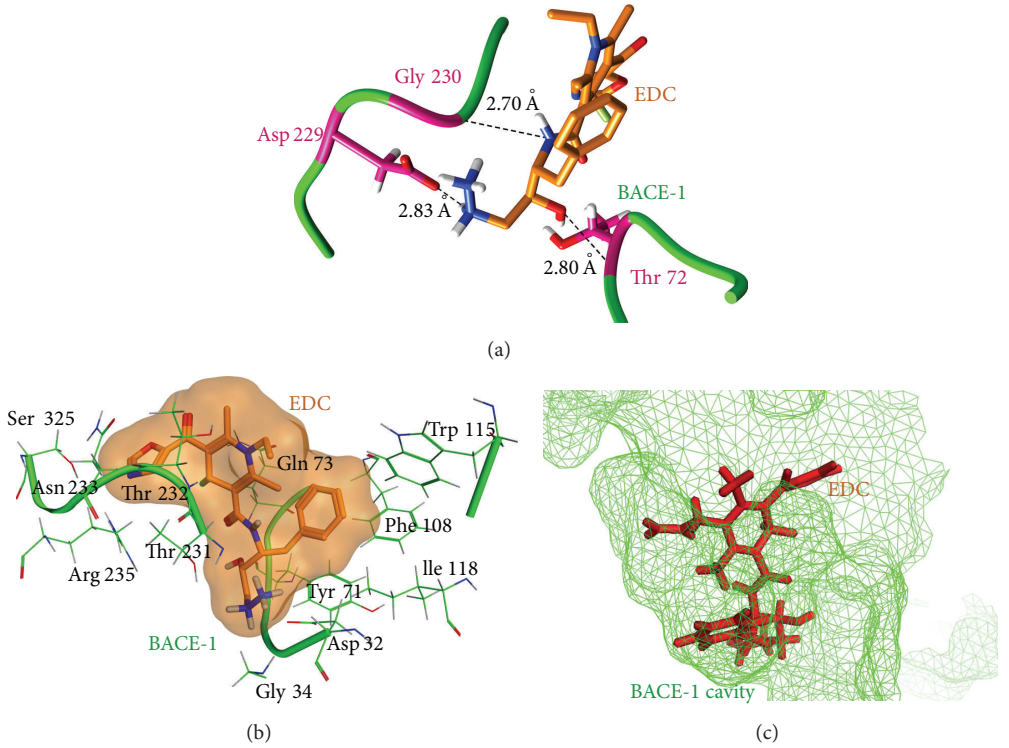



Figure 4: (a) Residues involved in hydrogen bond formation in EDC-BACE-1 complex. (b) Hydrophobically interacting amino acids in EDC-BACE-1 complex. (c) EDC bound in the active site of BACE-1.

at R3 and R4 site, respectively. Surprisingly, approximately all the high activity molecules were found to bear F atom at R2 site suggesting that the presence of F atom at R2 site plays a crucial role in activity enhancement. Therefore, constant value of 0.557 for extended topochemical descriptor R2-DeltaEpsilonA was observed. The constant low values of 1 and 0 for negatively contributing descriptors R1-NitrogensCount and R3-K3alpha depicted their role in activity enhancement.

*3.5. Docking Analysis.* Docking studies for 3405 molecules of combinatorial library were carried out against AChE and BACE-1. To filter out the chemically correct structures, molecules were converted into 3D format and then optimized using LigPrep module of Schrodinger suite which reduced the number of molecules for further analysis to 3238. Among these molecules, a total of 1310 and 1482 compounds having good binding affinity for BACE-1 and AChE, respectively, were identified using HTVS. After HTVS, the highest docking scores for both targets, BACE-1 and AChE, were found to be −10 kcal/mol and −12 kcal/mol, respectively. Compounds with Glide score above −8 kcal/mol for BACE-1 and −6 kcal/mol for AChE were then subjected to XP protocol for further refinement of Glide score. The two top scoring compounds showing dual inhibitory property against both targets were

selected for further evaluation of their mechanistic molecular mode of interaction with the target proteins.

*3.6. Interaction Mode Analysis of Docked Complexes.* The two top scoring compounds, namely, (4R)-1-ethyl-4-fluoro-N-[(2R,3S)-4-hydrazinyl-3-hydroxy-1-phenylbutan-2-yl]-2,6-dimethyl-5-(1,3-oxazole-5-carbonyl)-1,4-dihydropyridine-3-carboxamide and (4R)-4-fluoro-N-[(2R,3S)-4-hydrazinyl-3-hydroxy-1-phenylbutan-2-yl]-2,6-dimethyl-5-(1,2-oxazole-3-carbonyl)-1-(prop-2-en-1-yl)-1,4-dihydropyridine-3-carboxamide (further referred to as EDC and FDC, resp.) were found possessing dual target inhibitory capability. 2D structures of these compounds along with the common moiety are shown in Figure 3(b). The docking results revealed that EDC had the highest XP score of −15.20 kcal/mol against BACE-1 and a significant XP score of −11.92 kcal/mol against AChE. On the other hand, FDC was found to interact with strong binding affinity of −14.39 kcal/mol with BACE-1 and of −11.85 kcal/mol with AChE. Rest of all the docking parameters for these two ligand molecules with respect to both the targets were also taken into consideration and are summarized in Table 3. The $pIC_{50}$ value of both these lead compounds was 6.10 as predicted by the generated GQSAR model. The drug-like properties of the chosen compounds were also taken into account and both of the leads were found to have satisfactory values for all the essential drug-like properties such as logP value and molecular weight which are listed in Table 4.

*EDC-BACE-1 Complex.* In case of EDC-BACE-1 complex, EDC was found interacting with active site residues (Asp32,

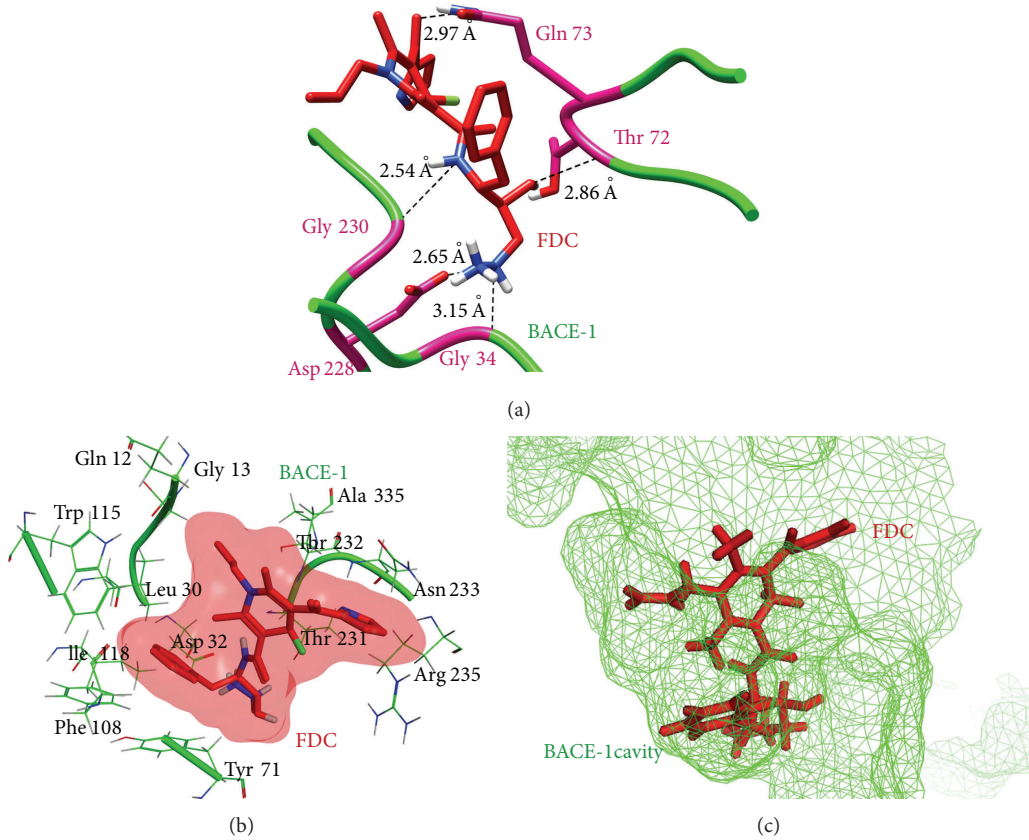



Figure 5: (a) Hydrogen bonds involved in the binding of FDC with BACE-1. (b) Residues of BACE-1 involved in the formation of hydrophobic contacts with FDC. (c) Binding of FDC molecule inside the cavity of BACE-1.

Table 3: Docking parameters for the complexes chosen after Lipinski filter.

| Complexes | Glide XP score (kcal/mol) | Glide Evdw (kcal/mol) | Glide Ecoul (kcal/mol) | Glide Emodel (kcal/mol) | Glide Energy (kcal/mol) |
|---|---|---|---|---|---|
| EDC-BACE | −15.20 | −32.97 | −23.79 | −96.58 | −56.76 |
| FDC-BACE | −14.39 | −28.88 | −30.66 | −97.88 | −59.55 |
| EDC-AChE | −11.92 | −46.03 | −18.82 | −106.16 | −64.85 |
| FDC-AChE | −11.85 | −42.17 | −16.71 | −86.54 | −58.88 |

Table 4: Molecular properties of two top scoring compounds.

| Molecular properties | Molecules | |
|---|---|---|
| | EDC | FDC |
| log *P* | 1.21 | 1.54 |
| HBD | 4 | 4 |
| HBA | 7 | 7 |
| Mol. wt. (Dalton) | 471.52 | 483.53 |
| Mol. refractivity | 124.69 | 129.09 |

HBD: hydrogen bond donar; HBA: hydrogen bond acceptor; Mol.: molecular; wt.: weight.

Gln73, Asp228, Gly230, Thr232, Asn233, and Arg235) of BACE-1 [4] with formation of four hydrogen bonds and 12 hydrophobic contacts. Among the residues lining the binding site, Asp228 and Gly230 were found participating in hydrogen bond formation with the ligand. The other residue participating in H bond formation was Thr72. The residues Asp32, Gln73, Thr232, Asn233, and Arg235 of the binding cleft along with numerous neighbouring amino acids, namely, Gly34, Tyr71, Phe108, Trp115, Ile118, Thr231, and Ser325, were observed to be involved in hydrophobic interactions with EDC. The involvement of binding site residues of BACE-1 with EDC would block the BACE-1 APP interaction, thereby preventing the processing of APP for A$\beta$ plaque formation. The binding mode of interactions can be well understood through the pictorial representation as shown in Figure 4.

*FDC-BACE-1 Complex*. Interaction analysis of this complex showed 5 hydrogen bonds and 13 hydrophobic interactions

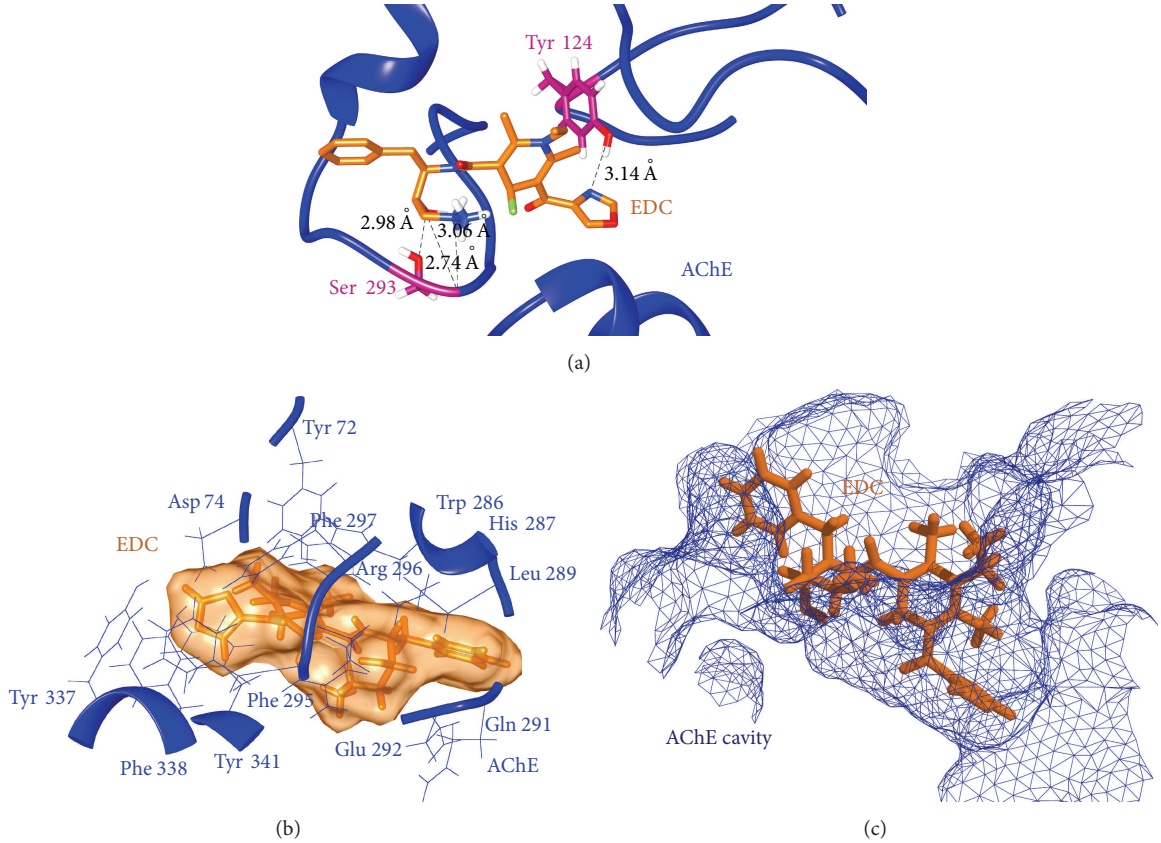



Figure 6: (a) Hydrogen bonds and their lengths as found in EDC-AChE complex. (b) Hydrophobic contacts formed between AChE residues and EDC ligand. (c) Binding of EDC molecule inside the peripheral anionic gorge of AChE.

between FDC and the binding site residues of BACE-1 as well as with some neighbouring amino acids that can be seen in Figure 5. BACE-1 residues involved in H-bond formation included Gly34, Thr72, Gln73, Asp228, and Gly230. Amino acids, Gln12, Gly13, Leu30, Asp32, Tyr71, Trp115, Ile118, Phe108, Thr231, Thr232, Asn233, Arg235, and Ser325, were making hydrophobic contacts. Binding of the ligand at this site would lead to blocking of protein-protein interactions between BACE-1 and APP.

*EDC-AChE Complex*. Since EDC was evaluated as a dual inhibitor of two different targets BACE-1 and AChE, the mechanistic mode of interaction was also analysed for EDC-AChE complex. In this complex, EDC was observed to form four hydrogen bonds and numerous hydrophobic contacts with PAS residues [23] along with some other surrounding amino acids. Two amino acids Tyr124 and Ser293 were involved in the formation of hydrogen bonds. The residues involved in hydrophobic contacts were Tyr72, Asp74, Trp286, His287, Leu289, Gln291, Glu292, Phe295, Arg296, Phe297, Tyr 337, Phe338, and Tyr341. Convincing docking score and high number of hydrogen bonds as well as hydrophobic interactions suggested EDC to be a significant inhibitor of AChE. Binding of EDC within the PAS of AChE is illustrated in Figure 6.

*FDC-AChE Complex*. Similar to EDC, the second lead molecule FDC was also evaluated for its dual inhibition property. Docking analysis for FDC-AChE complex showed that FDC was interacting with the PAS cavity of AChE. For this docked complex, three hydrogen bonds formed by two AChE residues (Glu292 and Tyr341) and FDC atoms were detected. A total of 13 hydrophobic contacts were identified with residues Tyr72, Asp74, Tyr124, Trp286, Leu289, Gln291, Ser293, Phe295, Arg296, Phe297, Tyr337, Phe338, and Gly342. The interaction mode of FDC-AChE complex showing hydrogen bonds with their respective bond length and hydrophobic interactions is illustrated in Figure 7.

## 4. Conclusion

This study is an attempt to identify novel dual inhibitors targeting BACE-1 and AChE enzymes. Structural characteristics of a set of dihydropyridine derivatives were studied using a novel group-based QSAR analysis. The GQSAR analysis revealed the importance of 2D descriptors and showed that the chemical group variations in the molecules substantially

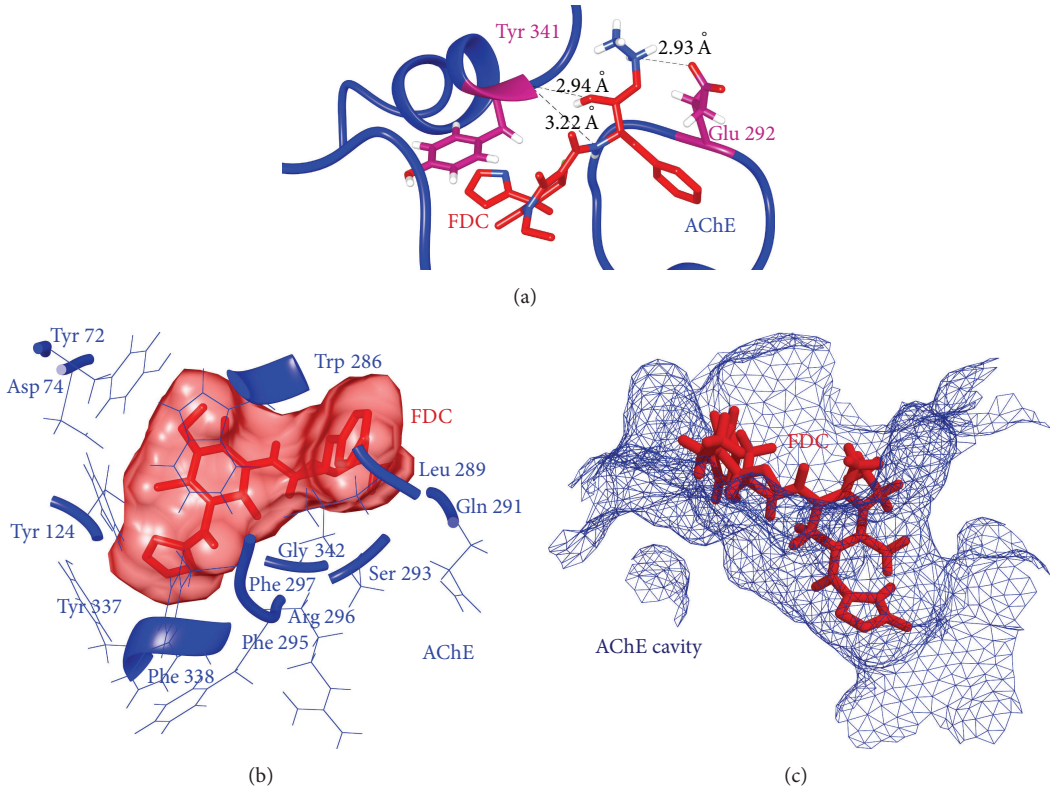



FIGURE 7: (a) Observed hydrogen bonds with their respective bond length in FDC-AChE complex. (b) FDC surrounded by hydrophobically interacting residues of AChE. (c) Ligand binding inside the PAS of FDC-AChE complex.

influenced their biological activity. We also generated a large combinatorial library of 86400 compounds by carrying out substitutions at four different sites of DHP. GQSAR model was utilized further for activity prediction of prepared combinatorial library. The two compounds (EDC and FDC) having high predicted inhibitory activity and the highest docking scores against both of the targets were identified as possessing dual inhibitory properties. We have also provided mechanistic insights into the binding mode of action of these leads. The enhanced predicted activity, high binding score, and the presence of crucial drug like molecular properties provide substantial evidence for consideration of these compounds as potent dual inhibitors for future prospective of AD treatment. This information could be of high value for design and development of novel multitargeted drugs against AD possessing improved binding properties and low toxicity to human cells.

## Conflict of Interests

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

## Acknowledgments

Abhinav Grover is thankful to Jawaharlal Nehru University for usage of all computational facilities. The work in this paper is supported by grants to Abhinav Grover from Science and Engineering Research Board, Department of Science and Technology, Government of India.

## References

[1] G. Benzi and A. Moretti, "Is there a rationale for the use of acetylcholinesterase inhibitors in the therapy of Alzheimer's disease?" *European Journal of Pharmacology*, vol. 346, no. 1, pp. 1–13, 1998.

[2] F. Belluti, M. Bartolini, G. Bottegoni et al., "Benzophenone-based derivatives: a novel series of potent and selective dual inhibitors of acetylcholinesterase and acetylcholinesterase-induced beta-amyloid aggregation," *European Journal of Medicinal Chemistry*, vol. 46, no. 5, pp. 1682–1693, 2011.

[3] D. M. Walsh and D. J. Selkoe, "Deciphering the molecular basis of memory failure in Alzheimer's disease," *Neuron*, vol. 44, no. 1, pp. 181–193, 2004.

[4] S.-J. Choi, J.-H. Cho, I. Im et al., "Design and synthesis of 1,4-dihydropyridine derivatives as BACE-1 inhibitors," *European Journal of Medicinal Chemistry*, vol. 45, no. 6, pp. 2578–2590, 2010.

[5] H.-W. Klafki, M. Staufenbiel, J. Kornhuber, and J. Wiltfang, "Therapeutic approaches to Alzheimer's disease," *Brain*, vol. 129, no. 11, pp. 2840–2855, 2006.

[6] P. D. Edwards, J. S. Albert, M. Sylvester et al., "Application of fragment-based lead generation to the discovery of novel, cyclic amidine $\beta$-secretase inhibitors with nanomolar potency,

cellular activity, and high ligand efficiency," *Journal of Medicinal Chemistry*, vol. 50, no. 24, pp. 5912–5925, 2007.

[7] A. Aguzzi and T. O'Connor, "Protein aggregation diseases: pathogenicity and therapeutic perspectives," *Nature Reviews Drug Discovery*, vol. 9, no. 3, pp. 237–248, 2010.

[8] C. A. Kelly, R. J. Harvey, and H. Cayton, "Drug treatments for Alzheimer's disease," *British Medical Journal*, vol. 314, no. 7082, pp. 693–694, 1997.

[9] P. J. Whitehouse, "Cholinergic therapy in dementia," *Acta Neurologica Scandinavica, Supplement*, vol. 88, no. 149, pp. 42–45, 1993.

[10] L. J. Scott and K. L. Goa, "Galantamine: a review of its use in Alzheimer's disease," *Drugs*, vol. 60, no. 5, pp. 1095–1122, 2000.

[11] A. Yan and K. Wang, "Quantitative structure and bioactivity relationship study on human acetylcholinesterase inhibitors," *Bioorganic & Medicinal Chemistry Letters*, vol. 22, no. 9, pp. 3336–3342, 2012.

[12] J. Birks, "Cholinesterase inhibitors for Alzheimer's disease," *Cochrane Database of Systematic Reviews*, no. 1, Article ID CD005593, 2006.

[13] S. A. Areosa, F. Sherriff, and R. McShane, "Memantine for dementia," *Cochrane Database of Systematic Reviews*, no. 2, Article ID CD003154, 2006.

[14] M. W. Weiner, C. Sadowsky, J. Saxton et al., "Magnetic resonance imaging and neuropsychological results from a trial of memantine in Alzheimer's disease," *Alzheimer's and Dementia*, vol. 7, no. 4, pp. 425–435, 2011.

[15] D. J. Selkoe, "Translating cell biology into therapeutic advances in Alzheimer's disease," *Nature*, vol. 399, supplement, pp. A23–A31, 1999.

[16] T. Guo and D. W. Hobbs, "Development of BACE1 inhibitors for Alzheimer's disease," *Current Medicinal Chemistry*, vol. 13, no. 15, pp. 1811–1829, 2006.

[17] L. Piazzi, A. Cavalli, F. Colizzi et al., "Multi-target-directed coumarin derivatives: hAChE and BACE1 inhibitors as potential anti-Alzheimer compounds," *Bioorganic & Medicinal Chemistry Letters*, vol. 18, no. 1, pp. 423–426, 2008.

[18] A. Cavalli, M. L. Bolognesi, A. Mìnarini et al., "Multi-target-directed ligands to combat neurodegenerative diseases," *Journal of Medicinal Chemistry*, vol. 51, no. 3, pp. 347–372, 2008.

[19] L. Piazzi, A. Rampa, A. Bisi et al., "3-(4-[benzyl(methyl)amino]methyl-phenyl)-6,7-dimethoxy-2H-2-chromenone (AP2238) inhibits both acetylcholinesterase and acetylcholinesterase-induced $\beta$-amyloid aggregation: a dual function lead for Alzheimer's disease therapy," *Journal of Medicinal Chemistry*, vol. 46, no. 12, pp. 2279–2282, 2003.

[20] S. L. Cole and R. Vassar, "BACE1 structure and function in health and Alzheimer's disease," *Current Alzheimer Research*, vol. 5, no. 2, pp. 100–120, 2008.

[21] R. E. Tanzi and L. Bertram, "Twenty years of the Alzheimer's disease amyloid hypothesis: a genetic perspective," *Cell*, vol. 120, no. 4, pp. 545–555, 2005.

[22] L. Zou, R. Yang, P. Zhang, and Y. Dai, "The enhancement of amyloid precursor protein and $\beta$-site amyloid cleavage enzyme 1 interaction: amyloid-$\beta$ production with aging," *International Journal of Molecular Medicine*, vol. 25, no. 3, pp. 401–407, 2010.

[23] R. Sheng, X. Lin, J. Zhang et al., "Design, synthesis and evaluation of flavonoid derivatives as potent AChE inhibitors," *Bioorganic & Medicinal Chemistry*, vol. 17, no. 18, pp. 6692–6698, 2009.

[24] G. Bottegoni, A. D. Favia, M. Recanatini, and A. Cavalli, "The role of fragment-based and computational methods in polypharmacology," *Drug Discovery Today*, vol. 17, no. 1-2, pp. 23–34, 2012.

[25] C. B. Breitenlechner, T. Wegge, L. Berillon et al., "Structure-based optimization of novel azepane derivatives as PKB inhibitors," *Journal of Medicinal Chemistry*, vol. 47, no. 6, pp. 1375–1390, 2004.

[26] Y. Luo, A. R. Shoemaker, X. Liu et al., "Potent and selective inhibitors of Akt kinases slow the progress of tumors in vivo," *Molecular Cancer Therapeutics*, vol. 4, no. 6, pp. 977–986, 2005.

[27] S. K. Deshpande, "Molecule fragmentation scheme and method for designing new molecules," in *Google Patents*, 2008.

[28] J. Verma, V. M. Khedkar, and E. C. Coutinho, "3D-QSAR in drug design—a review," *Current Topics in Medicinal Chemistry*, vol. 10, no. 1, pp. 95–115, 2010.

[29] *VLifeMDS: Molecular Design Suite*, Vlife Sciences Technologies Pvt. Ltd., Pune, India, 3rd edition, 2004.

[30] K. Baumann, "An alignment-independent versatile structure descriptor for QSAR and QSPR based on the distribution of molecular features," *Journal of Chemical Information and Computer Sciences*, vol. 42, no. 1, pp. 26–35, 2002.

[31] S. Ajmani and S. A. Kulkarni, "Application of GQSAR for scaffold hopping and lead optimization in multitarget inhibitors," *Molecular Informatics*, vol. 31, no. 6–7, pp. 473–490, 2012.

[32] R. D. Cramer III, D. E. Patterson, and J. D. Bunce, "Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins," *Journal of the American Chemical Society*, vol. 110, no. 18, pp. 5959–5967, 1988.

[33] A. Afantitis, G. Melagraki, H. Sarimveis, O. Igglessi-Markopoulou, and G. Kollias, "A novel QSAR model for predicting the inhibition of CXCR3 receptor by 4-N-aryl-[1,4] diazepane ureas," *European Journal of Medicinal Chemistry*, vol. 44, no. 2, pp. 877–884, 2009.

[34] A. Golbraikh and A. Tropsha, "Beware of q2!," *Journal of Molecular Graphics and Modelling*, vol. 20, no. 4, pp. 269–276, 2002.

[35] www.rcsb.org/pdb.

[36] Schrodinger, "Schrodinger suite," LLC, New York, NY, USA, 2009.

[37] A. C. Wallace, R. A. Laskowski, and J. M. Thornton, "LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions," *Protein Engineering*, vol. 8, no. 2, pp. 127–134, 1995.

[38] A. Nayyar, V. Monga, A. Malde, E. Coutinho, and R. Jain, "Synthesis, anti-tuberculosis activity, and 3D-QSAR study of 4-(adamantan-1-yl)-2-substituted quinolines," *Bioorganic and Medicinal Chemistry*, vol. 15, no. 2, pp. 626–640, 2007.

[39] D. Pal, C. Sengupta, and A. De, "A new topochemical descriptor (TAU) in molecular connectivity concept: part I—aliphatic compounds," *Indian Journal of Chemistry B*, vol. 27, pp. 734–739, 1988.

[40] D. K. Pal, C. Sengupta, and A. U. De, "Introduction of a novel topochemical index and exploitation of group connectivity concept to achieve predictability in QSAR and RDD," *Indian Journal of Chemistry B*, vol. 28, no. 3, pp. 261–267, 1989.

[41] K. Roy and R. N. Das, "On some novel extended topochemical atom (ETA) parameters for effective encoding of chemical information and modelling of fundamental physicochemical properties," *SAR and QSAR in Environmental Research*, vol. 22, no. 5-6, pp. 451–472, 2011.

[42] K. Roy and G. Ghosh, "Exploring QSARs with Extended Topochemical Atom (ETA) indices for modeling chemical and drug toxicity," *Current Pharmaceutical Design*, vol. 16, no. 24, pp. 2625–2639, 2010.

[43] L. H. Hall and L. B. Kier, "The molecular connectivity chi indexes and kappa shape indexes in structure-property modeling," *Reviews in Computational Chemistry*, vol. 2, pp. 367–422, 1991.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 248090, 8 pages
http://dx.doi.org/10.1155/2014/248090

*Research Article*

# MultiRankSeq: Multiperspective Approach for RNAseq Differential Expression Analysis and Quality Control

**Yan Guo, Shilin Zhao, Fei Ye, Quanhu Sheng, and Yu Shyr**

*Center for Quantitative Sciences, Vanderbilt University, Nashville, TN 37027, USA*

Correspondence should be addressed to Yan Guo; yan.guo@vanderbilt.edu and Yu Shyr; yu.shyr@vanderbilt.edu

Received 23 November 2013; Revised 1 February 2014; Accepted 15 March 2014; Published 27 May 2014

Academic Editor: Jason E. Mcdermott



*Background*. After a decade of microarray technology dominating the field of high-throughput gene expression profiling, the introduction of RNAseq has revolutionized gene expression research. While RNAseq provides more abundant information than microarray, its analysis has proved considerably more complicated. To date, no consensus has been reached on the best approach for RNAseq-based differential expression analysis. Not surprisingly, different studies have drawn different conclusions as to the best approach to identify differentially expressed genes based upon their own criteria and scenarios considered. Furthermore, the lack of effective quality control may lead to misleading results interpretation and erroneous conclusions. To solve these aforementioned problems, we propose a simple yet safe and practical rank-sum approach for RNAseq-based differential gene expression analysis named MultiRankSeq. MultiRankSeq first performs quality control assessment. For data meeting the quality control criteria, MultiRankSeq compares the study groups using several of the most commonly applied analytical methods and combines their results to generate a new rank-sum interpretation. MultiRankSeq provides a unique analysis approach to RNAseq differential expression analysis. MultiRankSeq is written in R, and it is easily applicable. Detailed graphical and tabular analysis reports can be generated with a single command line.

## 1. Introduction

Gene expression refers to the appearance of a characteristic or effect in the phenotype that can be attributed to a particular gene. The development of microarray technologies has helped biomedical researchers make significant advances in the last decade by allowing high-throughput gene expression screening on all known genes. The introduction of RNAseq technology further revolutionized the field of gene expression research with accurate measurements of transcripts instead of estimating relative measures and with the detection of structural variants such as splicing and gene fusion. RNAseq uses next-generation sequencing (NGS) technologies to sequence cDNA that has been reverse transcribed from RNA. It is commonly believed to be superior to microarray technology due to its ability to quantify gene expression at higher resolution (exon and CDS level) and detect structural variations. As early as 2008 [1], RNAseq has been hailed as the eventual replacement of microarray technology, and since then, multiple studies [2–6] have also illustrated the advantages of RNAseq and come to similar conclusions by analyzing real data or through thorough simulation study.

RNAseq technology introduces new and exciting opportunities to researchers in the field of biomedical research as well as stiff analysis challenges for bioinformaticians. The rich genomic information RNAseq technology contains gives RNAseq the decisive advantage over microarray but adds complication in the analysis phase. Several unique characteristics contribute to the difficulty of RNAseq data analysis. First, in RNAseq, the expression values are usually directly represented by the number of reads or adjusted number of reads aligned to a gene. For a nonexpressed gene, zero reads are aligned to the gene's genomic span. Because microarray technology is based on fluorescence intensity, there is always a nonzero background intensity, allowing microarray data to be log-transformed. In contrast, due to the large number of zeros for nonexpressed genes in RNAseq data (often around 50%), log transformation results in many invalid mathematical operations. The typical range of an RNAseq dataset is huge, between 0 and 10,000+ compared

to microarray's 2 to 15 after RMA normalization (with log 2 transformation). Because RNAseq's expression value starts from 0, large fold change can result from two very small expression values. For example, the fold change between 0.1 and 0.001 is 100, but both 0.1 and 0.001 should be considered nonexpressed. In addition, there are many sequencing and alignment artifact that can skew RNAseq data such as errors from demultiplexing and alignment ambiguity caused by highly homologous genomic regions.

These bioinformatic challenges create difficulty for RNAseq data analysis. In this study, we focus on the inconsistency of differential expression analyses and the lack of multiperspective quality control. First and foremost, to date the research community has yet to come to a consensus on the best of a multitude of different approaches for differential gene expression analysis of RNAseq data. The pioneer of RNAseq differential expression analysis, Cufflinks, is based on reads per kilobase per million mapped reads (RPKM) [7] and fragments per kilobase of transcript per million mapped reads (FPKM) [8]. A similar approach is RNAseq by expectation-maximization (RSEM) [9]. RPKM, FPKM, and RSEM can be classified as read normalization-based methods.

Another type of RNAseq differential gene expression analysis is based on read count. Many read count-based methods have been developed including DESeq [10], DEGseq [11], edgeR [12], baySeq [13], TSPM [14], NBPSeq [15], SAM-seq [16], and NOIseq [17]. Since they are dealing with RNAseq count data, the majority of them are based on Poisson or negative binomial distributions. But there are implementation details that separate them. For example, edgeR moderates dispersion estimates toward a trended mean, whereas DESeq takes the maximum of the individual dispersion estimates and the dispersion-mean trend, and baySeq uses an empirical Bayes approach assuming a negative binomial distribution of the data. Several studies [4, 18–21] have attempted to evaluate different normalization and differential gene comparison methods for RNAseq data. Although no final conclusion can be reached, through simulation analysis of real data, it has been found by multiple sources that DESeq, edgeR, and baySeq were able to maintain a reasonable false-positive rate without any loss of power. More recently, nonparametric approaches, such as SAMseq and NOIseq, were proposed aiming to overcome limitations of aforementioned parametric as they can be influenced by "outliers" in the data. In this paper, we focus on the currently widely applied parametric methods for RNAseq gene expression analysis, but it is easy to incorporate other methods including nonparametric approaches in MultiRankSeq.

In addition to the lack of consensus on the best statistical method, another issue associated with RNAseq data analysis is the lack of complete quality control. The majority of high-throughput sequencing quality control tools were designed exclusively for raw data. Previously, we have proposed a three-stage quality control [22] strategy for exome sequencing analysis that emphasizes the need to implement quality control at all stages of exome sequencing processing: raw data, alignment, and variant calling. The same idea can be easily adapted to the three stages of RNAseq analysis as well: raw data, expression quantification, and differential expression analysis. There have been several tools designed for RNAseq quality control such as RNA-SeQC [23] and RSeQC [24]. These tools generally target the raw data and expression quantification steps by calculating quality control parameters such as read coverage, and GC bias. However, quality control on differential expression analysis is often not considered.

In this paper, we propose a multimethod rank-sum approach for RNAseq expression analysis that combines multiple RNAseq differential expression analysis packages. Combining multiple methods of RNAseq data analysis has been previously suggested. For example, Robles et al. suggested that using a combination of multiple packages may overcome the possible bias susceptibility of a given package to a particular dataset of interest [20]. In another study by Soneson et al., the authors suggested the use of transformation-based approaches (the variance stabilizing transformation provided in the DESeq R package and the voom transformation from the limma R package) combined with LIMMA [25], which performed well under many conditions. In this study, we present a tool, MultiRankSeq, for RNAseq differential gene expression analysis. This tool offers rank-sum-based differential gene expression analysis, comprehensive diagnostic quality control assessment, and automated graphical reports. The input of MultiRankSeq is a read-count matrix. MultiRankSeq is implemented in R, and it is freely available for public use. MultiRankSeq can be downloaded from https://github.com/slzhao/MultiRankSeq.

## 2. Materials and Methods

Differential expression analysis can only be conducted between two phenotypes such as tumor versus normal or treated versus untreated. The ideal assumption for conducting differential expression analysis is that gene expression patterns are similar for samples within the same phenotype group (i.e., relatively homogeneous). Sometimes, however, this assumption does not hold true. A sample from one phenotype group may be more similar to the samples from the other phenotype groups based merely on expression profile. Unfortunately, the homogeneity of gene expression patterns within the same group is not always checked before conducting differential gene expression analysis. One simple yet effective way to check this assumption is through cluster analysis. Clustering refers to the task of grouping together a set of samples with similar gene expression patterns. To determine the pairwise sample gene expression pattern similarity, a similarity or distance measurement must be employed. In MultiRankSeq, we chose to use Spearman's correlation coefficient. Because the input of MultiRankSeq is read count, Spearman's correlation coefficient is used as it is more robust to handle skewness and outliers than a parametric method. MultiRankSeq performs unsupervised clustering using all genes to best represent the raw expression pattern of each sample. Samples clustered outside the true phenotype group are considered to be misclassified. This could occur due to sample contamination or other technical reasons. If the majority of the genes lack variation among

samples, the cluster may be unrepresentative of the true phenotype group. To alleviate this, MultiRankSeq performs additional cluster analyses on read counts filtered by the top 5% and 10% coefficient of variation. In theory, the clustering should improve as more stringent coefficient of variation cutoffs is used.

MultiRankSeq performs a gene expression integrity check by drawing the read count distribution and the normalized read count distribution. Normalization is done by dividing each gene's read count by the total read count of all genes in this sample. One of the unique optional features offered by MultiRankSeq is the ability to detect batch effect. Batch effects can be a problem with RNAseq data [26]. The most common sequencing failures often occur nonrandomly by lane, flow cell, run, or machine. MultiRankSeq recognizes and records the machine name, run ID, flow cell ID, and lane ID of an experiment from either the FASTQ file or BAM file. Based on this information, MultiRankSeq determines whether batch effect exists using the nonparametric Kruskal-Wallis [27] test and Fligner-Killeen test of homogeneity of variances [28]. MultiRankSeq uses boxplots and correlation matrices to demonstrate the expression variation between samples.

The idea behind MultiRankSeq's algorithm for integrating the results from multiple RNAseq analysis tools is based on the same analytic principle as the weighted flexible compound covariate method (WFCCM) [29]. WFCCM was designed to integrate the findings of multiple analysis methods (e.g., Kruskal-Wallis test, Fisher's exact test, permutation $t$-test, SAM, WGA, and modified info score) to identify the most significant gene expression associated with biological status and thereby allow for class-prediction modeling based on differential gene expression. In other words, WFCCM extends the compound covariate method by allowing for more than one statistical analysis method to be considered in the covariate and reduces the dimensionality of an analytic problem by generating a single covariate calculated as a weighted sum of the class predictors identified as most important.

Based on previous studies [18–21] and our own evaluation [30], we selected three methods for MultiRankSeq—DESeq, edgeR, and baySeq—and combined their algorithms in MultiRankSeq. The tabular report provided by MultiRankSeq includes log 2 fold change, raw $P$-value, and false discovery rate (FDR) adjusted $P$ value from all three methods except for baySeq because the Bayesian-based method does not calculate fold change. We rank the genes based on the raw $P$ value rather than FDR-adjusted $P$ value because the latter often has a large number of tied values. The sum of the rankings from all three methods is reported in the last column of the tabular report to serve as an overall ranking of genes. The sum of ranks can be used as a confidence level of differential expression. The smallest rank sums indicate differentially expressed genes are consistent among the three methods.

MultiRankSeq provides concordance analysis of the results from the three methods and detailed visualizations using various figures such as Venn diagram, heatmap, and scalable volcano plot to summarize and illustrate the analysis results. Venn diagrams demonstrate the logical relations between the three methods based on parameters such as fold change, adjusted $P$ value, and top ranked genes. The heatmap is used for visualization of gene expression patterns in a color scale. The correlation scatter plots depict the general consistency between the methods. The scalable volcano plot can help the user visualize the genes based on fold change, $P$ value, and ranking simultaneously. The rank of the gene is reflected by the size of the corresponding dot on the volcano plot.

## 3. Result and Discussion

*3.1. Results.* We demonstrate MultiRankSeq using two example datasets from the TCGA breast cancer and performed analysis using MultiRankSeq V1.1.2. This version of MultiRankSeq uses edgeR 3.4.2, DESeq 1.14.0, and baySeq 1.16.0 as the primary three differential expression analysis packages. The first example dataset contains RNAseq data from 3 tumors and 3 adjacent normal tissues from same patients (TCGA-A7-A0D9, TCGA-BH-A0B3, and TCGA-BH-A0BJ). This example is used to show the MultiRankSeq's cluster functionality.

When using unfiltered data, an adjacent normal sample was clustered with the tumor group (Figure 1(a)). Normally, we may consider this sample problematic and remove it from the analysis; however, cluster result using genes with the top 5% coefficient of variation showed the correct grouping (Figure 1(b)). The misclassified sample in Figure 1(a) is likely due to noise caused by genes that lack variation among samples. Therefore, part of its information can be used for the analysis instead of completely removing the sample. Figure 2 shows additional quality control matrix produced based on the example of dataset 1.

The second example dataset also contains RNAseq data from 3 tumors and 3 adjacent normal tissues from same patients (TCGA-BH-A0BM, TCGA-BH-A0C0, and TCGA-BH-A0DK). Using this example, we demonstrate the complete MultiRankSeq's functionality.

The example result figure produced by MultiRankseq using this example can be seen in Figure 3. The full HTML reports of MultiRankSeq from the example data can be found at the tool's hosting website. The complete R command used to generate the results can be viewed as follows:

```
library(MultiRankSeq);
#Load the downloaded data into R, and generate group definition;
Figure 1<read.csv("TcgaFigure 1.csv",header=T,row.names=1,check.names=F);
Figure 3<read.csv("TcgaFigure 3.csv",header=T,row.names=1,check.names=F);
group=c(0,0,0,1,1,1);
#Generate report;
reportF1<-MultiRankSeqReport; (output="report Figure 1.html",rawCounts=TcgaFigure 1, group=group);
```

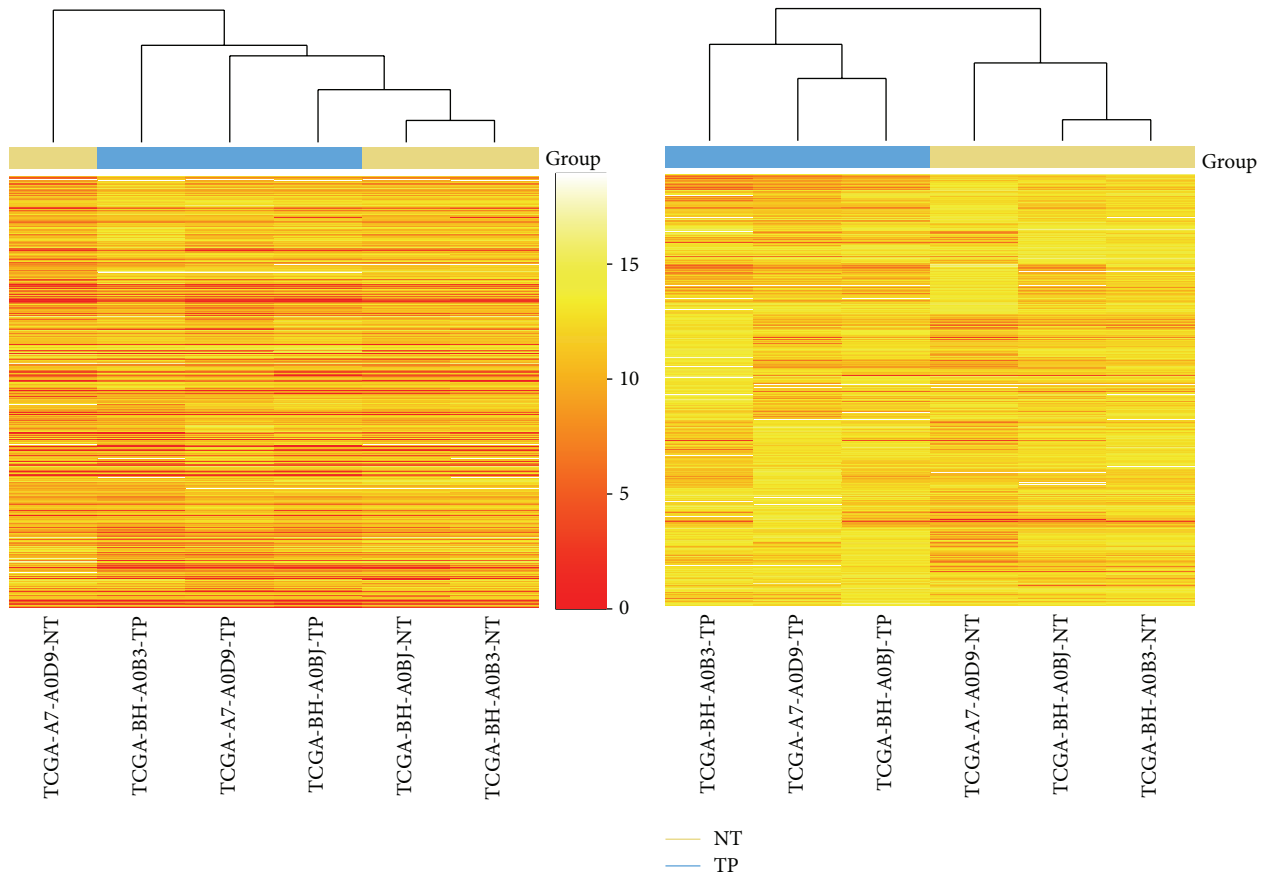



Figure 1: (a) Cluster result using all genes shows control 1 clustered together with disease group. (b) Cluster results using genes with top 5% coefficient of variation, control group, and disease group are now clustered correctly.

reportF3<-MultiRankSeqReport; (output="report Figure 3.html",rawCounts=TcgaFigure 3, group= group).

Even though, majority of the time, DESeq, edgeR, and baySeq agree with each other, there is still occasional disagreement. This can be demonstrated through the second example. We observed that when there is a large variation in the read counts, especially when the homogeneity assumption seems to be violated, the 3 methods can disagree with each other significantly. One particular example is the gene IGHG2 (Table 1). Based on the FDR-adjusted $P$ value, only edgeR considered it to be significant. The FDR-adjusted $P$ values were 0.047, 0.28, and 0.91, respectively, for edgeR, DESeq, and baySeq. We then performed an additional analysis using Cuffdiff [31]. Cuffdiff agreed with edgeR with an FDR-adjusted $P$ value <0.001. The log 2 fold changes of IGHG2 produced also spans a large range (from 2.92 to 5.83). After adjusting for the total number of reads, the variation becomes less obvious with each of the methods (Table 2). However, in this particular case, edgeR seems to have performed more effective variation stabilization.

In terms of number of winner genes (adjusted $P < 0.05$) identified, the three methods differ hugely in example 2 (DESeq = 1118, edgeR = 743, and baySeq = 63). We performed network, pathway, and biological functionality analysis using ingenuity. The results are split into seven categories: genes identified by DESeq, edgeR, baySeq, singleton genes that identified each of the three methods, and overlapped genes among the three methods. Singleton gene means this gene is only identified by one method. The top five networks, biological functionalities, and canonical pathways are reported (see Table S1 in Supplementary Material available online at http://dx.doi.org/10.1155/2014/248090). The functionality results of DESeq and edgeR are similar because they have large overlap; the results of baySeq are more unique because of less overlap with other methods.

Table 1: Analysis difference for IGHG2.

| Method | Adjusted $P$ value | log 2 FC | Rank |
|---|---|---|---|
| DESeq | 0.278 | 3.00 | 2572 |
| edgeR | 0.047 | 2.92 | 712 |
| baySeq | 0.907 | NA | 24962 |
| Cuffdiff | <0.001 | 5.83 | 13 |

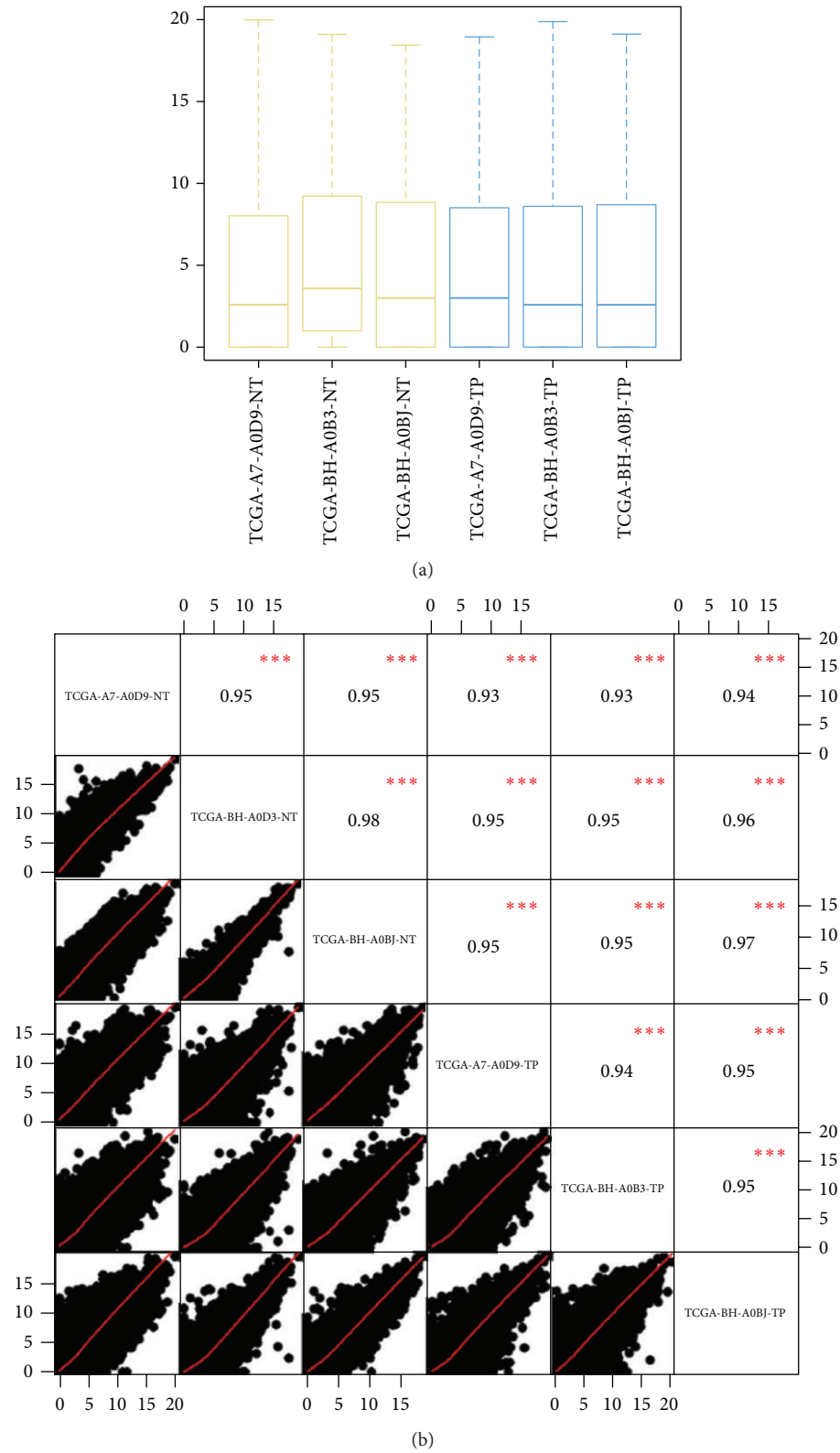



Figure 2: (a) Boxplots of gene raw read count. (b) Correlation matrix of all genes between all pairs of samples using raw read count.

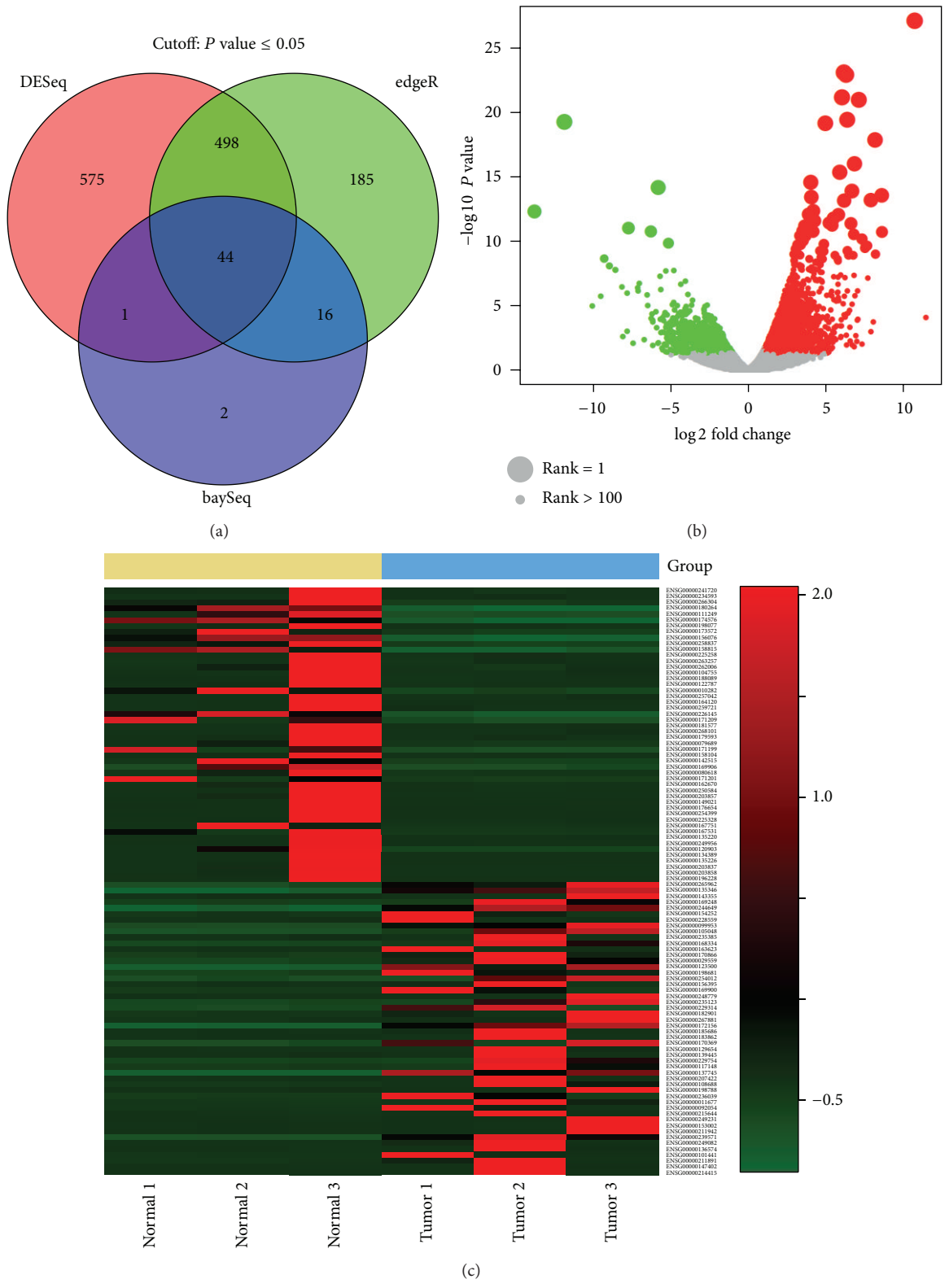



FIGURE 3: (a) Venn diagram of differential expression analyses by DESeq, edgeR, and baySeq. The Venn diagram can be drawn based on *P* value, fold change, or rank. (b) Scalable volcano plot representing fold change, *P* value, and rank. Rank is presented as the size of the circle, and larger size denotes higher ranking. (c) Heatmap of top differentially expressed genes. MultiRankSeq produces heatmap based on *P* value, fold change, and rank; only genes selected by fold change are shown here.

Table 2: Read count of samples for IGHG2 gene.

| | Disease 1 | Disease 2 | Disease 3 | Control 1 | Control 2 | Control 3 |
|---|---|---|---|---|---|---|
| Read count (IGHG2) | 391 | 2038 | 338 | 634 | 10282 | 1764 |
| Total read count | 49870084 | 65550902 | 71454121 | 35641084 | 44863975 | 49052840 |
| Adjusted read Count[1] | 78 | 311 | 47 | 178 | 2292 | 360 |

[1] Adjusted read count of gene A is computed as read count of a gene A divided by total read count of the sample times a constant.

*3.2. Discussion.* Performing traditional comparative statistical analysis methods such as $t$-test or Wilcoxon rank-sum test requires at least 3 samples for variation to generate meaningful results. Such limitations also apply to DESeq, edgeR, and baySeq. Cuffdiff, however, can assign $P$ values even for 1 sample versus 1 sample. In order to do this, it makes the assumption that similarly expressed genes have similar variance and the majority of the genes are not differentially expressed. In many studies, these assumptions will hold true, but including multiple samples in a group will always generate more robust results. The current version of MultiRankSeq only considers methods based on read count data. However, it is our goal to incorporate Cuffdiff in the future research.

RNAseq data is difficult to analyze and sometimes is methodology-dependent as previously discussed. MultiRankSeq tackles this problem from a different perspective by combining ranked results from multiple well-rated RNAseq analysis methods. This approach brings more confidence to the selection of truly differentially expressed genes. Another novelty that MultiRankSeq brings is the bridging of the gap between quality control and statistical analysis. The report generated by MultiRankSeq is comprehensive and helps the user better appreciate the power and complexity of RNAseq data. MultiRankSeq is based on an intuitive idea of combining multiple methods yet very practical. Because MultiRankSeq is designed with user friendliness and flexibility in mind, additional RNAseq analysis programs can be easily added to it in the future if needed. In conclusion, MultiRankSeq is a simple framework of RNAseq data analysis which provides tremendous convenience and an alternative perspective for researchers who conduct routine RNAseq analysis.

We do not claim that combining the results of multiple methods will always produce more accurate result. There are the scenarios when the minority method is corrected. Thus combining three methods may still produce false positive results. However, if multiple methods agree, the probability of generating the true positive results will most likely to increase. The goal of MultiRankSeq is to provide the user with higher confidence to pick a gene that is significantly differentially expressed with high probability.

## Conflict of Interests

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

## Authors' Contribution

Yan Guo and Shilin Zhao contributed equally to this paper.

## References


[1] J. Shendure, "The beginning of the end for microarrays?" *Nature Methods*, vol. 5, no. 7, pp. 585–587, 2008.

[2] Y. W. Asmann, E. W. Klee, E. A. Thompson et al., "3′ tag digital gene expression profiling of human brain and universal reference RNA using illumina genome analyzer," *BMC Genomics*, vol. 10, article 531, 2009.

[3] N. Cloonan, A. R. Forrest, G. Kolle et al., "Stem cell transcriptome profiling via massive-scale mRNA sequencing," *Nature Methods*, vol. 5, no. 7, pp. 613–619, 2008.

[4] Y. Guo, Q. Sheng, J. Li, F. Ye, D. C. Samuels, and Y. Shyr, "Large scale comparison of gene expression levels by microarrays and RNAseq using TCGA data," *PLoS ONE*, vol. 8, no. 8, Article ID e71462, 2013.

[5] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad, "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays," *Genome Research*, vol. 18, no. 9, pp. 1509–1517, 2008.

[6] Z. Wang, M. Gerstein, and M. Snyder, "RNA-seq: a revolutionary tool for transcriptomics," *Nature Reviews Genetics*, vol. 10, no. 1, pp. 57–63, 2009.

[7] 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.

[8] C. Trapnell, B. A. Williams, G. Pertea et al., "Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation," *Nature Biotechnology*, vol. 28, no. 5, pp. 511–515, 2010.

[9] B. Li and C. N. Dewey, "RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome," *BMC Bioinformatics*, vol. 12, article 323, 2011.

[10] S. Anders and W. Huber, "Differential expression analysis for sequence count data," *Genome Biology*, vol. 11, no. 10, article R106, 2010.

[11] L. Wang, Z. Feng, X. Wang, 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, 2009.

[12] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: a bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, no. 1, pp. 139–140, 2010.

[13] T. J. Hardcastle and K. A. Kelly, "BaySeq: empirical Bayesian methods for identifying differential expression in sequence count data," *BMC Bioinformatics*, vol. 11, article 422, 2010.

[14] P. L. Auer and R. W. Doerge, "A two-stage poisson model for testing RNA-seq data," *Statistical Applications in Genetics and Molecular Biology*, vol. 10, no. 1, article 26, 2011.

[15] Y. Di, D. W. Schafer, J. S. Cumbie, and J. H. Chang, "The NBP negative binomial model for assessing differential gene expression from RNA-seq," *Statistical Applications in Genetics and Molecular Biology*, vol. 10, no. 1, article 24, 2011.

[16] J. Li and R. Tibshirani, "Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-seq data," *Statistical Methods in Medical Research*, vol. 22, no. 5, pp. 519–536, 2011.

[17] S. Tarazona, F. García-Alcalde, J. Dopazo, A. Ferrer, and A. Conesa, "Differential expression in RNA-seq: a matter of depth," *Genome Research*, vol. 21, no. 12, pp. 2213–2223, 2011.

[18] M. A. Dillies, A. Rau, J. Aubert et al., "A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis," *Brief Bioinform*, vol. 14, no. 6, pp. 671–683, 2012.

[19] V. M. Kvam, P. Liu, and S. Yaqing, "A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data," *The American Journal of Botany*, vol. 99, no. 2, pp. 248–256, 2012.

[20] J. A. Robles, S. E. Qureshi, S. J. Stephen, S. R. Wilson, C. J. Burden, and J. M. Taylor, "Efficient experimental design and analysis strategies for the detection of differential expression using RNA-sequencing," *BMC Genomics*, vol. 13, article 484, 2012.

[21] C. Soneson and M. Delorenzi, "A comparison of methods for differential expression analysis of RNA-seq data," *BMC Bioinformatics*, vol. 14, article 91, 2013.

[22] Y. Guo, F. Ye, Q. Sheng, T. Clark, and D. C. Samuels, "Three-stage quality control strategies for DNA re-sequencing data," *Briefings in Bioinformatics*, 2014.

[23] D. S. DeLuca, J. Z. Levin, A. Sivachenko et al., "RNA-SeQC: RNA-seq metrics for quality control and process optimization," *Bioinformatics*, vol. 28, no. 11, pp. 1530–1532, 2012.

[24] L. Wang, S. Wang, and W. Li, "RSeQC: quality control of RNA-seq experiments," *Bioinformatics*, vol. 28, no. 16, pp. 2184–2185, 2012.

[25] I. Diboun, L. Wernisch, C. A. Orengo, and M. Koltzenburg, "Microarray analysis after RNA amplification can detect pronounced differences in gene expression using limma," *BMC Genomics*, vol. 7, article 252, 2006.

[26] K. D. Hansen, R. A. Irizarry, and Z. Wu, "Removing technical variability in RNA-seq data using conditional quantile normalization," *Biostatistics*, vol. 13, no. 2, pp. 204–216, 2012.

[27] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American Statistical Association*, vol. 47, no. 260, pp. 583–621, 1952.

[28] W. J. Conover, M. E. Johnson, and M. M. Johnson, "A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data," *Technometrics*, vol. 23, no. 4, pp. 351–361, 1981.

[29] Y. Shyr and K. Kim, "Weighted flexible compound covariate method for classifying microarray data," in *A Practical Approach to Microarray Data Analysis*, D. Berrar, W. Dubitzky, and M. Granzow, Eds., pp. 186–200, Springer, New York, NY, USA, 2003.

[30] Y. Guo, C. I. Li, F. Ye, and Y. Shyr, "Evaluation of read count based RNAseq analysis methods," *BMC Genomics*, vol. 14, supplement 8, article S2, 2013.

[31] C. Trapnell, A. Roberts, L. Goff et al., "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks," *Nature Protocols*, vol. 7, pp. 562–578, 2012.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 873436, 9 pages
http://dx.doi.org/10.1155/2014/873436

*Research Article*

# A Diverse Stochastic Search Algorithm for Combination Therapeutics

**Mehmet Umut Caglar[1,2] and Ranadip Pal[2]**

[1] *Department of Physics, Texas Tech University, P.O. Box 41051, Lubbock, TX 79409, USA*
[2] *Department of Electrical and Computer Engineering, Texas Tech University, P.O. Box 43102, Lubbock, TX 79409, USA*

Correspondence should be addressed to Ranadip Pal; ranadip.pal@ttu.edu

Received 17 November 2013; Revised 20 January 2014; Accepted 6 February 2014; Published 12 March 2014

Academic Editor: Hua Xu



*Background*. Design of drug combination cocktails to maximize sensitivity for individual patients presents a challenge in terms of minimizing the number of experiments to attain the desired objective. The enormous number of possible drug combinations constrains exhaustive experimentation approaches, and personal variations in genetic diseases restrict the use of prior knowledge in optimization. *Results*. We present a stochastic search algorithm that consisted of a parallel experimentation phase followed by a combination of focused and diversified sequential search. We evaluated our approach on seven synthetic examples; four of them were evaluated twice with different parameters, and two biological examples of bacterial and lung cancer cell inhibition response to combination drugs. The performance of our approach as compared to recently proposed adaptive reference update approach was superior for all the examples considered, achieving an average of 45% reduction in the number of experimental iterations. *Conclusions.* As the results illustrate, the proposed diverse stochastic search algorithm can produce optimized combinations in relatively smaller number of iterative steps. This approach can be combined with available knowledge on the genetic makeup of the patient to design optimal selection of drug cocktails.

## 1. Introduction

Biological networks are complex and stochastic by nature. They are also robust and incorporate redundancy in their functionality. Thus, from the perspective of intervention, targeting individual proteins or pathways may not be sufficient for achieving a desirable outcome. For instance, solid tumors often fail to respond to monotherapy due to redundant pathways being able to carry on proliferation [1, 2]. Thus, combination therapy is often considered where multiple proteins and pathways are targeted to reduce tumor growth and avoid resistance to therapy [3, 4]. The primary concern with this approach is the enormous increase in the possible candidate concentrations that needs to be experimentally tested. One possible solution can be detailed modeling of the cellular system and design of the combination therapy based on analytical optimization and simulation. However, the kind of detailed model that captures the synergy or antagonism of drugs at different levels can require enormous amount of experimentation to infer the model parameters. Furthermore, this kind of approach may only work for molecularly targeted drugs where the specific drug targets are known, but modeling chemotherapeutic drug synergies can often be difficult. Existing approaches to drug sensitivity prediction based on genomic signatures often suffer from low accuracy [5–7]. Thus for generation of optimal drug cocktails, systematic empirical approaches tested on in-vitro patient tumor cultures are often considered. Some existing approaches include (i) systematic screening of combinations [8–10] which requires numerous test combinations, (ii) medicinal algorithmic combinatorial screen (MACS) based on laboratory drug screen for multiple drug combinations guided by sequential search using a fitness function [11], and (iii) deterministic and stochastic optimized search algorithms [12–15]. The systematic search approach should be focused on locating the global maximum instead of getting stuck in a local maximum. Furthermore, the optimization algorithm needs to be effective in search spaces, without existing prior

knowledge, and easily adaptable to higher dimensional systems. Since the knowledge of the drug sensitivity distribution is unknown, the algorithm should be effective over a number of unrelated search spaces. A common problem related to the stochastic search algorithms in the literature is the normalization issue mentioned in [15]. Some of the search algorithms like Gur Game require proper normalization of the search space without having any prior information about it [13]. In order to overcome this problem, proposed algorithms should be adaptive to nonnormalized search spaces.

In this paper, we propose a diversified stochastic search algorithm (termed DSS) that does not require prior normalization of the search space and can find optimum drug concentrations efficiently. At this point, it is important to emphasize that the primary objective of the algorithm is to minimize the number of experimental steps and not the CPU time. The critical problem is the cost of the experiments that are necessary to find the most efficient combination, since the cost of biological experiments is significantly higher than the cost of computation in terms of time and money. As the number of biological experiments necessary to find the global maximum is equal to the number of steps, the objective is to minimize the experimental steps in order to reduce the cost of the overall process. With our proposed algorithm, we significantly decrease the number of steps necessary to find the maximum. The proposed iterative algorithm is based on estimating the sensitivity surface and incorporating the response of previous experimental iterations. By generating an estimate of the sensitivity distribution based on currently available response data, we are able to make larger moves in the search space as compared to smaller steps for gradient based approaches. The proposed algorithm is composed of two parts: (a) the first part consists of generating a rudimentary knowledge of the search space, (b) and in the second part, we utilize the crude knowledge of the sensitivity distribution to run a focused search to locate the exact maximums or run a diverse search to gather further information on the sensitivity distribution.

For comparing the accuracy of our proposed approach, we compare the performance of our algorithm to the recently proposed *adaptive reference update* (ARU) [15] algorithm which has been shown to outperform earlier stochastic search approaches for drug cocktail generation [13, 14]. We consider seven diverse example functions that represent possible drug interaction surfaces and also test our algorithm on two experimentally generated synergistic drug combinations. Even though our algorithm is not constrained to discretized drug levels, we have considered discretized drug concentration levels for our examples to be able to compare our results with previous studies. The results illustrate that the proposed algorithm is more efficient than the ARU algorithm for all the considered drug response surfaces. We also present a theoretical analysis of the proposed search algorithm to explain the algorithm performance.

The paper is organized as follows: the *Results* section contains the detailed performance analysis for the 9 examples, and 4 of them were analyzed for 2 different parameter sets; the *Discussion* section includes the theoretical analysis of the algorithm along with conclusions; and the search algorithm along with the surface estimation algorithm are presented in the *Methods* section.

## 2. Results

In this section, we present the performance of our proposed algorithm for nine different examples. The numbers of drugs considered in the examples are 2, 3, 4, and 5 with 21, 11, 11, and 11 discretized concentration levels, respectively, resulting in search space sizes of $21^2$, $11^3$, $11^4$, and $11^5$ for the synthetic examples and search grid sizes of $9^2$ and $10^2$ for the experimentally generated examples with two drugs and number of discretization levels of 9 and 10. As mentioned earlier, our results are compared with the latest stochastic search algorithm for drug cocktail generation (termed ARU algorithm) [15] which was shown to outperform earlier approaches [13, 14]. Similar to comparisons in [15], two parameters are primarily considered (a) *Cost*: average number of steps required to reach within 95% of the maximum sensitivity and (b) *Success Rate*: percentage of times that the search algorithm reaches 95% of the maximum sensitivity within a fixed number of steps. The details of the example functions, search parameters, and performance of our approach and ARU approach are shown in supplementary Tables 1–9 (see Supplementary Material available online at http://dx.doi.org/10.1155/2014/873436). Each table contains the problem dimensions, intervals, grid points, algorithmic parameters, and the performance comparison in terms of success percentage and average number of iterations (termed score) for our proposed approach and ARU. Two of the presented examples are based on experimentally generated data. (i) Supplementary Table 8 reports the results for bacterial (*S. aureus*) inhibition response for the drugs Trimethoprim and Sulfamethoxazole that has a synergistic effect as shown in [10]. The data surface is shown in supplementary Table 10. (ii) Supplementary Table 9 considers lung cancer inhibition response using the drugs Pentamidine and Chlorpromazine [16]. Both of these compounds have moderate antiproliferative activities on their own in-vitro in A549 lung carcinoma cells. But neither pentamidine (an antiprotozoal agent) nor chlorpromazine (an antipsychotic agent) is used clinically as a cancer drug. On the other hand, because of the synergy between them, they can prevent the growth of A549 lung carcinoma cells; in addition to that, in proper concentrations, combination is more effective than some of the commonly used cancer drugs. The data surface is shown in supplementary Table 11. The performance of our approach as compared to ARU algorithm for the nine examples is summarized in Table 1. The results indicate that we achieve 100% success rate for all nine examples (thirteen different evaluations), whereas ARU has slightly lower success rate in 2 of these examples. The primary benefit of our approach is the lower average number of iterations to reach within 95% of the maximum sensitivity. For all the examples considered, we require significantly lower average number of iterations to reach within 95% of the maximum. Note that the standard deviation of the number of iterations required to reach within 95% of the maximum is relatively small as compared to the difference

Table 1: Summary of the results for seven synthetic and two experiment based examples. The table illustrates that the cost for our proposed algorithm is significantly lower than the ARU algorithm [15] which has been shown to be efficient than other existing algorithms. Furthermore, the success rate for our algorithm is also better than the ARU algorithm.

| | Number of points with $\geq 0.95 \times \text{Max}_{\text{efficacy}}$ | Search grid size | ARU [15] cost | ARU [15] Success Rate | DSS cost | DSS STD | DSS worst case | DSS Success Rate | Initial LHC points |
|---|---|---|---|---|---|---|---|---|---|
| Synthetic examples | | | | | | | | | |
| 2 DeJong | 2 | $21^2$ | 46.2 | 99% | 16.0 | 7.99 | 48 | 100% | 5 |
| 3a | 4 | $11^3$ | 74 | 100% | 24.7 | 11.73 | 72 | 100% | 10 |
| 3b | 1 | $11^3$ | 79.4 | 100% | 52.9 | 32.20 | 149 | 100% | 10 |
| 4a | 1 | $11^4$ | 136.8 | 100% | 65.3 | 14.11 | 106 | 100% | 40 |
| 4a | 1 | $11^4$ | 136.8 | 100% | 50.7 | 21.81 | 159 | 100% | 10 |
| 4b | 12 | $11^4$ | 91.6 | 100% | 52.7 | 8.80 | 85 | 100% | 40 |
| 4b | 12 | $11^4$ | 91.6 | 100% | 28.3 | 9.17 | 57 | 100% | 10 |
| 5a | 4 | $11^5$ | 80.6 | 100% | 79.3 | 23.25 | 157 | 100% | 40 |
| 5a | 4 | $11^5$ | 80.6 | 100% | 61.8 | 27.58 | 176 | 100% | 10 |
| 5b | 8 | $11^5$ | 216.8 | 100% | 159.5 | 90.51 | 402 | 100% | 40 |
| 5b | 8 | $11^5$ | 216.8 | 100% | 194.2 | 150.15 | 647 | 100% | 10 |
| Experiment based examples | | | | | | | | | |
| Bacterial inhibition [10] | 34 | $9^2$ | 4.8 | 100% | 1.85 | 0.78 | 3 | 100% | 3 |
| Lung cancer [16] | 7 | $10^2$ | 12.4 | 98% | 5.97 | 4.74 | 23 | 100% | 3 |

in average iterations between ARU and DSS. For instance, the first example in 2 dimensions requires an average of 16 iterations for our proposed approach as compared to 46.2 iterations for ARU approach. The standard deviation ($\sigma$) in 100 runs of DSS algorithm is 7.99 and thus the difference in the mean runs between DSS and ARU is more than $3.7\sigma$. The ARU algorithm has earlier been shown to outperform other existing algorithms such as Gurgame and its variants. Please refer to Tables 1 and 2 of [15] for the detailed comparison results of ARU with Gurgame. This strongly illustrates that the proposed algorithm is able to generate high sensitivity drug combinations in lower number of average iterations than existing approaches.

To further illustrate the significance of the proposed approach, let us consider one of the experimental example results. The experimental data on lung cancer contains the sensitivity for $102 = 100$ drug concentration combinations where each drug is assumed to have one of 10 discrete concentrations. This data has been utilized to study the efficacy of the proposed algorithm. For instance, an exhaustive search approach will experimentally test the sensitivity of each of these 100 concentrations and select the one with the highest sensitivity and thus it will require 100 experimental steps. On the other hand, the stochastic search algorithms such as ARU and proposed DSS will start with random drug concentration combinations and try to sequentially select drug concentrations that will provide an improved knowledge of the sensitivity surface over these two drugs. As Table 1 shows, ARU will require an average of 12.4 sequential steps to reach a drug combination that has sensitivity within 95% of the maximum sensitivity, whereas the proposed DSS will require an average of 5.97 sequential steps to reach within 95% of the maximum sensitivity. Thus, DSS will reach within 95% of the maximum sensitivity on an average of 5.97 experimental steps, whereas ARU will require 12.4 experimental steps and exhaustive search will require 100 experimental steps. Note that since ARU and DSS are stochastic approaches, the number of sequential steps required can vary with each experimental run and the numbers 12.4 and 5.97 represent the mean of multiple experimental runs. The experimental data has been used here to provide the sensitivities for specific drug concentrations requested by the algorithms.

For analyzing the behavior of our algorithm during the iteration process, we analyzed the minimal distance of the optimal point(s) from the DSS selected points. Let us consider $n$ drugs and 0 to $T$ discretization levels for each drug. Figure 1 represents the minimum $L_1$ distance of the points selected for experimentation to any of the optimal point(s) for the synthetic Example 6 (the simulation details are included in supplementary Table 6) with two different parameter sets (the number of initial points is 40 and 10, resp.) for 100 repeated experiments. Note that $n = 4$ and $T = 10$ for the example and thus the maximum possible $L_1$ distance is 40. The number of optimal points for this example is 1. The red vertical line represents the value of $m$ which is 40 and 10, respectively, for two different solutions of this example. The black vertical line represents the average number of iterations required to reach an optimal point for the specific response function. The cyan vertical dotted line represents the worst situation out of 100 different runs. The average minimum distance of the experimental points to the optimal point is shown in green in Figure 2. The solid blue line represents the analytically calculated expected minimum $L_1$ distance. The theoretical analysis of the minimum distance is included in Section 3. Note that there is a change in the shape of the blue curve after the end of Step 1 (iteration 40 for Figure 2(a) and iteration 10 for Figure 2(b)).

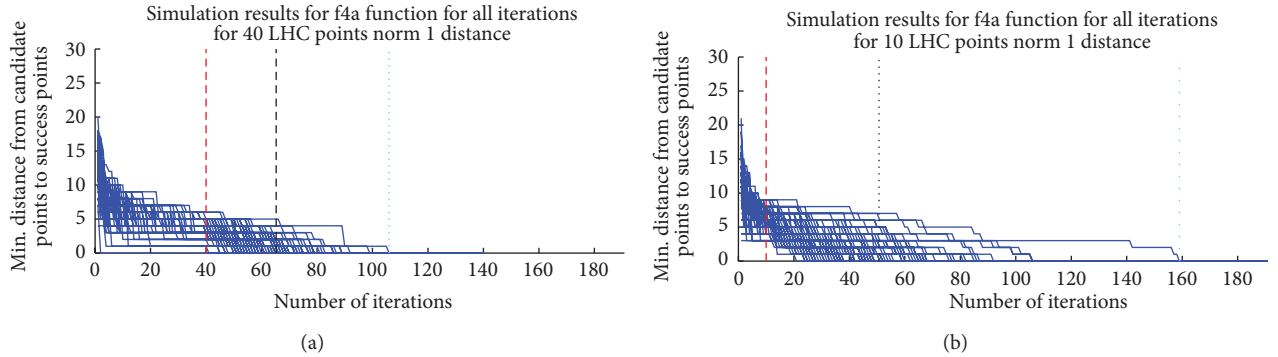



Figure 1: Minimum distance to optimal points for function f4a. We simulate the f4a function (see Supplementary Table 6) 100 times with two different LHC numbers. First figure represents the simulation with LHC equal to 40 and second figure represents the simulation LHC equal to 10. The analyzed function has 1 optimal point. The blue line represents the minimum norm 1 distance between the optimal points and DSS selected points. The red vertical line represents the end of Step 1, that is, Latin Hypercube Numbers, which is equal to 40th iteration in first simulation set and 10th iteration in the second simulation set; and the black vertical dotted line represents the average value of iterations (cost of proposed algorithm) required to find one of the points with $\geq 0.95 \times \text{Max}_{\text{efficacy}}$ (equal to 79.25). The cyan vertical dotted line represents the worst situation out of 100 different runs.

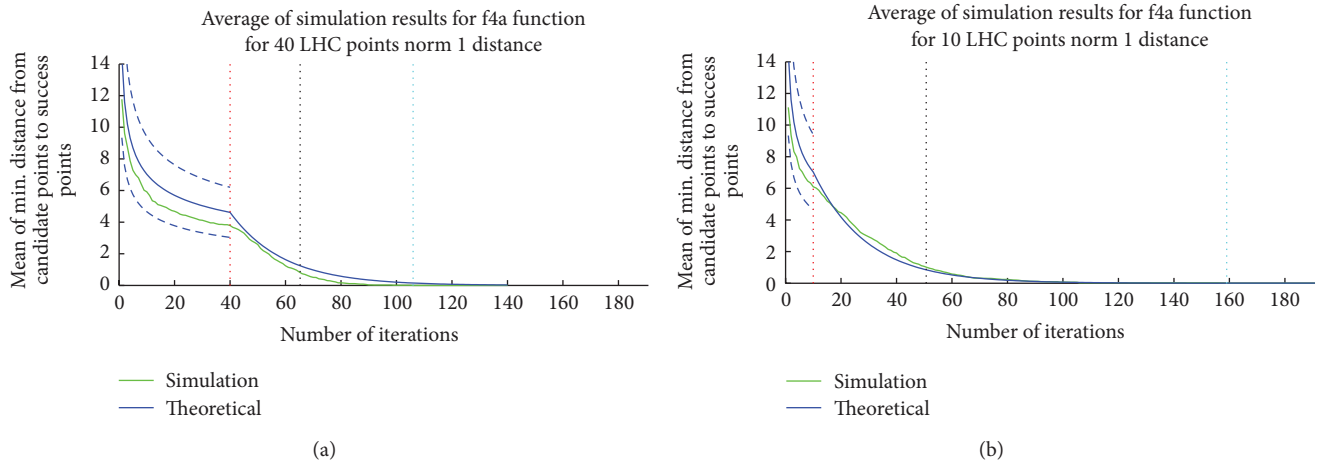



Figure 2: Simulation average and theoretical expected minimum distance to optimal points for function f4a (Supplementary Table 6). The blue line represents the theoretical values for $L_1$ distance and the dashed blue lines represent the error margins ($\mu \pm \sigma$) for the analytically calculated values for Step 1 of the iteration. First figure represents the simulation with LHC equal to 40 and second figure represents the simulation LHC equal to 10. The red vertical line represents the end of Step 1, that is, Latin Hypercube Numbers, which is equal to 40th iteration; and the black vertical line represents the average value of iterations (cost of proposed algorithm) required to find one of the points with $\geq 0.95 \times \text{Max}_{\text{efficacy}}$. The perpendicular cyan line represents the worst situation out of 100 different runs.

The dotted blue curve denotes the analytically calculated $\mu \pm \sigma$ where $\mu$ and $\sigma$ denote the mean and standard deviation for the minimum distance. Figures 1 and 2 illustrate that the minimum distance of the selected points to the optimal points decreases with successive iteration and closely matches the analytical predictions.

## 3. Discussion

In this section, we provide a generalized analysis of the proposed search algorithm followed by conclusions and future research directions.

*3.1. Theoretical Analysis.* In this subsection, we will attempt to theoretically analyze the distance of the point with the global sensitivity maximum from the points that are tested by the proposed algorithm. We will consider that each drug is discretized from 0 to $T$ levels and that we are considering $n$ drugs. Thus, any drug cocktail can be represented by a $n$ length vector $V = \{V(1), V(2), \ldots, V(n)\}$, where $V(i) \in \{0, 1, 2, \ldots, T\}$ for $i \in \{0, 1, \ldots, n\}$. Thus, the search space of drug cocktails (denoted by $\Omega$) is of size $(T + 1)^n$ and represents points in an $n$-dimensional hypercube of length $T$. Let $V_{\text{max}}$ denote the drug cocktail with the maximum sensitivity among the $(T + 1)^n$ points. The mapping from the drug cocktail to sensitivity will be denoted by the function $f : \Omega \rightarrow [0\ \ 1]$; that is, the maximum sensitivity will be given by $f(V_{\text{max}})$. We will assume that if the distance of the test point ($V$) from the point with the global maximum ($V_{\text{max}}$) is small, the sensitivity will be close to the global maximum;

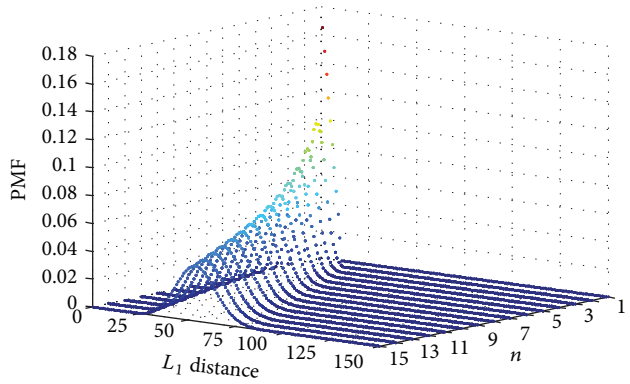



Figure 3: Distribution of random variable $R_1$ (denoting $L_1$ distance from optimal point) for $T = 10$ and different values of $n$.

that is, a small $|V_{\max} - V|$ will imply a small $|f(V_{\max}) - f(V)|$. We will primarily analyze the $L_1$ norm of $|V_{\max} - V|$.

Note that $|V_{\max} - V|_1 = \sum_{i=1}^{n} |V_{\max}(i) - V(i)|$. The first $m$ points in our algorithm are chosen randomly in the search space and thus we will consider that $V(i)$ has a uniform distribution between 0 and $T$. $V_{\max}$ can also be situated in any portion of the search space and thus we will consider $V_{\max}$ to have a uniform distribution between 0 and $T$. Thus, the probability mass function of the random variable $Z = V(i) - V_{\max}(i)$ will be given by

$$f_Z(z) = \begin{cases} \dfrac{T + 1 - |z|}{(T+1)^2} & z = \{-T, -T+1, \ldots, T-1, T\} \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Subsequently, the PMF of the random variable $W = |Z|$ will be given by

$$f_W(w) = \begin{cases} \dfrac{2(T+1-w)}{(T+1)^2} & w = \{0, \ldots, T-1, T\} \\ \dfrac{1}{T+1} & w = 0 \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

The random variable $R_1$ denoting the $L_1$ norm $|V_{\max} - V|_1 = \sum_{i=1}^{n} |V_{\max}(i) - V(i)|$ will be a sum of $n$ random variables with PMF given by (2). The distribution for the sum of any two random variables consists of the convolution of the individual distributions of the random variables. Thus, the probability distribution of $R_1$ can be calculated by convolving $f_W$ $n$ times. The distribution of $R_1$ for $T = 10$ and $n = \{1, \ldots, 15\}$ is shown in Figure 3.

At the beginning of our algorithm, we are selecting $m$ points in random. Thus, the nearest neighbor distance from the optimal point will be given by the random variable $R_2$ that denotes the minimum of $m$ random variables $X_1, X_2, \ldots, X_m$ selected independently based on the distribution of $R_1$. Thus,

Table 2: Expectation and variance of the minimum distances from the optimal point for various values of $n$, $T$, and $m$.

| $n$ | $T$ | $m$ | Mean | Variance |
|---|---|---|---|---|
| 5 | 5 | 20 | 4.15 | 1.77 |
| 5 | 5 | 40 | 3.42 | 1.29 |
| 5 | 5 | 60 | 3.04 | 1.09 |
| 10 | 5 | 20 | 11.35 | 4.29 |
| 10 | 5 | 40 | 10.20 | 3.32 |
| 10 | 5 | 60 | 9.59 | 2.90 |
| 15 | 5 | 20 | 19.16 | 6.88 |
| 15 | 5 | 40 | 17.69 | 5.43 |
| 15 | 5 | 60 | 16.92 | 4.79 |
| 5 | 10 | 20 | 8.15 | 5.25 |
| 5 | 10 | 40 | 6.86 | 3.71 |
| 5 | 10 | 60 | 6.21 | 3.06 |
| 10 | 10 | 20 | 21.75 | 13.35 |
| 10 | 10 | 40 | 19.70 | 10.21 |
| 10 | 10 | 60 | 18.62 | 8.83 |
| 15 | 10 | 20 | 36.45 | 21.76 |
| 15 | 10 | 40 | 33.83 | 17.02 |
| 15 | 10 | 60 | 32.45 | 14.94 |

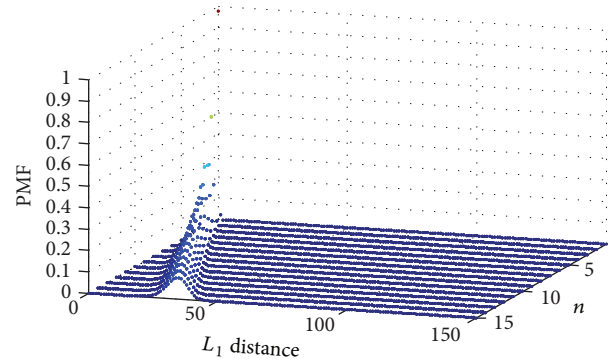



Figure 4: Distribution of random variable $R_2$ (denoting the minimal $L_1$ distance from the optimal point for $m = 40$) for $T = 10$ and different values of $n$.

the cumulative distribution function (CDF) of $R_2$ is given by [17]

$$\begin{aligned} P(R_2 \le x) &= 1 - P(X_1 > x, \ldots, X_m > x) \\ &= 1 - P(X_1 > x) * \cdots * P(X_m > x) \\ &= 1 - \left(1 - \mathrm{CDF}_{R_1}(x)\right)^m. \end{aligned} \quad (3)$$

The PMF of $R_2$ given by $\mathrm{PMF}_{R_2}(x) = \mathrm{CDF}_{R_2}(x) - \mathrm{CDF}_{R_2}(x-1)$ for $i = 1, 2, \ldots nT$ and $\mathrm{PMF}_{R_2}(0) = \mathrm{CDF}_{R_2}(0)$ is shown in Figure 4.

For example, the expected minimum distance from the optimal point for $m = 40$, $T = 10$ is 6.86 for $n = 5$. The mean $\mu(n, T, m)$ and variance $\sigma(n, T, m)^2$ of the minimum distance from the optimal point for different values of $n$, $T$, and $m$ are shown in Table 2. Note that if there are $k$ optimal points in diverse locations, the mean $\mu_k(n, T, m)$ and variance

$\sigma_k(n,T,m)^2$ of the minimum distance of the selected points from any of the optimal points are given by $\mu_k(n,T,m) = \mu(n,T,k*m)$ and $\sigma_k(n,T,m)^2 = \sigma(n,T,k*m)^2$. This is because when there are $k$ optimal points, the minimum distance will consist of the minimum of $m \times k$ distances ($m$ distances from each optimal point). If there are multiple optimal points in one hill with small distances between each other, they will be considered as one single optimal point for the minimum distance analysis.

As Table 2 shows, the $L_1$ distance will increase with increasing $n$ and $T$ and following Step 1 of the algorithm, our point with highest experimental sensitivity may not be close to the optimal point but rather may belong to another hill with a local optima. However, based on the nearest neighbor $L_1$ distances, we would expect to have at least one point close to the optimal point in the top $k$ optimal points. Thus, if we keep selecting $\rho$ points for further experimentation from around the top $k$ experimental points sequentially, we expect that on an average $\rho_1 = \rho/k$ points will be selected around the optimal point.

Consider that the $L_1$ distance from the optimal point was given by the random variable $R_2$ and if a point is selected randomly between the experimental point and the randomly selected point, the subsequent nearest neighbor distance from the optimal point will be given by the random variable $R_3 = R_2 * G_1$, where $G_1$ is a uniform random variable between 0 and 1. The distance in each dimension will be reduced by a number selected based on a uniform random variable, and consequently, we will approximate the $L_1$ distance (sum of $n$ such distances) to be reduced by a number selected based on a uniform random variable. Thus, after $\rho_1$ points have been selected sequentially around the optimal point, the distance to the optimal point will be given by the random variable $R_{\rho_1} = R_2 * G_1 * G_2 * \cdots * G_{\rho_1}$. The probability distribution function of the multiplication of $\rho_1$ random variables with uniform distribution between $[0,1]$ is given by [18]

$$f_{G_1 * \cdots * G_{\rho_1}}(x) = \frac{(\ln(1/x))^{n-1}}{(n-1)!}. \quad (4)$$

Thus, if the expected distance from the optimal point after the initial selection of $m$ points is $D$ and we select $\rho_1$ points sequentially between the optimal point and its current nearest neighbor, the expected nearest neighbor distance from the optimal point will be $D/2^{\rho_1}$.

As an example, if $n = 10$, $T = 10$, and $m = 40$, we have $D = 19.7$ from Table 2. The expected $L_1$ distance from the optimal point at the end of $40 + 20 = 60$ iterative steps will be $19.7/2^6 = 0.3078$ assuming a single hill and a 0.3 probability for the focused search (path $a$ of Step 2 of the algorithm). Based on the focused search probability, at the end of 60 iterations, we expect to have $(60 - 40) * 0.3 = 6$ points selected around the optimal point.

*3.2. Conclusions.* In this paper, we proposed a diverse stochastic search (DSS) algorithm that consisted of a parallel and sequential phase that outperformed existing efficient algorithms for drug cocktail generations on nine different examples (thirteen different evaluations). Our results show that the DSS algorithm is more efficient than the previous algorithms in terms of decreasing the number of experiments required to generate the optimum drug combination (i.e., cost of the algorithm) which in turn reduces the total cost of the drug combination search process in terms of time and money. Note that the primary costs in each sequential biological experimental step are in personnel effort to prepare the drug combination and the time involved to generate the combination drug response and thus the goal is to reduce the number of sequential steps. One of the limitations of the current approach is the computational complexity when the number of drugs ($n$) is large. The proposed method is suitable for selection of optimal drug concentrations when the number of candidate drugs has been reduced from hundreds to around ten. A number of approaches can be applied to achieve the selection of candidate drugs. For instance, an application of a drug screen to measure cell viability of the tumor culture can be utilized to narrow down the drugs to be included in a combination drug cocktail [19, 20]. Note that algorithms [19, 20] are convenient for selecting the small set of drugs for combination therapy but not for deciding on the optimal drug concentrations of the selected drugs. For possible clinical application, available genetic information can be utilized to narrow down the possible drugs to be tested and the proposed DSS algorithm can be applied to tumor cell cultures to generate the optimal concentrations of the drugs. In this paper, we also presented a theoretical analysis of the search based on minimum distance between the optimal point and the DSS selected points. Future research directions will consider incorporating the cost of drug application in the optimization process and the effect of data extraction noise on the search algorithm. The cost will be a measure of toxicity or side effects of the drug combination. One approach to incorporate the cost can consist of changing the sensitivity surface by negating the cost from the sensitivity. The cost can be simplified to be proportional to the linear addition of the individual drug concentrations. Another approach for incorporating the cost entails restricting the search space by limiting the search to areas that have cost lower than a toxicity threshold.

## 4. Methods

In this section, we present the search algorithm along with the surface mapping algorithm based on currently available information. Finally, we discuss the reasoning behind the selection of the algorithm related parameters.

*4.1. The Search Algorithm.* The primary objective of the search algorithm is to locate the global maximum in minimum number of iteration steps. Numerous approaches can be considered for this purpose and our proposed method is based on a combination of stochastic and deterministic approaches. We expect that the efficiency in terms of average number of iteration steps can be increased by large jumps over the search space rather than using traditional step-by-step gradient descent approach. Our algorithm consists of two parts: an initial parallel part and a subsequent iterative

segment. The objective of the initial part is to generate a rudimentary idea of the search space. The objective of the iteration part is twofold: (a) it tries to find the exact maximum using the currently available knowledge and (b) it searches the space further to add new knowledge, that is, it attempts to find new hills that the previous iterations could not locate.

Step-by-step schema of the search algorithm is described as follows.

*Step 1.* Generation of Latin Hypercube Numbers (LHNs)

(1) In this step, $m$ points in the given grid are selected for drug response experiments. We first generate $m$ points in the continuous search space based on LHN generation approach with the criterion of maximizing the minimum distance between these points. This approach assists in distributing the points homogeneously in the search space such that the maximum possible distance between a given target point and the nearest point whose coordinates are represented by LHN will be minimum. Consequently, we map these points to the nearest grid points and term these mapped points as approximate Latin hypercube points. We considered this continuous-discrete grid mapping to compare our results with the previous studies that utilized a grid structure for the search space.

(2) In this step, experiments are conducted to determine the efficiency of the $m$ drug combinations determined by the approximate Latin hypercube points.

*Step 2.* Iterative Segment

(1) Normalize the experimental drug efficacy results to numbers between 0 and 1. Then the $(n-1)$th power of the normalized drug efficacies are considered where $n$ denotes the number of drugs. The power step emphasizes the hills of the distribution and the value $n-1$ is termed as *Power used for the inputs*.

(2) Estimate the drug efficacies of the unknown grid points using the sensitivity surface estimation algorithm. The details of the surface estimation algorithm are explained in subsequent sections. At the end of this procedure, we have estimates for the efficacies of each and every point on the search grid. The grid points are classified into two groups: known points from experimental data and estimated points based on interpolation and extrapolation.

(3) Decide the objective of the iteration step based on a probabilistic approach. For our case, the algorithm follows *path a* to find the exact maximum based on previous knowledge with a 0.3 probability and follows *path b* with a 0.7 probability to explore the search space with a diversified approach.

(4) *Path a* (Focused Search)

(a) The main idea of the focused search is to experimentally search the estimated maximums generated following the surface estimation mapping. The algorithm also tries to avoid focusing on an individual local maximum by exploring geographically apart multiple estimated local maximums. To achieve this purpose, we employ a tracking algorithm to label the local maximums and avoid prolonged emphasis on individual maximum points. The individual steps of the Focused Search part are described as follows.

(b) Sort the grid sensitivities (both experimental and estimated sensitivities) from higher to lower sensitivity values.

(c) Check if the location corresponding to the highest sensitivity is an experimental point or an estimated point. If it is an estimated point, generate the experimental sensitivity for this grid location.

(d) If the highest sensitivity point is an experimental point, check the second highest point. If the second highest point is an estimated point, generate the experimental value for this grid location.

(e) If the second highest point is also an experimental point, generate the gradient from the second highest point based on the mapped surface. If the upward path from the second highest point leads up to the highest point, label both the points as 1, which implies that they belong to the same hill. Otherwise, label the highest point as 1 and label the second highest point as 2, which indicates that they belong to different hills.

(f) Repeat this procedure till an estimated point is located. Meanwhile, keep labeling the experimental points with respect to the hill they belong to and the order of the point on the hill (ex: 3rd highest point in hill 2, etc.).

(g) If a hill's highest $\xi$ points are experimental points, then label the hill as discovered, which indicates that we have enough information on this hill and collecting information on other hills might be more beneficial.

(h) If the search continues till 1% of grid points without finding a suitable candidate for experimentation, halt the search. Locate all the considered points that are inside a sphere of volume 1/500 of the whole search space with center being the highest sensitivity point. Assign a value of "0" for the sensitivities of all points inside this sphere. (Maintain the record of their actual values in another place). Then go to the beginning of Step 2.

(5) *Path b* (Diverse Search)

(a) The aim of the diverse search is to explore the space to locate new possible candidate hills that were not discovered in the previous searches.

(b) Assume that the surface generated by the experimental and estimated points is a probability distribution function (PDF).

(c) Generate points by sampling this distribution. For the generation process, we use the Gibbs sampling algorithm. The number of points generated by the Gibbs algorithm is termed as *Number of points to generate the Gibbs sampling*. Since the points are generated from sampling the PDF, the points are denser around the hills and less dense at locations where the efficacy estimate is close to 0.

(d) Randomly select one of the generated points as the candidate point and generate its sensitivity experimentally.

*4.2. Sensitivity Surface Estimation Algorithm.* The sensitivity surface estimation algorithm used for our approach is established on the $n$ dimensional application of the penalized least square regression analysis based on discrete cosine transform (LS-DTC) proposed by Garcia [21, 22]. The code is generated to compute missing values in data sets. The estimation algorithm contains a parameter $s$, termed as *smoothing parameter* that determines the smoothness of the output. For our case, the smoothness parameter is adjusted to a small value so that the result of surface estimation goes through the actual experimental points. The core of the algorithm is based on minimizing the equation $F(y) = wRSS + s * P(y)$, where wRSS corresponds to weighted residual sum of squares and $P$ is the penalty function related to the smoothness of the output. wRSS can be written explicitly as $\|W^{1/2} \cdot (\widehat{y} - y)\|^2$, where $y$ represents the actual data with missing values and $\widehat{y}$ provides the estimate of the data without any missing values. $W$ is a diagonal matrix, whose entries represent the reliability of the points and can take values between 0 and 1. For our case, the missing values represent unknown points and are assigned a value of 0 and the experimental points are reliable points which are assigned a value of 1. The solution to $\widehat{y}$ that minimizes $F(y)$ can be generated based on an iterative process starting from an arbitrary initial point $\widehat{y}_0$.

*4.3. Choice of Parameters.* The implementation of the proposed algorithm includes several parameters that can affect the performance of the search process. In this subsection, we present the guiding principles behind the selection of the parameters based on the dimensionality and the total number of grid points in the search space.

*4.3.1. Latin Hypercube Numbers (LHNs).* Denote the number of points that will be tested in Step 1 of the algorithm. These points are supposed to provide an initial estimate of the search space. Based on simulations and theoretical analysis, we observe that increasing the number of LHNs provides limited benefit in terms of reaching the maximum sensitivity combination after a certain point. On the other hand, keeping this number too low will cause the program to start the second step with limited knowledge and to search low sensitivity locations. Thus, there is an optimum number of Latin hypercube numbers to maximize the benefit of the algorithm. Although this optimum number depends on the search space; our simulations for 4 surfaces with two different LHNs (10 and 40) illustrate that the proposed algorithm provides better results than ARU algorithm for a fairly large interval of LHNs.

*4.3.2. Latin Hypercube Iterations.* The Latin hypercube numbers are distributed homogeneously through an iterative algorithm. The iterations maximize the minimum distance between the points. It is desirable to have a higher number of iterations, but after a point, the benefits of increasing the iterations become negligible. For our simulations, we selected a threshold point following which the increase in the maximum minimum distance is negligible.

*4.3.3. Number of Iterations to Generate the Sensitivity Surface Estimate.* This parameter is related to sensitivity surface estimation algorithm and describes the number of iterations used to find a smooth surface passing through the given points in high dimensional space. A higher value for this parameter will provide a smoothed surface (that still passes through the experimental points) but will carry a high computational time cost. Furthermore, the benefits of increasing the iterations become negligible after a threshold and the output surface becomes more stable. For our examples, we have fixed this number to 100.

*4.3.4. Probability of Focused Search.* Denote the probability that the search algorithm follows *path a*. *Path a* attempts to discover the exact local maximum of a hill, and *path b* attempts to learn new hills. For all our simulations, this parameter has been assigned a value of 0.3.

*4.3.5. Power Used by Inputs.* This parameter attempts to emphasize the hills. After normalizing the experimental points, we take the $(n - 1)$th power of the values so that the high peaks are emphasized as compared to dips or grid points with average values in the estimated surface. Thus, the probability of point selection around hills is increased during the Gibbs sampling process.

*4.3.6. Number of Points Generated by the Gibbs Sampling.* This parameter describes the number of points generated by the Gibbs sampling in *path b* of Step 2 of the algorithm. More points provide a better representation of the estimated surface. After a level, the number of points is sufficient to represent the probability distribution and the benefits of increasing the iterations become negligible. We achieve better sampling by increasing this parameter. This parameter is required to be large for problems in higher dimensions or problems containing a huge number of grid points. For our examples, if the number of grid points is below 7500, this parameter has been assigned a value equal to twice the number of grid points. Otherwise, we have fixed the number of the Gibbs sampling points to 15,000.

*4.3.7. Clustering Related Parameters.* The clustering concept is introduced to avoid the search being stuck in one dominant hill.

*4.3.8. Cluster Threshold $\xi$.* This denotes the maximum number of experimental points in an individual hill. Further exploration of the hill is paused once this value is reached. For our examples, if the number of drugs (dimensions) $n$ is less than 5, $\xi$ is assigned a value of $2n - 1$. Otherwise, the parameter is fixed at 7.

*4.3.9. Cluster Break.* This parameter denotes the maximum number of high efficacy point estimates in a single hill. If this condition is reached, we assign a value of 0 sensitivity for points around the known top of the hill. This parameter is considered to be around 1% of the total grid points.

*4.3.10. Cluster Distance.* This parameter represents the radius of the sphere around the hill top for which any grid point within the sphere is assigned a value of 0. The Cluster Distance is selected such that the volume of the sphere is 0.2% of the total volume. The parameter considers that the algorithm has no knowledge of the hills that are narrower than the 0.2% of the total search space.

## Conflict of Interests

The authors declare that they have no competing interests.

## Authors' Contribution

Mehmet Umut Caglar and Ranadip Pal conceived and designed the algorithms, Mehmet Umut Caglar performed the simulation experiments, Mehmet Umut Caglar and Ranadip Pal analyzed the results, and Mehmet Umut Caglar and Ranadip Pal wrote the paper. All authors read and approved the final paper.

## Acknowledgment

This work was supported in part by NSF Grant CCF 0953366.

## References

[1] C. L. Corless, A. Schroeder, D. Griffith et al., "PDGFRA mutations in gastrointestinal stromal tumors: frequency, spectrum and in vitro sensitivity to imatinib," *Journal of Clinical Oncology*, vol. 23, no. 23, pp. 5357–5364, 2005.

[2] J. A. Engelman, K. Zejnullahu, T. Mitsudomi et al., "MET amplification leads to gefitinib resistance in lung cancer by activating ERBB3 signaling," *Science*, vol. 316, no. 5827, pp. 1039–1043, 2007.

[3] A. Dubrovska, J. Elliott, R. J. Salamone et al., "Combination therapy targeting both tumor-initiating and differentiated cell populations in prostate carcinoma," *Clinical Cancer Research*, vol. 16, no. 23, pp. 5692–5702, 2010.

[4] B. Al-Lazikani, U. Banerji, and P. Workman, "Combinatorial drug therapy for cancer in the post-genomic era," *Nature Biotechnology*, vol. 30, no. 7, pp. 679–692, 2012.

[5] J. Barretina, G. Caponigro, N. Stransky et al., "The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity," *Nature*, vol. 483, pp. 603–607, 2012.

[6] M. J. Garnett, E. J. Edelman, S. J. Heidorn et al., "Systematic identification of genomic markers of drug sensitivity in cancer cells," *Nature*, vol. 483, pp. 570–575, 2012.

[7] M. L. Sos, K. Michel, T. Zander et al., "Predicting drug susceptibility of non-small cell lung cancers based on genetic lesions," *Journal of Clinical Investigation*, vol. 119, no. 6, pp. 1727–1740, 2009.

[8] A. A. Borisy, P. J. Elliott, N. W. Hurst et al., "Systematic discovery of multicomponent therapeutics," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 13, pp. 7977–7982, 2003.

[9] M. Wadman, "The right combination," *Nature*, vol. 439, pp. 390–391, 2006.

[10] G. R. Zimmermann, J. Lehár, and C. T. Keith, "Multi-target therapeutics: when the whole is greater than the sum of the parts," *Drug Discovery Today*, vol. 12, no. 1-2, pp. 34–42, 2007.

[11] R. G. Zinner, B. L. Barrett, E. Popova et al., "Algorithmic guided screening of drug combinations of arbitrary size for activity against cancer cells," *Molecular Cancer Therapeutics*, vol. 8, no. 3, pp. 521–532, 2009.

[12] D. Calzolari, S. Bruschi, L. Coquin et al., "Search algorithms as a framework for the optimization of drug combinations," *PLoS Computational Biology*, vol. 4, no. 12, Article ID e1000249, 2008.

[13] K. W. Pak, F. Yu, A. Shahangian, G. Cheng, R. Sun, and C.-M. Ho, "Closed-loop control of cellular functions using combinatory drugs guided by a stochastic search algorithm," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 13, pp. 5105–5110, 2008.

[14] B.-J. Yoon, "Enhanced stochastic optimization algorithm for finding effective multi-target therapeutics," *BMC Bioinformatics*, vol. 12, supplement 1, article S18, 2011.

[15] M. Kim and B. J. Yoon, "Adaptive reference update (ARU) algorithm. A stochastic search algorithm for efficient optimization of multi-drug cocktails," *BMC Genomics*, vol. 13, supplement 6, article S12, 2012.

[16] A. A. Borisy, P. J. Elliott, N. W. Hurst et al., "Systematic discovery of multicomponent therapeutics," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 13, pp. 7977–7982, 2003.

[17] D. P. Bertsekas and J. N. Tsitsikilis, *Introduction to Probability*, Athena Scientific, 2nd edition, 2008.

[18] C. P. Dettmann and O. Georgiou, "Product of n independent uniform random variables," *Statistics and Probability Letters*, vol. 79, no. 24, pp. 2501–2503, 2009.

[19] N. Berlow, L. E. Davis, E. L. Cantor, B. Seguin, C. Keller, and R. Pal, "A new approach for prediction of tumor sensitivity to targeted drugs based on functional data," *BMC Bioinformatics*, vol. 14, article 239, 2013.

[20] R. Pal and N. Berlow, "A kinase inhibition map approach for tumor sensitivity prediction and combination therapy design for targeted drugs," in *Proceedings of the Pacific Symposium on Biocomputing*, vol. 17, pp. 351–362, World Scientific, 2012.

[21] D. Garcia, "Robust smoothing of gridded data in one and higher dimensions with missing values," *Computational Statistics and Data Analysis*, vol. 54, no. 4, pp. 1167–1178, 2010.

[22] G. Wang, D. Garcia, Y. Liu, R. de Jeu, and A. Johannes Dolman, "A three-dimensional gap filling method for large geophysical datasets: application to global satellite soil moisture observations," *Environmental Modelling & Software*, vol. 30, pp. 139–142, 2012.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 240403, 6 pages
http://dx.doi.org/10.1155/2014/240403

*Research Article*

# Evaluating Word Representation Features in Biomedical Named Entity Recognition Tasks

**Buzhou Tang,[1,2] Hongxin Cao,[3] Xiaolong Wang,[1] Qingcai Chen,[1] and Hua Xu[2]**

[1] *Department of Computer Science, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, Guangdong 518055, China*
[2] *School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA*
[3] *Department of Medical Informatics, Second Military Medical University, Shanghai 200433, China*

Correspondence should be addressed to Hua Xu; hua.xu@uth.tmc.edu

Received 23 November 2013; Revised 25 January 2014; Accepted 3 February 2014; Published 6 March 2014

Academic Editor: Bing Zhang



Biomedical Named Entity Recognition (BNER), which extracts important entities such as genes and proteins, is a crucial step of natural language processing in the biomedical domain. Various machine learning-based approaches have been applied to BNER tasks and showed good performance. In this paper, we systematically investigated three different types of word representation (WR) features for BNER, including clustering-based representation, distributional representation, and word embeddings. We selected one algorithm from each of the three types of WR features and applied them to the JNLPBA and BioCreAtIvE II BNER tasks. Our results showed that all the three WR algorithms were beneficial to machine learning-based BNER systems. Moreover, combining these different types of WR features further improved BNER performance, indicating that they are complementary to each other. By combining all the three types of WR features, the improvements in $F$-measure on the BioCreAtIvE II GM and JNLPBA corpora were 3.75% and 1.39%, respectively, when compared with the systems using baseline features. To the best of our knowledge, this is the first study to systematically evaluate the effect of three different types of WR features for BNER tasks.

## 1. Introduction

Biomedical Named Entity Recognition (BNER), which extracts important biomedical concepts such as genes and proteins, is a crucial step of natural language processing (NLP) in the biomedical domain. Because of the complexity of biomedical nomenclature, BNER has been a challenging task. First, the same biomedical named entities can be expressed in various forms. For example, gene names often contain alphabets, digits, hyphens, and other characters, thus having many variants (e.g., "HIV-1 enhancer" versus "HIV 1 enhancer"). Moreover, many abbreviations (e.g., "IL2" for "Interleukin 2") have been used for biomedical named entities. Sometimes, the same entity can have very different aliases (e.g., "PTEN" and "MMAC1" refer to the same gene) [1]. Another challenge of BNER is the ambiguity problem. The same word or phrase can refer to more than one type of entities or does not refer to an entity depending on context (e.g., "TNF alpha" can refer to a protein or DNA). All these phenomena make the named entity recognition (NER) task in the biomedical domain more difficult than that in open domains such as newswire.

Considerable efforts have been devoted to BNER research, including some shared-task challenges, such as JNLPBA (Joint Workshop on Natural Language Processing in Biomedicine and its Applications) in 2004 [2] and BioCreAtIvE (Critical Assessment for Information Extraction in Biology Challenge) II GM (gene mention) in 2007 [3]. Different methods have been developed for BNER, mainly falling into three categories: (1) dictionary-based methods [4]; (2) rule-based methods [5, 6]; and (3) machine learning-based approaches [7]. Among them, machine learning-based methods have demonstrated their advantage and showed better performance than the other two categories of methods when a large annotated corpus is available. For example, all the systems in the JNLPBA challenge used one or more machine learning algorithms and greatly outperformed the dictionary-based baseline system [2].

Various machine learning algorithms have been used in BNER, including hidden Markov models (HMM) [8, 9], maximum entropy Markov models (MEMM) [10, 11], conditional random fields (CRF) [12, 13], and support vector machines (SVM) [14, 15]. Among them, CRF have been recognized as a reliable, high-performance algorithm for different BNER-shared tasks [12, 16, 17]. Another important aspect for machine learning-based BNER approaches is features used for building the classification models. Current BNER systems often use different types of linguistic features including morphological, syntactic, semantic information of words, and domain-specific features from biomedical terminologies such as BioThesaurus [18] and UMLS (Unified Medical Language System) [19]. More recently, there is an interest in using new features from unlabeled corpora to improve machine learning-based NER systems. One of the most representative techniques is word representation (WR) [20], which uses unsupervised learning algorithms to generate word-level back-off features from an unlabeled corpus. Those WR features could contain latent syntactic/semantic information of a word. Currently, only very few studies have applied WR features to BNER tasks. For example, Kuksa and Qi investigated the effect of distributed WR features for BNER and their evaluation using BioCreativeII GM corpus showed a significant improvement when adding these features [21].

A large number of techniques have been proposed to extract WR features, such as hyperspace analogue to language (HAL) [22], LSA (latent semantic analysis) [23], latent Dirichlet allocation (LDA) [24], random indexing (RI) [25], canonical correlation analysis (CCA) [26], Brown clustering [27], and neural language models [28–32]. According to a review by Turian et al. [20], WR features can be divided into three categories: (1) clustering-based methods such as Brown clustering [27]; (2) distributional representations, such as LSA [23], LDA [24], and random indexing [25]; and (3) word embeddings (also called distributed representations), such as neural language models [28]. Recently, WR techniques have been widely used to improve various machine learning-based NLP tasks, such as part-of-speech (POS), chunking, and NER in newswire domain [20], and entity recognition in clinical text [33–35]. Word embeddings have also been applied to the biomedical domain and showed improvement on entity recognition in biomedical literature [21]. Nevertheless, the contribution of different types of WR features to BNER has not been extensively investigated yet.

The goal of this study is to systematically evaluate three types of WR features, as well as their combinations, on BNER tasks. We selected one algorithm from each of the three types of WR features and applied them to the JNLPBA and BioCreAtIvE II BNER tasks. Our results showed that all the three WR algorithms were beneficial to machine learning-based BNER systems. Moreover, these different WR features were also complementary to each other. By combining all the three types of WR features, the improvements in $F$-measure on the BioCreAtIvE II GM and JNLPBA corpora were 3.75% and 1.39%, respectively, when compared with the systems using baseline features. To the best of our knowledge, this is the first study to systematically evaluate the effect of three different types of WR features for BNER tasks.

## 2. Materials and Methods

*2.1. Data Sets.* Our experiments were conducted on the BioCreAtIvE II GM corpus and JNLPBA corpus. The BioCreAtIvE II GM corpus consists of 20,000 sentences (15,000 sentences for training and 5,000 sentences for test) from MEDLINE citations, where gene/protein names were manually annotated. The JNLPBA corpus consists of 22,402 sentences from MEDLINE (18,546 sentences for training and 3,856 for test), where five categories of entities (protein, DNA, RNA, cell line, and cell type) were manually annotated. Table 1 shows the counts of different types of entities in two corpora. Sentences are pretokenized in the JNLPBA but not in the BioCreAtIvE II GM corpus. In our experiments, we used GENIA tagger (http://www.nactem.ac.uk/GENIA/tagger/) to perform tokenization for the BioCreAtIvE II GM corpus.

*2.2. Machine Learning Algorithm.* Given the tokenized text, the NER task can be modeled as a sequence labeling problem by assigning each token to a label to determinate the boundaries of named entities, such as B = beginning of an entity, I = inside an entity, and O = outside of an entity (see examples in Table 2). In this study, we used conditional random fields (CRF), a probabilistic undirected graphical model, for two BNER tasks. CRF have been widely used in NER tasks in various domains including biomedicine and have shown the state-of-the-art performance. For example, almost all top-ranked teams in BioCreAtIvE II GM and JNLPBA challenges utilized CRF [2, 3].

*2.3. Features.* In this study, we included four types of features: one set of basic features such as bag-of-word and part-of-speech (POS) and three types of WR features. Although any unlabeled MEDLINE corpus can be used to generate WR features, in this study, we treated the BioCreAtIvE II GM and JNLPBA corpora as unlabeled collections to generate WR features. Details of each type of features are described as follows.

*2.3.1. Basic Features.* Basic features include stemmed words in a context window of [−2, 2], including unigrams, bigrams, and trigrams. Porter stemming algorithm was used to extract the stem of each normalized word. In addition, we also added part-of-speech (POS) tags of words in the same window as features. POS tagging was done by GENIA tagger (http://www.nactem.ac.uk/GENIA/tagger/).

*2.3.2. Clustering-Based WR.* The clustering-based WR induces clusters over words in an unlabeled corpus and represents a word by cluster(s) it belongs to. The idea is that words that are semantically/syntactically similar tend to be in the same or close clusters. Similar to [34], we adopted the Brown clustering algorithm [27] (https://github.com/percyliang/brown-cluster/), a hierarchical clustering algorithm. We ran the Brown clustering algorithm and generated hierarchical clusters of all the words in each corpus, represented by a binary tree, whose leaf nodes are all the words. Figure 1 shows a fragment of a hierarchical cluster containing 7 words from the JNLPBA corpus. The numbers in the squares (e.g., 00) represent the subpaths starting from the root of

Table 1: Counts of different types of entities in two corpora used in this study.

| Corpus | BioCreAtIvE II GM | | JNLPBA | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Gene/protein | Total | Protein | DNA | RNA | Cell line | Cell type | Total |
| Training | 18,265 | 18,265 | 30,269 | 9,534 | 951 | 3,830 | 6,718 | 51,301 |
| Test | 6,331 | 6,331 | 5,067 | 1,056 | 118 | 500 | 1,921 | 8,662 |

Table 2: Examples of named entities represented by BIO labels. The first sentence comes from the JNLPBA corpus and the second sentence comes from the BioCreAtIvE II GM corpus.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Example 1 | Token | IL-2 | gene | expression | and | NF-kappa | B | activation | ⋯ |
| | Label | B-DNA | I-DNA | O | O | B-protein | I-protein | O | ⋯ |
| Example 2 | Token | Comparison | with | alkaline | phosphatases | and | 5 | — | nucleotidase |
| | Label | O | O | B-GM | I-GM | O | B-GM | I-GM | I-GM |

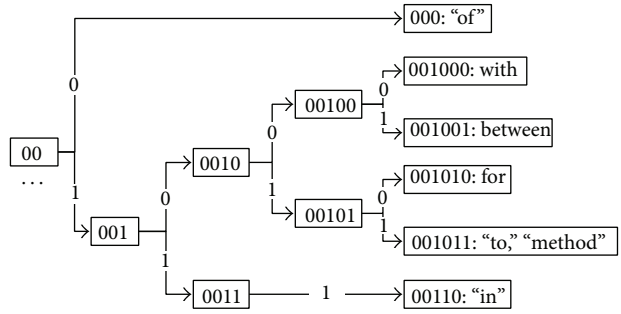



Figure 1: A hierarchical structure fragment generated by Brown clustering for 7 words from the JNLPBA corpus.

the cluster encoded with a binary sequence, and words that share more similar subpaths are semantically closer. In our experiments, all subpaths from the root to a word (i.e., a leaf node) were used as its features. For example, the following features were extracted for the word "for" (001010): {"0," "00," "001," "0010," "00101," and "001010"}. The number of clusters for running Brown clustering algorithm was selected from the set of {50, 100, 200, 500, 1000, and 2000}. The optimized cluster numbers were 500 and 200 on the BioCreAtIvE II GM and JNLPBA corpora, respectively.

*2.3.3. Distributional WR.* The distributional WR is a word cooccurrence-based approach to latent semantics, which uses statistical approximations to reduce a word cooccurrence matrix of high dimensionality to a latent semantic matrix of low dimensionality. Then, a semantic thesaurus can be constructed from the semantic matrix by computing similarities of each word pair or clusters by clustering algorithms. Finally, a word can be represented by other words in the semantic thesaurus or cluster(s) it belongs to. In this study, we reduced dimension of cooccurrence matrix using random indexing [25] and then built a semantic thesaurus using cosine function for semantic similarity computing. Finally, a word was represented by its nearest semantic words (with similarity) in the semantic thesaurus. Table 3 shows a fragment of the semantic thesaurus of 3 words in the JNLPBA corpus. The word in the first row of each column (e.g., "zymosan-tr") is a word in the corpus, and other words in the same column (e.g., "interferon-tr") are words in the semantic thesaurus, sorted by semantic similarity score (e.g., "0.276595744681"). In our experiments, each word was represented by $N$-nearest semantic words, where $N$ was selected from the set of {5, 10, 20, and 50}. The optimized $N$s were 10 and 50 on the BioCreAtIvE II GM and JNLPBA corpora, respectively. For example, the following features were extracted for the word "zymosan-tr": {"interferon-tr": 0.276595744681, "jak-1-defici": 0.243902439024, "p388": 0.236842105263, "ald-induc": 0.228571428571, and "alpha-prolif": 0.22}.

*2.3.4. Word Embeddings.* Word embeddings (also called distributed word representations) induce a real valued latent syntactic/semantic vector for each word from large unlabeled corpus by continuous space language models. A word can be directly represented by its vector and similar words are likely to have similar vectors. In our experiments, we adopted the method in [32] (https://code.google.com/p/word2vec/), a neural network language model to generate word embeddings (shown in Table 4). The dimension of each word vector was selected from the set of {50, 100, 200, and 300}. The optimized dimensions of each word vector were 50 and 100 on the BioCreAtIvE II GM and JNLPBA corpora, respectively.

*2.4. Experiments and Evaluation.* In this study, we started with a baseline system that adopted basic features such as bag-of-word and POS mentioned in the previous section. Then, we evaluated the effect of three types of WR features: clustering-based, distributional word representations, and word embeddings, by adding each of them individually to the baseline system. Furthermore, we evaluated different combinations of three types of WR features. All WR features were derived from the entire unlabeled corpora of BioCreAtIvE II GM and JNLPBA.

We used CRFsuite (http://www.chokkan.org/software/crfsuite/) as an implementation of CRF and optimized its parameters on the training set of each corpus by 10-fold crossvalidation. The optimum number for each type of WR features was also determined during 10-fold crossvalidation. The performance of different approaches was evaluated using the test set of each corpus and reported as standard precision, recall and $F$-measure, calculated using the official evaluation tool provided by the organizers of the two challenges [2, 3].

Table 3: A fragment of the semantic thesaurus of 3 words in the JNLPBA corpus, after running random indexing.

| zymosan-tr | zymogen | ym268 |
|---|---|---|
| interferon-tr: 0.276595744681 | monocyte/b-cell-specif: 0.359477124183 | jak-l: 0.272425249169 |
| jak-l-defici: 0.243902439024 | tubulointerstitium: 0.314720812183 | forskolin: 0.272388059701 |
| p388: 0.236842105263 | c-fms: 0.284768211921 | nf-a 1: 0.265560165975 |
| ald-induc: 0.228571428571 | simplest: 0.282608695652 | icp0: 0.261467889908 |
| alpha-prolif: 0.22 | isotype-specif: 0.277777777778 | betal: 0.25 |
| ... | ... | ... |

Table 4: Word embeddings of 4 words in the JNLPBA corpus. Each number denotes the feature value in a latent semantic/syntactic space.

| |
|---|
| the: 0.067476 −0.017934 0.036855 0.348073 0.063362 −0.138005 −0.144527 −0.014324 0.161269 0.152643 ... |
| of: 0.067905 −0.074922 0.012121 0.050542 0.327945 0.098191 −0.087244 0.194758 0.218592 −0.115941 ... |
| gene: −0.254542 0.100417 −0.124032 0.084818 −0.279409 0.081752 −0.378949 −0.068434 −0.050847 0.142284 ... |
| transcript: −0.157966 −0.303626 0.010010 −0.081133 −0.111763 −0.088829 −0.160671 0.185505 0.097515 −0.014036 ... |

## 3. Results

Table 5 shows the performance of CRF-based BNER approaches on the test sets of BioCreAtIvE II GM and JNLPBA corpora, when three different types of WR features were added individually or in combination. As shown in the table, each individual type of WR features improved the performance of BNER systems. When the clustering-based, distributional, and word embedding WR features were individually added to the basic features, the $F$-measures were improved by 2.1%, 2.86%, and 1.53% on the BioCreAtIvE II GM corpus and by 1.2%, 0.55%, and 0.49% on the JNLPBA corpus, respectively. Different types of WR features seemed to be complementary to each other. BNER systems with any two types of WR features outperformed these with a single type of WR features. For example, when both clustering-based and distributional WR features were used, the $F$-measures were improved by 3.38% on the BioCreAtIvE II GM corpus (versus improvements of 2.1% and 2.86% when either clustering-based or distributional WR features were added to the baseline) and 1.38% on the JNLPBA corpus (versus improvements of 1.2% and 0.55% when either clustering-based or distributional WR features were individually added to the baseline). When all three types of WR features were used, the BNER systems achieved the best performance on both the BioCreAtIvE II GM and JNLPBA corpora, with the highest $F$-measures of 80.96% and 71.39% (improvements of 3.75% and 1.39% compared to the baseline), respectively.

## 4. Discussion

In this paper, we investigated the effect of three types of WR features, including clustering-based representation, distributional representation, and word embeddings, on machine learning-based BNER systems. Evaluation on both the BioCreAtIvE II GM and JNLPBA corpora showed that each type of WR features was beneficial to the CRF-based BNER systems, with an $F$-measure improvement ranging from 0.49% to 2.86%. Moreover, our results also demonstrated that combining different types of WR features further improved BNER performance, indicating that these different types of WR features were complementary to each other. All these findings provide valuable insight into efficient use of WR features in BNER tasks.

Another interesting finding is that the improvements by different WR features varied among different corpora. For example, the distributional WR features achieved the highest improvement on the BioCreAtIvE II GM corpus (i.e., 2.86% in $F$-measure), while it was the clustering-based features that achieved the highest improvement on the JNLPBA corpus (i.e., 1.2% in $F$-measure). We also noticed that the performance gain by WR features was mainly from higher recalls, because unsupervised word representation features could help detect more entities that do not appear in the training data set. For example, the "Baseline+WR1+WR2+WR3" system detected additional 476 entities (288 entities were correct) on the JNLPBA corpus, when compared with the "Baseline" system.

To compare our system with other state-of-the-art BNER systems, we further included additional features to our best systems, including word shape, prefixes, suffixes, orthographic features, and morphological features, all of which were widely used in previously developed BNER systems [9]. The best $F$-measures with all the features were 85.83% and 72.74% on the BioCreAtIvE II GM and JNLPBA corpora, respectively. As expected, WR features were still helpful, though the improvements by WR features were much less (0.2% and 0.3% $F$-measures, resp.) when all other features were used. Anyway, these results are competitive; for example, the $F$-measure on the JNLPBA corpus (72.74%) was higher than the best system in the JNLPBA 2004 challenge. However, our system's performance on BioCreAtIvE II GM was still not as good as others such as [3, 18, 36, 37]. The main reason is that those systems used extensive domain knowledge, ensemble approaches, or postprocessing modules. We believe that adding WR features to these existing systems would further improve their performance.

This study has limitations. For each type of WR features, only one algorithm was implemented and evaluated. It is worth investigating other algorithms in each type of WR features, which is one of our future works. In addition, we treated the annotated corpora as unlabeled data sets to generate WR

Table 5: Performance of CRF-based BNER systems when different types of WR features were used.

| System | BioCreAtIvE II GM (%) | | | JNLPBA (%) | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | *F*-measure | Precision | Recall | *F*-measure |
| Baseline | 87.31 | 69.20 | 77.21 | 71.37 | 68.68 | 70.00 |
| Baseline + WR1 | 86.55 | 73.18 | 79.31 | 70.96 | 71.44 | 71.20 |
| Baseline + WR2 | 87.34 | 73.91 | 80.07 | 71.59 | 69.55 | 70.55 |
| Baseline + WR3 | 86.56 | 72.22 | 78.74 | 71.11 | 69.88 | 70.49 |
| Baseline + WR1 + WR2 | 86.56 | 75.39 | 80.59 | 70.99 | 71.77 | 71.38 |
| Baseline + WR1 + WR3 | 85.77 | 74.65 | 79.82 | 70.77 | 71.87 | 71.31 |
| Baseline + WR2 + WR3 | 87.03 | 74.90 | 80.51 | 71.19 | 70.41 | 70.80 |
| Baseline + WR1 + WR2 + WR3 | 86.54 | 76.05 | 80.96 | 70.78 | 72.00 | 71.39 |

*WR1, WR2, and WR3 denote three different types of word representation features: clustering-based, distributional, and word embeddings features, respectively.

features. In reality, we could generate WR features from a much larger unlabeled corpus such as MEDLINE, which may achieve even higher performance.

## 5. Conclusions

In this study, we investigated the use of three different types of WR features in biomedical entity recognition. Our evaluation on the BioCreAtIvE II GM and JNLPBA corpora showed that not only individual types of WR features were beneficial to BNER tasks but also different types of WR features could be combined and further improve the performance of BNER systems.

## Conflict of Interests

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

## Authors' Contribution

The work presented here was carried out in collaboration between all authors. Buzhou Tang, Hongxin Cao, and Hua Xu designed the methods and experiments. Buzhou Tang, Hongxin Cao, Qingcai Chen, and Xiaolong Wang carried out the experiments. Buzhou Tang, Hongxin Cao, and Hua Xu analyzed the data, interpreted the results, and wrote the paper. All authors have attributed to, seen, and approved the paper. Buzhou Tang and Hongxin Cao contributed equally to this work.

## Acknowledgments

This study is supported in part by Grants CPRIT (Cancer Prevention and Research Institute of Texas) no. R1307, NSFC (National Natural Science Foundation of China) no. 612762383, and CPSF (China Postdoctoral Science Funding) no. 2011M500669. The authors also thank the organizers of the BioCreAtIvE II GM and JNLPBA 2004 challenges.

## References

[1] U. Leser and J. Hakenberg, "What makes a gene name? Named entity recognition in the biomedical literature," *Briefings in Bioinformatics*, vol. 6, no. 4, pp. 357–369, 2005.

[2] J.-D. Kim, T. Ohta, Y. Tsuruoka, Y. Tateisi, and N. Collier, "Introduction to the bio-entity recognition task at JNLPBA," in *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pp. 70–75, Stroudsburg, Pa, USA, 2004.

[3] L. Smith, L. K. Tanabe, R. Ando et al., "Overview of BioCreative II gene mention recognition," *Genome Biology*, vol. 9, 2, article S2, 2008.

[4] R. Gaizauskas, G. Demetriou, and K. Humphreys, "Term Recognition and Classification in Biological Science Journal Articles," in *Proceedings of the Computional Terminology for Medical and Biological Applications Workshop of the 2nd International Conference on NLP*, pp. 37–44, 2000.

[5] K. Fukuda, A. Tamura, T. Tsunoda, and T. Takagi, "Toward information extraction: identifying protein names from biological papers," *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pp. 707–718, 1998.

[6] D. Proux, F. Rechenmann, L. Julliard, V. V. Pillet, and B. Jacq, "Detecting gene symbols and names in biological texts: a first step toward pertinent information extraction," *Genome Informatics Work. Genome Informatics*, vol. 9, pp. 72–80, 1998.

[7] C. Nobata, N. Collier, and J. Tsujii, "Automatic term identification and classification in biology texts," in *Proceedings of the 5th NLPRS*, pp. 369–374, 1999.

[8] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[9] S. Zhao, "Named entity recognition in biomedical texts using an HMM model," in *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pp. 84–87, Stroudsburg, Pa, USA, 2004.

[10] A. Mccallum, D. Freitag, and F. Pereira, "Maximum entropy markov models for information extraction and segmentation," in *Proceedings of the 17th International Conference on Machine Learning*, pp. 591–598, 2000.

[11] J. Finkel, S. Dingare, H. Nguyen, M. Nissim, C. Manning, and G. Sinclair, "Exploiting context for biomedical entity recognition:

from syntax to the web," in *Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its applications (JNLPBA '04)*, 2004.

[12] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: probabilistic models for segmenting and labeling sequence data," in *Proceedings of the 18th International Conference on Machine Learning*, pp. 282–289, San Francisco, Calif, USA, 2001.

[13] B. Settles, "Biomedical named entity recognition using conditional random fields and rich feature sets," in *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pp. 104–107, Stroudsburg, Pa, USA, 2004.

[14] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[15] L. Si, T. Kanungo, and X. Huang, "Boosting performance of bio-entity recognition by combining results from multiple systems," in *Proceedings of the 5th International Workshop on Bioinformatics*, pp. 76–83, New York, NY, USA, 2005.

[16] N. Ponomareva, P. Rosso, F. Pla, and A. Molina, *Conditional Random Fields Vs. Hidden Markov Models in a Biomedical Named Entity Recognition Task*, 2007.

[17] F. Liu, Y. Chen, and B. Manderick, "Named entity recognition in biomedical literature: a comparison of support vector machines and conditional random fields," in *Enterprise Information Systems*, J. Filipe, J. Cordeiro, and J. Cardoso, Eds., pp. 137–147, Springer, Berlin, Germany, 2009.

[18] H. Liu, Z.-Z. Hu, J. Zhang, and C. Wu, "BioThesaurus: a web-based thesaurus of protein and gene names," *Bioinformatics*, vol. 22, no. 1, pp. 103–105, 2006.

[19] O. Bodenreider, "The Unified Medical Language System (UMLS): integrating biomedical terminology," *Nucleic Acids Research*, vol. 32, no. supplement 1, pp. D267–D270, 2004.

[20] J. Turian, L. Ratinov, and Y. Bengio, "Word representations: a simple and general method for semi-supervised learning," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL '10)*, pp. 384–394, Stroudsburg, Pa, USA, July 2010.

[21] P. P. Kuksa and Y. Qi, "Semi-supervised bio-named entity recognition with word-codebook learning," in *Proceedings of the SIAM International Conference on Data Mining (SDM '10)*, pp. 25–36, Columbus, Ohio, USA, April 2010.

[22] K. Lund and C. Burgess, "Producing high-dimensional semantic spaces from lexical co-occurrence," *Behavior Research Methods, Instruments, and Computers*, vol. 28, no. 2, pp. 203–208, 1996.

[23] T. Hofmann, "Probabilistic latent semantic analysis," in *Proceedings of the Uncertainty in Artificial Intelligence (UAI '99)*, pp. 289–296, 1999.

[24] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, no. 4-5, pp. 993–1022, 2003.

[25] P. Kanerva, J. Kristoferson, and A. Holst, "Random indexing of text samples for latent semantic analysis," in *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, pp. 103–106, 2000.

[26] D. R. Hardoon, S. Szedmak, O. Szedmak, and J. Shawe-taylor, *Canonical Correlation Analysis; An Overview with Application to Learning Methods*, 2007.

[27] P. F. Brown, P. V. deSouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai, "Class-based n-gram models of natural language," *Computational Linguistics*, vol. 18, pp. 467–479, 1992.

[28] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *Journal of Machine Learning Research*, vol. 3, no. 6, pp. 1137–1155, 2003.

[29] Y. Bengio, H. Schwenk, J. -S. Senécal, F. Morin, and J.-L. Gauvain, "Neural probabilistic language models," in *Innovations in Machine Learning*, P. D. E. Holmes and P. L. C. Jain, Eds., pp. 137–186, Springer, Berlin, Germany, 2006.

[30] T. Mikolov, M. Karafiát, L. Burget, C. Jan, and S. Khudanpur, "Recurrent neural network based language model," in *Proceedings of the 11th Annual Conference of the International Speech Communication Association: Spoken Language Processing for All (INTERSPEECH '10)*, pp. 1045–1048, September 2010.

[31] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of Machine Learning Research*, vol. 12, pp. 2493–2537, 2011.

[32] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, 2013.

[33] B. Tang, H. Cao, Y. Wu, M. Jiang, and H. Xu, "Clinical entity recognition using structural support vector machines with rich features," in *Proceedings of the ACM 6th International Workshop on Data and Text Mining in Biomedical Informatics*, pp. 13–20, New York, NY, USA, 2012.

[34] B. Tang, H. Cao, Y. Wu, M. Jiang, and H. Xu, "Recognizing clinical entities in hospital discharge summaries using Structural Support Vector Machines with word representation features," *BMC Medical Informatics and Decision Making*, vol. 13, no. supplement 1, p. S1, 2013.

[35] B. Tang, Y. Wu, M. Jiang, Y. Chen, J. C. Denny, and H. Xu, "A hybrid system for temporal information extraction from clinical text," *Journal of the American Medical Informatics Association*, 2013.

[36] R. K. Ando, "BioCreative II gene mention tagging system at IBM watson," in *Proceedings of the 2nd Biocreative Challenge Evaluation Workshop*, pp. 101–104, Madrid, Spain, 2007.

[37] K. Ganchev, K. Crammer, F. Pereira et al., "Penn/UMass/CHOP Biocreative II systems," in *Proceedings of the 2nd Biocreative Challenge Evaluation Workshop*, pp. 119–124, 2007.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 907420, 8 pages
http://dx.doi.org/10.1155/2014/907420

*Research Article*

# Integrative Analysis of miRNA-mRNA and miRNA-miRNA Interactions

**Li Guo, Yang Zhao, Sheng Yang, Hui Zhang, and Feng Chen**

*Department of Epidemiology and Biostatistics, School of Public Health, Nanjing Medical University, Nanjing 211166, China*

Correspondence should be addressed to Feng Chen; fengchen@njmu.edu.cn

Received 5 October 2013; Revised 24 November 2013; Accepted 16 December 2013; Published 12 February 2014

Academic Editor: Yufei Huang



MicroRNAs (miRNAs) are small, noncoding regulatory molecules. They are involved in many essential biological processes and act by suppressing gene expression. The present work reports an integrative analysis of miRNA-mRNA and miRNA-miRNA interactions and their regulatory patterns using high-throughput miRNA and mRNA datasets. Aberrantly expressed miRNA and mRNA profiles were obtained based on fold change analysis, and qRT-PCR was used for further validation of deregulated miRNAs. miRNAs and target mRNAs were found to show various expression patterns. miRNA-miRNA interactions and clustered/homologous miRNAs were also found to contribute to the flexible and selective regulatory network. Interacting miRNAs (e.g., miRNA-103a and miR-103b) showed more pronounced differences in expression, which suggests the potential "restricted interaction" in the miRNA world. miRNAs from the same gene clusters (e.g., miR-23b gene cluster) or gene families (e.g., miR-10 gene family) always showed the same types of deregulation patterns, although they sometimes differed in expression levels. These clustered and homologous miRNAs may have close functional relationships, which may indicate collaborative interactions between miRNAs. The integrative analysis of miRNA-mRNA based on biological characteristics of miRNA will further enrich miRNA study.

## 1. Introduction

MicroRNAs (miRNAs) are small (~22 nts) endogenous noncoding RNAs (ncRNAs). They have many biological roles and act by negatively regulating mRNA expression at the post-transcriptional level [1–3]. They suppress gene expression via interaction with their target messenger RNAs (mRNAs) and either block the translation process or initiate cleavage. These small regulators have important roles in multiple essential biological processes, including cell differentiation and apoptosis [4]. They are also involved in pathological processes and contribute to occurrence and the development of some cancers [5–7]. Abnormal expression of the small ncRNAs may lead to cell death or abnormal cell phenotypes via miRNA-mRNA interactions [8]. Specifically, abnormally expressed miRNAs have been shown to be crucial contributors and may serve as biomarkers in many human diseases.

Bioinformatics analysis indicates that a specific miRNA can regulate expression of up to thousand mRNAs through miRNA-mRNA association, and a specific mRNA can be regulated by multiple miRNAs. miRNAs may also be regulated as potential targets *in vivo* [9]. The flexible regulatory pattern should exist between the two coding (mRNA) and noncoding (miRNA) RNA molecules. Numerous reports have shown that miRNA-mRNA interaction is more complex than we had believed, and a series of studies have been performed to predict miRNA-mRNA modules and correlation networks using miRNA and mRNA expression profiles [10–13]. miRNA-miRNA interactions can also be observed between natural sense and antisense miRNAs [14–18]. These miRNAs always more pronounced divergences in expression, because they may complementarily bind and restrict each other. Many miRNAs are not randomly distributed but rather clustered on chromosomes and cotranscribed as a single polycistronic transcript [19, 20]. Some of these clustered miRNAs can be considered homologous miRNAs (members of the same miRNA gene family). Their sequences are more similar to each other than those of other miRNAs. This is especially true of miRNAs with conserved seed sequences

(nucleotides 2–8) [21]. These miRNA gene clusters and gene families always have close functional relationships and coregulate or coordinately regulate multiple biological processes [22–24].

Integrative analyses that are based on miRNA-mRNA interactions always aim to develop algorithms or tools [25, 26]. Few studies have addressed the biological characteristics of miRNA in miRNA-mRNA interactions. For example, miRNAs are prone to cluster on the chromosomes, some miRNAs show more sequence similarity than others, and a single miRNA locus can yield a cluster of isomiRs with various 5′ and 3′ ends and length distributions [23, 27, 28]. In the present study, an integrative analysis of miRNA-mRNA was performed based on miRNA and mRNA expression profiles in human HepG2 and L02 cells by applying high-throughput techniques. HepG2 cells are human hepatoma cell lines, and they are a suitable model to study occurrence of development of human hepatocellular carcinoma *in vitro*. L02 is the normal human liver cell line, which is always used as control cell lines of HepG2 cells. The purpose of this study was to improve understanding of miRNA-mRNA interactions in regulatory networks. The patterns of expression of potential miRNA-miRNA pairs were also analyzed comprehensively, and the patterns of expression of miRNAs with potential functional relationships, including members of the same miRNA gene clusters and gene families, were surveyed.

## 2. Materials and Methods

*2.1. miRNA and mRNA Profiling Using High-Throughput Techniques.* HepG2 and L02 cells were obtained from the American Type Tissue Collection. miRNA expression profiles were generated from Illumina Genome Analyzer IIx, and then analyzed using Novoalign software (http://www.novocraft.com/, v2.07011) based on the latest known human pre-miRNAs in the miRBase database (Release 19.0, http://www.mirbase.org/) [29]. To further understand expression patterns of target mRNAs of miRNAs, mRNA expression profiles were assessed using microarray hybridization. Hybridization was performed in Agilent's SureHyb Hybridization Chambers (Human LncRNA Array v2.0, 8 × 60 K, Arraystar).

*2.2. Data Analysis.* Aberrantly expressed miRNAs and mRNAs in HepG2 cells were surveyed and identified via fold change analysis. To filter out rare species with lower levels of relative expression, fold change values were estimated by adding additional units (10 units). A detailed flow chart showing the integrative analysis of miRNA-mRNA is given in Figure 1. The main steps were as follows: (1) Abnormal miRNAs and mRNAs were first surveyed through bioinformatics analysis. miRNA expression analysis was also performed at the isomiR level, including the different selections of isomiRs (the most abundant isomiR, sum of all isomiRs, and the canonical miRNA sequence) [28]. (2) Several deregulated miRNAs were further experimentally validated using qRT-PCR. (3) The potential expression and functional relationships among miRNAs were evaluated through analysis of the patterns of expression of clustered and homologous miRNAs based on miRNA gene clusters and families. miRNA-miRNA pairs with potential interactions were also screened and analyzed. (4) GO/pathway terms were enriched based on deregulated mRNAs and the target mRNAs of miRNAs, and miRNA-mRNA regulatory patterns were predicted based on expression profiles.

The experimentally validated target mRNAs of those abnormal miRNAs were obtained from the miRTarBase and Tarbase databases [30]. Common target mRNAs were subjected to functional enrichment analysis using CapitalBio Molecule Annotation System V4.0 and compared to abnormally expressed mRNA profiles from microarray datasets (MAS, http://bioinfo.capitalbio.com/mas3/). GO and pathway analyses were used to determine the biological roles of deregulated miRNA and mRNA species. Potential miRNA-mRNA and miRNA-miRNA interactions and miRNA/mRNA expression profiles were used to construct functional interaction networks using Cytoscape v2.8.2 Platform [31].

*2.3. qRT-PCR Validation.* Abnormal miRNAs were further validated using quantitative real-time reverse transcription PCR (qRT-PCR) using SYBR premix Ex Taq (Takara, Japan). Samples were amplified using the Mastercycler ep realplex2 system (Eppendorf, Hamburg, Germany). qPCR was performed using specifically designed primers and used to detect hsa-miR-15b/103a/106b (Bulge-Loop miRNA qRT-PCR Primer Set, RiboBio, Guangzhou, China), and U6 served as an internal control. The relative amount of each miRNA was measured using the $2^{(-\Delta\Delta CT)}$ method [32]. All qRT-PCR reactions were carried out in triplicate, and data were presented as the mean ± standard deviation. The two-tailed Student's *t* test was used to compare the expression difference between tumor and normal cells.

## 3. Results

*3.1. Overview of miRNA/mRNA Expression Profiles and Further Experimental Validation.* Upregulated and downregulated miRNAs/mRNAs were identified using the fold change values (log 2) based on the control sample. Many miRNAs and mRNAs were found to be differentially expressed (see Figure S1 in the Supplementary Material available online at http://dx.doi.org/10.1155/2014/907420). miRNA expression patterns were further analyzed at the isomiR level. Fold change values were found to diverge based on the different selections of isomiRs (the most abundant isomiR, sum of all isomiRs, and the canonical miRNA sequence) (Figure 2(a)). Differences in fold change values rarely affected the selection of deregulated miRNA species. The canonical or annotated miRNA sequences were not always the most dominant species in the miRNA locus. They had even lower levels of expression. The qRT-PCR primers used here were designed according to the canonical miRNA sequences in the miRBase database (Release 19.0, http://www.mirbase.org/) [29]. For this reason, in order to further validate deregulated miRNAs using qRT-PCR technique, we randomly selected several abnormally expressed miRNAs (miR-15b, miR-103a,

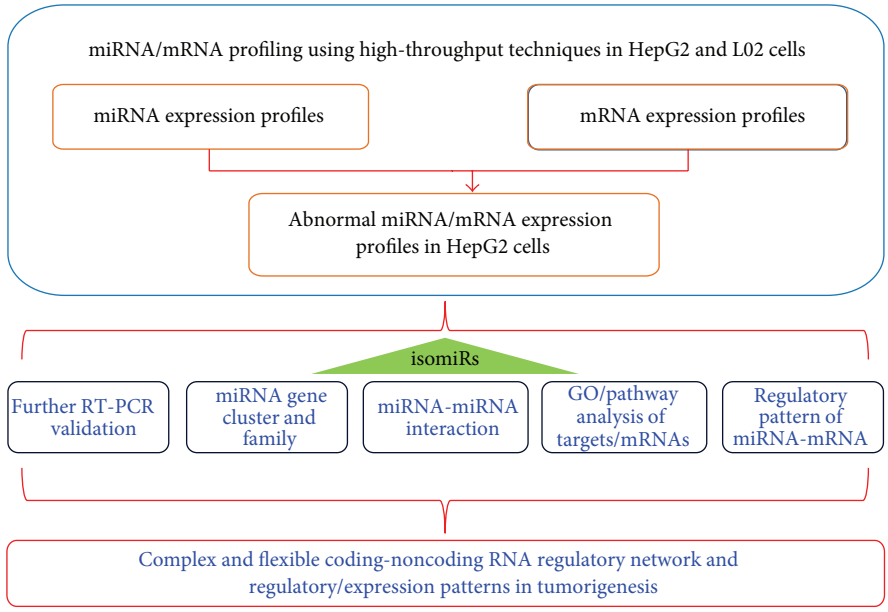



Figure 1: The miRNA-mRNA integrative analysis.

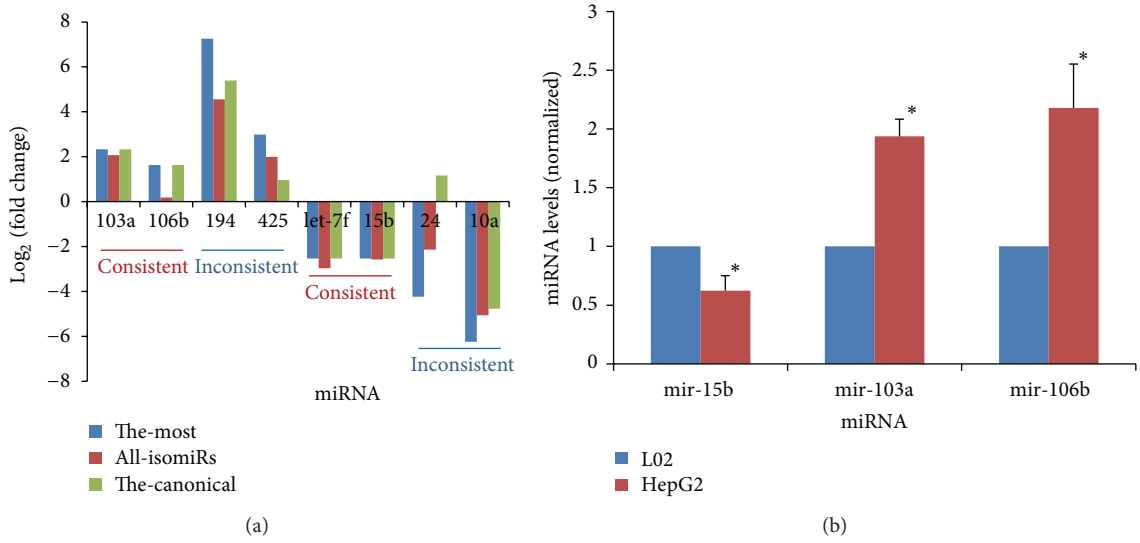



Figure 2: (a) miRNA expression analysis and (b) and further qRT-PCR validation. (a) The fold change values (log 2) differ in the variety of miRNA sequences involved. "The-most" indicates the most abundant and dominant isomiR sequence. "All-isomiRs" indicates sum of all isomiRs. "The-canonical" indicates the reference miRNA sequence in the miRBase database. The canonical miRNA sequence may be consistent or inconsistent with the most abundant isomiR sequence. Different methods of estimation may produce different fold change values (log 2), but they always show consistent deregulation patterns. (b) Further RT-PCR validation is performed for miR-15b, miR-103a, and miR-106b, and the experimental results show consistent deregulation patterns. "*" indicates that the $P$ value is less than 0.05.

and miR-106b; their canonical miRNA sequences were the most abundant isomiRs) for further experimental validation (Figure 2). Bioinformatic analysis showed that miR-103a and miR-106b were upregulated in tumor cells, while miR-15b was identified as downregulated species (Figure 2(a)). As expected, qRT-PCR experimental validation showed consistent results (Figure 2(b)).

*3.2. Expression Patterns of miRNA-miRNA Pairs and miRNA Gene Clusters and Families.* The expression patterns of miRNA-miRNA pairs that can form miRNA-miRNA duplexes were also analyzed [18]. Eight miRNA-miRNA pairs were found to be abundantly expressed in HepG2 or L02 cells. Expression analysis showed one member of each natural miRNA-miRNA pair to be abundantly expressed

Table 1: Differences in expression between natural sense and antisense miRNAs.

| miRNA/miRNA | The most abundant isomiR | | Sum of all isomiRs | | The canonical miRNA | |
|---|---|---|---|---|---|---|
| | HepG2 | L02 | HepG2 | L02 | HepG2 | L02 |
| 103a/103b | 9232/— | 1833/— | 10639/— | 2525/— | 9232/— | 1833/— |
| 122/3591 | 163/— | 938/— | 837/— | 3550/— | 10/— | 17/— |
| 203/3545 | —/— | 216/— | —/— | 705/— | —/— | 12/— |
| 24/3074 | 787/2 | 14998/— | 6618/3 | 29094/— | 3592/— | 1597/— |
| 423-5p/3184-3p | 1208/— | 3159/— | 1882/— | 5790/— | 1208/— | 3159/— |
| 423-3p/3184-5p | 981/— | 5036/1 | 1934/— | 8290/1 | 981/— | 5036/— |
| 7-5p/3529-3p | 1132/— | 939/— | 1931/— | 2397/— | 238/— | 386/— |
| 374b-5p/374c-3p | 137/— | 67/— | 318/— | 203/— | 137/— | 67/— |

Based on the different methods of estimation, the most abundant isomiR, sum of all isomiRs, and the canonical miRNA, relative expression levels of these pairs of miRNA pairs were determined. They are presented here using normalized data. One member of each pair was always far more abundantly expressed than the other. "—" indicates an undetectable miRNA.

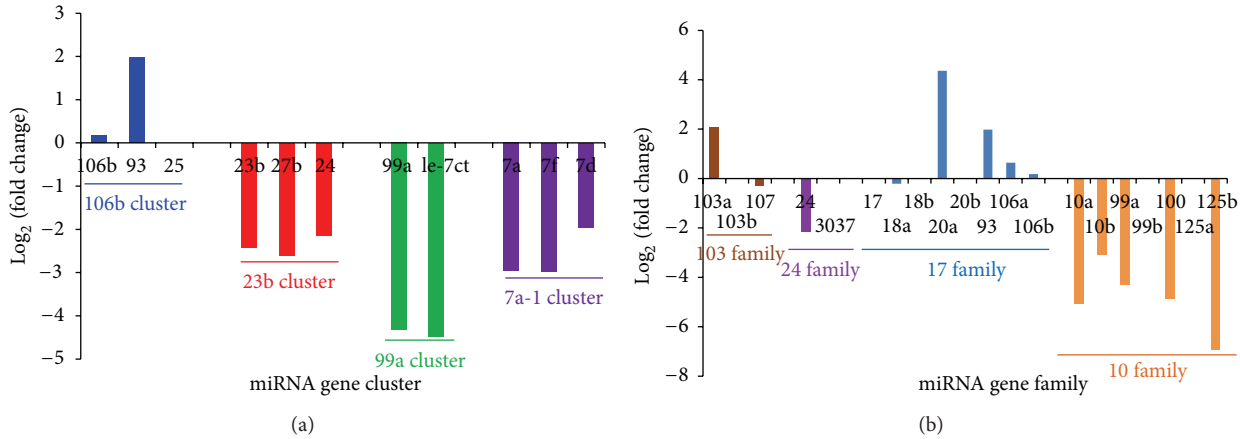



Figure 3: Examples of (a) deregulated miRNA gene clusters and (b) gene families. (a) Clustered and (b) homologous miRNAs are always consistently upregulated or downregulated in tumor cells, although they can differ in fold change values (log 2) and relative expression levels. miRNAs shown here to have zero change (such as miR-25) are not detected or did not show significant differences between tumor and normal cells.

and the other to be quite rare (Table 1). For example, the miR-103a/miR-103b pairs showed a pronounced difference in the degree of expression: miR-103a was abundantly expressed (normalized sequence count was more than 9,232 in tumor cells), and miR-103b was not detected. Pronounced differences in degree of expression were quite common between these two members of each miRNA-miRNA pair (Table 1).

The expression patterns of miRNAs that might have potential functional relationships were also analyzed. Clustered and homologous miRNAs always showed consistent patterns of deregulation (Figure 3), although they could differ in relative level of expression, sometimes showing large differences. These differences in expression may have led to the various fold change values observed between these related miRNA members (Figure 3). For example, the miRNA in the miR-23b gene cluster were downregulated, showing similar fold change values, and those of the miR-106b gene cluster showed highly different fold change values (Figure 3(a)).

*3.3. Expression and Regulatory Patterns of miRNAs/mRNAs and Functional Enrichment Analysis.* Although each aberrantly expressed miRNA can negatively regulate target mRNAs via miRNA-mRNA association, their potential targets always show dramatically different expression patterns (Figure 4). Common target mRNAs might be detected between different deregulated miRNAs, even between upregulated and downregulated miRNAs (according to validated miRNA-mRNA interaction, *E2F3* can be negatively regulated by upregulated miR-106b and downregulated miR-125b, Figure 4). Targets of miRNAs of the same gene clusters and families also showed complex expression patterns, although these related miRNAs were downregulated in tumor cells (Figure 4(b)). These homologous and clustered miRNAs were always simultaneously upregulated or downregulated. They might negatively target the same mRNAs (Figure 4(b)).

Functional enrichment analysis based on the deregulated target mRNAs suggested multiple biological roles (Figures S2, S3, and S4). They were found to contribute to many biological

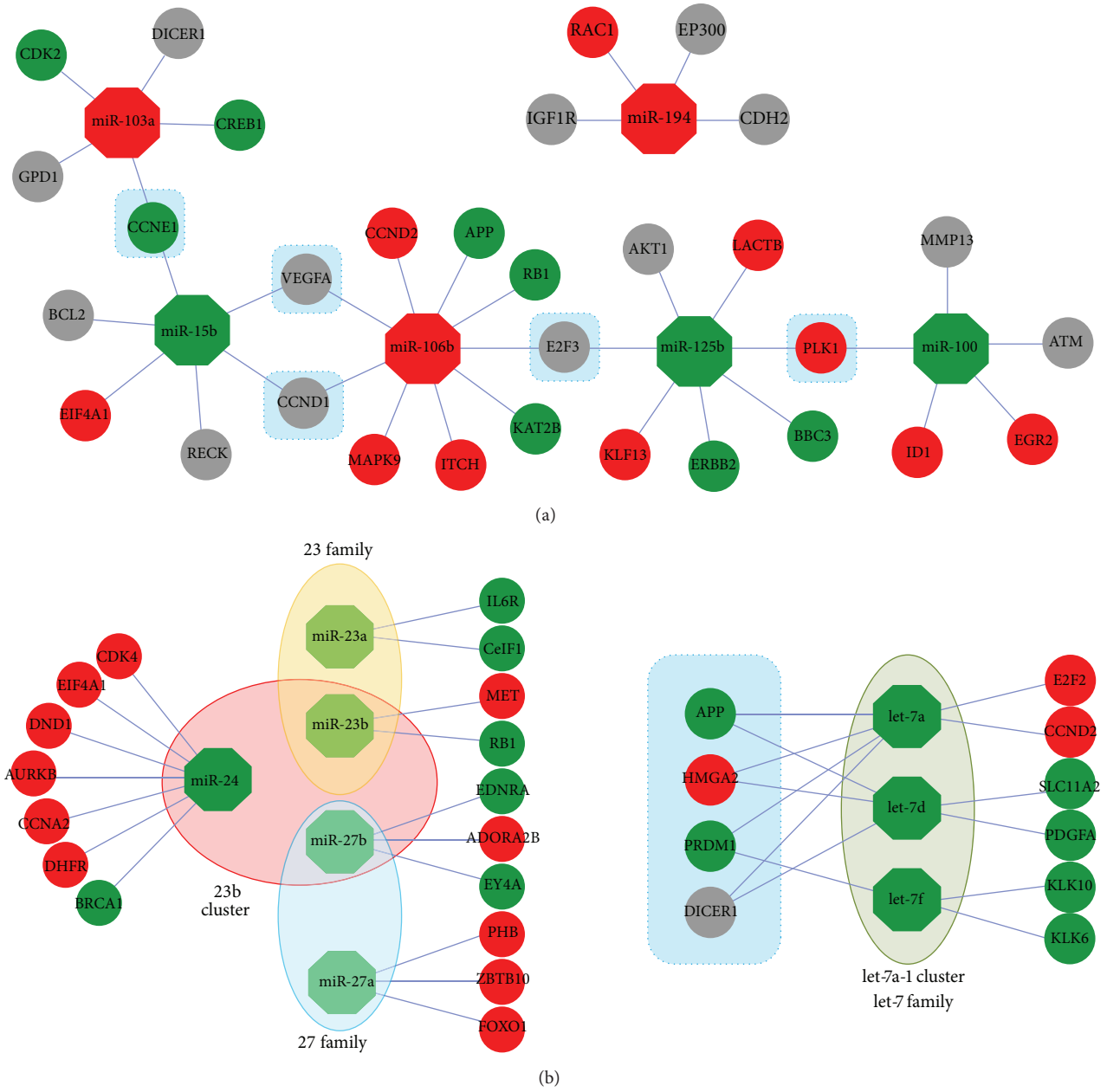



Figure 4: Examples of flexible and selective regulatory network between miRNAs and mRNAs. (a) Selected overexpressed (miR-103a, miR-106b, and miR-194) and underexpressed (miR-15b, miR-100, and miR-125b) miRNAs are used to reconstruct the regulatory network. Their experimentally validated target mRNAs show various expression patterns: some are stably expressed, and others are upregulated or downregulated. Overexpressed miRNAs and mRNAs are here highlighted in red octagons and ellipse, respectively, and underexpressed miRNAs and mRNAs are highlighted in green octagon and ellipse, respectively. Grey ellipses indicate stably expressed mRNAs and mRNAs are not detected in the present study. The targets common to different miRNAs are highlighted in blue rectangles. (b) Selected underexpressed miRNA gene clusters (miR-23b and let-7a-1) and gene families (miR-23 and miR-27) also show complex regulatory networks. These clustered and homologous members are consistently downregulated in tumor cells, and their validated targets show various expression patterns. miRNAs in the let-7a-1 gene cluster are also members of the let-7 gene family. The targets common to these miRNAs have shown upregulated, downregulated, and stable patterns of expression.

processes, such as the cell cycle, calcium signaling pathway, p53 signaling pathway, and T cell receptor signaling pathway. These aberrantly expressed mRNA species are also involved in some human diseases, including pancreatic cancer, renal cell carcinoma, prostate cancer, and colorectal cancer.

## 4. Discussion

In the study, integrative analysis of miRNA-mRNA is performed using biological characteristic of miRNAs, and miRNA-miRNA interaction is simultaneously analyzed based

on the relationships between different miRNAs (Figure 1). Compared to other algorithms or tools of miRNA-mRNA analysis [25, 26], the approach aims to track miRNA-mRNA and miRNA-miRNA interactions based on characteristic of miRNAs. Specifically, (1) miRNAs are prone to detected homologous miRNAs with higher level of sequence similarity, (2) miRNAs are prone to cluster together with close physical distance, (3) some miRNAs are located on sense and antisense strands of specific genomic regions, and (4) miRNA locus can generate multiple isomiRs with various sequences and expression levels, and so forth. Although these specific features of miRNAs have been widely concerned in miRNA study, they are rarely mentioned or involved in miRNA-mRNA analysis. Indeed, many miRNAs coordinately contribute to biological processes, and one specific biological pathway will involved in a series of mRNAs and regulatory miRNAs. Therefore, it is quite necessary to study miRNA-mRNA interactions using characteristic of miRNAs, especially homologous and/or clustered miRNAs are prone to have functional relationships. More importantly, the canonical or annotated miRNA sequence is only one specific member of the multiple isomiRs, and the study at the isomiR level will enrich miRNA study. IsomiR expression patterns contribute to tracking pre-miRNA processing and miRNA maturation processes and understanding regulatory network at the isomiR levels.

According to the integrative analysis method, firstly, aberrantly expressed miRNA and mRNA profiles were collected based on fold change analysis. To further validate these deregulated miRNA species, several deregulated miRNAs that had been experimentally validated using qRT-PCR were randomly selected. As expected, qRT-PCR experiments showed results consistent with those of bioinformatic analysis (Figure 2). As in other reports, miR-103a, and miR-106b were overexpressed in hepatocellular carcinoma (HCC) and served as important negative regulators [33, 34]. However, miR-15b was found to be upregulated [35]. The overexpression of miR-15b may restrict cell proliferation and increase the rate of cellular apoptosis, and abundant expression may indicate a low risk of HCC recurrence [36]. The dynamic expression of miR-15b may play multiple biological roles in tumorigenesis.

Many reports have shown that multiple isomiRs (miRNA variants) can be detected at the same miRNA locus. This is due to imprecise and alternative cleavage of Drosha and Dicer [23, 27, 28]. According to three different methods of estimation methods, the most abundant isomiR, the sum of all isomiRs, and the canonical miRNA, the phenomenon of the multiple miRNA variants may influence the relative expression levels and lead to various fold change values (Figure 2(a)) [24, 28]. This is mainly because of differences among isomiR repertoires and expression patterns, although they are always well conserved across different tissues and animal species [28, 37, 38]. Differences in isomiR expression profiles may play a role in occurrence and development of disease [28]. Generally, consistent deregulated miRNAs could be identified using different methods of estimation methods, even if they have different fold change values (Figure 2(a)). However, if abnormal miRNA expression profiles are collected using the typical methods of analysis of canonical miRNA or the sum of all isomiRs, the difference in fold change values may affect the collection of deregulated miRNA species and may require further experimental validation. Among multiple isomiRs, the canonical miRNAs are not always the most abundant (Figure 2(a)). Some of them can be very rare. Other abundant isomiR species, especially isomiRs with novel 5′ ends and seed sequences (5′ isomiRs), may also be regulatory molecules. These 5′ isomiRs may have novel potential target mRNAs and may contribute to the regulation of previously unknown biological processes. Collectively, it may be best to observe deregulated miRNAs through bioinformatic analysis at the miRNA level using the most abundant and dominant isomiR sequence and isomiR profiles through bioinformatic analysis at the isomiR level based on variations in sequence and expression levels.

miRNAs negatively regulate mRNA expression and contribute to many biological processes through complementary binding to their target mRNAs. Some miRNAs can interact with the 3′-untranslated region (UTR) of target mRNA and reduce the level of mRNA expression [39]. An attempt was here made to reconstruct the coding-noncoding RNA regulatory network according to negative regulation and the deregulation of miRNAs and target mRNAs. Although miRNAs can be either downregulated or upregulated in tumor cells, their experimentally validated and predicted targets may show consistent or inconsistent deregulation patterns (Figure 4). Abnormal miRNA and mRNA expression profiles complicate the regulatory network, although they showed close functional relationships by forming miRNA-mRNA duplexes. A single miRNA can regulate multiple target mRNAs and vice versa. The fact that a single miRNA can engage in many possible miRNA-mRNA interactions can render regulatory networks highly complex. Flexible regulatory patterns indicate that a specific miRNA may regulate selected specific targets and so contribute to specific stages of development. miRNA-mRNA may affect the spatial-temporal expression patterns of miRNAs, but these interactions can also be more strictly regulated during specific stages of development. The selection of regulated target mRNAs may have been driven by functional pressure in cellular environments through complex regulatory mechanisms. In this way, overexpressed, underexpressed, and stably expressed target mRNAs can be identified for specific upregulated and downregulated miRNAs (Figure 4). A single mRNA can be negatively regulated by selected specific miRNAs. The coding-noncoding RNA regulatory network is more complexity than previously thought, especially for complicated and selective multiple interactions of miRNAs and mRNAs (Figure 4).

Functional miRNA groups also contribute to the complexity of regulatory networks. miRNAs that have completely or partially complementary structures can form miRNA-miRNA duplexes through reverse complementary binding events. They can also form miRNA:miRNA* or miRNA-#-5p:mRNA-#-3p duplexes [14, 16–18]. miRNA:miRNA interactions are specific phenomenon. They are especially common between natural or endogenous sense and antisense miRNAs. Possibly because of restricted interactions,

these miRNA-miRNA pairs show greater differences in the level expression than other miRNAs do: one member typically has a far higher level of enrichment than the other, which can be quite rare (Table 1). This indicates that restricted interactions may be a regulatory pattern in the miRNA world. Another, very different, type of interaction between miRNAs, termed coordinated interaction, also contributes to the pronounced efficiency of the regulatory process. Some miRNAs, such as clustered and homologous miRNA species, may coregulate or coordinately regulate biological processes [19, 40]. They may be located close to another (clustered in the same genomic region, miRNA gene cluster) or may share sequence similarity (homologous miRNAs, miRNA gene family). Some clustered miRNAs share sequence similarity and are identified as both members of the same cluster and of the same family. These phenomena are not random but rather derived from functional and evolutionary pressures. These related miRNAs always show similar or consistent patterns of deregulation (Figure 3), although they may have different levels of enrichment because of maturation and degradation mechanisms. Deregulation patterns may cause functional relationships. This indicates that collaborative interactions may take place within the coding-noncoding RNA regulatory network. Therefore, related miRNAs further complicate the regulatory patterns, especially when they share specific target mRNAs. In summary, coordinated interactions and restricted interactions both exist in the world of small, noncoding RNA. Although they can be thought of as indirect and direct interactions, respectively, these interactions represent the versatility and complexity of the functional and evolutionary relationships among different miRNAs. miRNA-miRNA interactions enrich and complicate the coding-noncoding RNA regulatory network and contribute to the robustness of the regulatory network in organism.

## Conflict of Interests

The authors declare no potential conflict of interests with respect to the authorship and/or publication of this paper.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China (nos. 61301251, 81072389, and 81373102), the Research Fund for the Doctoral Program of Higher Education of China (no. 211323411002 and 20133234120009), the China Postdoctoral Science Foundation funded project (no. 2012M521100), the key Grant of the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (no. 10KJA33034), the National Natural Science Foundation of Jiangsu (no. BK20130885), the Natural Science Foundation of the Jiangsu Higher Education Institutions (nos. 12KJB310003 and 13KJB330003), the Jiangsu Planned Projects for Postdoctoral Research Funds (no. 1201022B), the Science and Technology Development Fund Key Project of Nanjing Medical University (no. 2012NJMU001), and the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).

## References

[1] R. W. Carthew and E. J. Sontheimer, "Origins and mechanisms of miRNAs and siRNAs," *Cell*, vol. 136, no. 4, pp. 642–655, 2009.

[2] H. Guo, N. T. Ingolia, J. S. Weissman, and D. P. Bartel, "Mammalian microRNAs predominantly act to decrease target mRNA levels," *Nature*, vol. 466, no. 7308, pp. 835–840, 2010.

[3] E. Huntzinger and E. Izaurralde, "Gene silencing by microRNAs: contributions of translational repression and mRNA decay," *Nature Reviews Genetics*, vol. 12, no. 2, pp. 99–110, 2011.

[4] L. B. Frankel, N. R. Christoffersen, A. Jacobsen, M. Lindow, A. Krogh, and A. H. Lund, "Programmed cell death 4 (PDCD4) is an important functional target of the microRNA miR-21 in breast cancer cells," *The Journal of Biological Chemistry*, vol. 283, no. 2, pp. 1026–1033, 2008.

[5] P. M. Voorhoeve, C. le Sage, M. Schrier et al., "A genetic screen implicates miRNA-372 and miRNA-373 as oncogenes in testicular germ cell tumors," *Cell*, vol. 124, no. 6, pp. 1169–1181, 2006.

[6] W. C. S. Cho, "OncomiRs: the discovery and progress of microRNAs in cancers," *Molecular Cancer*, vol. 6, article 60, 2007.

[7] S. M. Hammond, "MicroRNAs as tumor suppressors," *Nature Genetics*, vol. 39, no. 5, pp. 582–583, 2007.

[8] B. J. Reinhart, F. J. Slack, M. Basson et al., "The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans," *Nature*, vol. 403, no. 6772, pp. 901–906, 2000.

[9] 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, 2010.

[10] X. Peng, Y. Li, K.-A. Walters et al., "Computational identification of hepatitis C virus associated microRNA-mRNA regulatory modules in human livers," *BMC Genomics*, vol. 10, article 373, 2009.

[11] B. Liu, L. Liu, A. Tsykin et al., "Identifying functional miRNA-mRNA regulatory modules with correspondence latent dirichlet allocation," *Bioinformatics*, vol. 26, no. 24, pp. 3105–3111, 2010.

[12] C. Girardi, C. de Pittà, S. Casara et al., "Analysis of miRNA and mRNA expression profiles highlights alterations in ionizing radiation response of human lymphocytes under modeled microgravity," *PLoS ONE*, vol. 7, no. 2, Article ID e31293, 2012.

[13] W. Zhang, A. Edwards, W. Fan, E. K. Flemington, and K. Zhang, "miRNA-mRNA correlation-network modules in human prostate cancer and the differences between primary and metastatic tumor subtypes," *PLoS One*, vol. 7, Article ID e40130, 2012.

[14] E. C. Lai, C. Wiel, and G. M. Rubin, "Complementary miRNA pairs suggest a regulatory role for miRNA:miRNA duplexes," *RNA*, vol. 10, no. 2, pp. 171–175, 2004.

[15] C. F. Hongay, P. L. Grisafi, T. Galitski, and G. R. Fink, "Antisense transcription controls cell fate in Saccharomyces cerevisiae," *Cell*, vol. 127, no. 4, pp. 735–745, 2006.

[16] A. Stark, N. Bushati, C. H. Jan et al., "A single Hox locus in Drosophila produces functional microRNAs from opposite DNA strands," *Genes & Development*, vol. 22, no. 1, pp. 8–13, 2008.

[17] L. Guo, T. Liang, W. Gu, Y. Xu, Y. Bai, and Z. Lu, "Cross-mapping events in miRNAs reveal potential miRNA-Mimics and evolutionary implications," *PLoS ONE*, vol. 6, no. 5, Article ID e20517, 2011.

[18] L. Guo, B. Sun, Q. Wu, S. Yang, and F. Chen, "miRNA-miRNA interaction implicates for potential mutual regulatory pattern," *Gene*, vol. 511, pp. 187–194, 2012.

[19] L. P. Lim, M. E. Glasner, S. Yekta, C. B. Burge, and D. P. Bartel, "Vertebrate microRNA genes," *Science*, vol. 299, no. 5612, p. 1540, 2003.

[20] V. N. Kim and J.-W. Nam, "Genomics of microRNA," *Trends in Genetics*, vol. 22, no. 3, pp. 165–173, 2006.

[21] A. A. Aravin, M. Lagos-Quintana, A. Yalcin et al., "The small RNA profile during Drosophila melanogaster development," *Developmental Cell*, vol. 5, no. 2, pp. 337–350, 2003.

[22] J. Yu, F. Wang, G.-H. Yang et al., "Human microRNA clusters: genomic organization and expression profile in leukemia cell lines," *Biochemical and Biophysical Research Communications*, vol. 349, no. 1, pp. 59–68, 2006.

[23] P. Landgraf, M. Rusu, R. Sheridan et al., "A mammalian microRNA expression atlas based on small RNA library sequencing," *Cell*, vol. 129, no. 7, pp. 1401–1414, 2007.

[24] L. Guo and Z. Lu, "Global expression analysis of miRNA gene cluster and family based on isomiRs from deep sequencing data," *Computational Biology and Chemistry*, vol. 34, no. 3, pp. 165–171, 2010.

[25] N. Hecker, C. Stephan, H. J. Mollenkopf, K. Jung, R. Preissner, and H. A. Meyer, "A new algorithm for integrated analysis of miRNA-mRNA interactions based on individual classification reveals insights into bladder cancer," *PLoS One*, vol. 8, Article ID e64543, 2013.

[26] M. Khorshid, J. Hausser, M. Zavolan, and E. van Nimwegen, "A biophysical miRNA-mRNA interaction model infers canonical and noncanonical targets," *Nature Methods*, vol. 10, pp. 253–255, 2013.

[27] R. D. Morin, M. D. O'Connor, M. Griffith et al., "Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells," *Genome Research*, vol. 18, no. 5, pp. 610–621, 2008.

[28] L. Guo, Q. Yang, J. Lu et al., "A comprehensive survey of miRNA repertoire and 3′ addition events in the placentas of patients with pre-eclampsia from high-throughput sequencing," *PLoS ONE*, vol. 6, no. 6, Article ID e21072, 2011.

[29] A. Kozomara and S. Griffiths-Jones, "MiRBase: integrating microRNA annotation and deep-sequencing data," *Nucleic Acids Research*, vol. 39, no. 1, pp. D152–D157, 2011.

[30] T. Vergoulis, I. S. Vlachos, P. Alexiou et al., "TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support," *Nucleic Acids Research*, vol. 40, pp. D222–D229, 2012.

[31] M. E. Smoot, K. Ono, J. Ruscheinski, P.-L. Wang, and T. Ideker, "Cytoscape 2.8: new features for data integration and network visualization," *Bioinformatics*, vol. 27, no. 3, pp. 431–432, 2011.

[32] K. J. Livak and T. D. Schmittgen, "Analysis of relative gene expression data using real-time quantitative PCR and the 2-ΔΔCT method," *Methods*, vol. 25, no. 4, pp. 402–408, 2001.

[33] R. R. Wei, G. L. Huang, M. Y. Zhang et al., "Clinical significance and prognostic value of microRNA expression signatures in hepatocellular carcinoma," *Clinical Cancer Research*, vol. 19, no. 17, pp. 4780–4791, 2013.

[34] W. L. Yau, C. S. Lam, L. Ng et al., "Over-expression of miR-106b promotes cell migration and metastasis in hepatocellular carcinoma by activating epithelial-mesenchymal transition process," *PLoS One*, vol. 8, Article ID e57882, 2013.

[35] F. An, B. Gong, H. Wang et al., "miR-15b and miR-16 regulate TNF mediated hepatocyte apoptosis via BCL2 in acute liver failure," *Apoptosis*, vol. 17, pp. 702–716, 2012.

[36] G. E. Chung, J.-H. Yoon, S. J. Myung et al., "High expression of microRNA-15b predicts a low risk of tumor recurrence following curative resection of hepatocellular carcinoma," *Oncology Reports*, vol. 23, no. 1, pp. 113–119, 2010.

[37] A. M. Burroughs, Y. Ando, M. J. L. de Hoon et al., "A comprehensive survey of 3′ animal miRNA modification events and a possible role for 3′ adenylation in modulating miRNA targeting effectiveness," *Genome Research*, vol. 20, no. 10, pp. 1398–1410, 2010.

[38] 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.

[39] J. Soh, J. Iqbal, J. Queiroz, C. Fernandez-Hernando, and M. M. Hussain, "MicroRNA-30c reduces hyperlipidemia and atherosclerosis in mice by decreasing lipid synthesis and lipoprotein secretion," *Nature Medicine*, vol. 19, pp. 892–900, 2013.

[40] J. Z. Xu and C. W. Wong, "A computational screen for mouse signaling pathways targeted by microRNA clusters," *RNA*, vol. 14, no. 7, pp. 1276–1283, 2008.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 258784, 9 pages
http://dx.doi.org/10.1155/2014/258784

*Research Article*

# Network-Assisted Prediction of Potential Drugs for Addiction

**Jingchun Sun,[1] Liang-Chin Huang,[1] Hua Xu,[1] and Zhongming Zhao[2,3,4]**

[1] *School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA*
[2] *Department of Biomedical Informatics, Vanderbilt University School of Medicine, 2525 West End Avenue, Suite 600, Nashville, TN 37203, USA*
[3] *Department of Psychiatry, Vanderbilt University School of Medicine, Nashville, TN 37212, USA*
[4] *Department of Cancer Biology, Vanderbilt University School of Medicine, Nashville, TN 37232, USA*

Correspondence should be addressed to Zhongming Zhao; zhongming.zhao@vanderbilt.edu

Received 22 November 2013; Accepted 30 December 2013; Published 9 February 2014

Academic Editor: Yufei Huang



Drug addiction is a chronic and complex brain disease, adding much burden on the community. Though numerous efforts have been made to identify the effective treatment, it is necessary to find more novel therapeutics for this complex disease. As network pharmacology has become a promising approach for drug repurposing, we proposed to apply the approach to drug addiction, which might provide new clues for the development of effective addiction treatment drugs. We first extracted 44 addictive drugs from the NIDA and their targets from DrugBank. Then, we constructed two networks: an addictive drug-target network and an expanded addictive drug-target network by adding other drugs that have at least one common target with these addictive drugs. By performing network analyses, we found that those addictive drugs with similar actions tended to cluster together. Additionally, we predicted 94 nonaddictive drugs with potential pharmacological functions to the addictive drugs. By examining the PubMed data, 51 drugs significantly cooccurred with addictive keywords than expected. Thus, the network analyses provide a list of candidate drugs for further investigation of their potential in addiction treatment or risk.

## 1. Introduction

Drug addiction is a chronic and relapsing brain disease that causes compulsive drug seeking and abuse. The disease affects the brain functions and behavior of many people of all ages. The subjects suffer harmful consequences of drug addiction, which generates an enormous medical, financial, social, and emotional burden on individuals, their families, and our society. During the past several decades, investigators have made numerous efforts to understand the neuronal effects of addictive drugs and the molecular mechanisms of addiction. Such knowledge has facilitated the uncovering of novel targets and drugs for both treating and preventing addictive disorders.

The large body of studies has revealed that genetic and environmental factors contribute to the development of addiction [1]. The genetic studies of twins and families have suggested that genetic factors might account for 30–60% of the overall risk for the development of drug addiction [2, 3]. The recent advent of high-throughput experimental technologies, such as gene expression profiling, genome-wide association studies (GWAS), and next-generation sequencing (NGS), has revolutionized biomedical research and generated a massive amount of data for addiction research [4, 5]. This provides valuable information for further development of addiction treatment. Even so, an effective treatment of drug addiction patients is still unavailable.

Currently, medication and behavioral therapy, especially when combined, are the major therapeutic treatment approaches for addiction [6]. Thus, the discovery of effective drugs with fewer side effects is crucial to provide effective treatment and prevent relapse. During the past decade, advancements in target-based approaches have provided us with a promising direction for further treatment development [7]. Therefore, systematic investigations of addictive drugs and their targets might provide deeper insights into the relationship between individual addictive drugs and nonaddictive drugs. However, the absence of comprehensive

drug-target data is a major limitation in performing a systematic investigation. Recently, DrugBank has provided a comprehensive collection of drugs and their targets [8], which largely eases this problem. Other drug-target databases have become available to assist further computational analyses [9, 10]. Furthermore, the concept of network medicine has been proposed and various approaches have been developed to assist with drug-drug and drug-target discovery [11–14]. We recently applied network pharmacology approach to exploring the features of antipsychotic and illicit drugs as well as their targets and found some interesting drug-target interaction features [13–15]. Here, we expanded our work to perform a systematic investigation of the relationships between multiple addictive drugs and their targets, as well as other drugs that have targets in common with these addictive drugs. The inclusion of addiction-related drugs might help us predict other addiction-related drugs from available drugs through drug repurposing approaches. We hypothesize that some of the addiction-related drugs that have not been assigned as addictive drugs might have the potential to treat addiction, while others might cause addictive effects.

We mainly focused on the addictive drugs annotated by the National Institute on Drug Abuse (NIDA). We extracted their targets from the DrugBank database. We first constructed a basic addictive drug-target network from which we attempted to find the unique connectivity as a proof of the concept of the network pharmacology approach for addictive drugs. Next, we built an expanded addictive drug-target network by recruiting non-addictive drugs. These non-addictive drugs have at least one target in common with at least one addictive drug. We analyzed these two networks by examining their network topological characteristics, which allowed us to explore whether some of the non-addictive drugs in the network might have the potential to either be addictive themselves or have the potential to treat addiction. Finally, to explore some lines of evidence from previous studies, we examined cooccurrence of drugs and addiction-related keywords to evaluate the association of non-addictive drugs with addiction. This preliminary study demonstrated that the network-assisted approach is promising in the prediction of drug repurposing.

## 2. Materials and Methods

*2.1. Addictive Drugs and Drug Targets.* In this study, we define addictive drugs as those abused drugs and prescribed drugs that can cause addiction disease once they are abused by humans. We manually obtained a list of the abused drugs from the Commonly Abused Drugs Chart and the prescribed drugs from the Prescription Drugs Abuse Chart created by the National Institute on Drug Abuse (NIDA) (http://drugabuse.gov/). The two charts contain addictive drugs along with their common and street names. These addictive drugs could be grouped into six categories according to similarities between how they work and what effects they produce in the human body, especially in the brain. These six categories are depressants, dissociative anesthetics, hallucinogens, opioids and morphine derivatives, stimulants, and other compounds.

We extracted the drug target data from DrugBank, a publically available database [8]. DrugBank includes 6712 drugs and 150 corresponding data fields for each drug. To match the addictive drugs collected from NIDA to DrugBank, we first manually searched the DrugBank website (http://www.drugbank.ca/) by using the drugs' common names and then collected their DrugBank accession numbers. The "Accession Number" is the unique DrugBank ID consisting of a two-letter prefix (DB) and a five-digit suffix. Next, we obtained their targets and the non-addictive drugs that share at least one target with at least one addictive drug from the DrugBank XML file (version 3.0) downloaded in July, 2013. We extracted the corresponding data from the following fields: "Name," "Groups," and "Targets." The "Name" field includes the standard name of a drug as provided by the drug manufacturer. The "Groups" field represents the legal status of a drug such as "Approved," "Experimental," "Nutraceutical," "Illicit," and "Withdrawn" (detailed information can be found on the DrugBank website). The "Targets" field contains drug targets to which one drug can bind, including proteins, macromolecules, nucleic acids, or small molecules. In this study, we primarily extracted human proteins with UniProtKB identifiers and then mapped them to Entrez gene symbols and gene IDs using the UniProt ID mapping service (http://www.uniprot.org/mapping/).

*2.2. Drug ATC Classification.* To systematically examine drug classifications of addictive and non-addictive drugs, we further employed the Anatomical Therapeutic Chemical (ATC) classification (http://www.whocc.no/atc_ddd_index/). The classification system categorizes active drugs into five different levels based on the organ or system on which they act as well as their therapeutic and chemical characteristics. For each drug, the ATC classification information was extracted from the DrugBank XML file (version 3.0) or the Kyoto Encyclopedia of Genes and Genomes (KEGG) DRUG "htext" file, which was downloaded from KEGG Anatomical Therapeutic Chemical (ATC) classification website (http://www.genome.jp/kegg-bin/get_htext?br08303.keg) in July, 2013.

*2.3. Functional Analysis of Targets.* To characterize the functionality of those addictive drugs' targets, we performed an enrichment analysis of KEGG canonical pathways using the online tool Web-Based Gene Set Analysis Toolkit (WebGestalt) [16]. After the genes of interest were input into the WebGestalt system, it mapped the genes to the KEGG annotation and performed hypergeometric tests. To reduce the type I error, we conducted the Benjamini-Hochberg correction for multiple testings [17]. Using this approach, we calculated the adjusted $P$ values to assess the overrepresentation of these input genes in each biological pathway. Here, we selected the pathways with adjusted $P$-values of less than 0.01 as the significantly enriched pathways. To further ensure a biologically meaningful analysis, we considered only those KEGG pathways that contained at least five target genes [18].

*2.4. Network Construction, Visualization, and Analyses.* We constructed two addiction-related networks. The first one is the addictive drug-target network, in which the nodes represent addictive drugs or their targets and edges represent the associations between these drugs and targets. The second network is an expanded addictive drug-target network, in which nodes include addictive drugs, their targets, and non-addictive drugs that have at least one target in common with addictive drugs. We employed the software Cytoscape (version 3.01) [19] to visualize and analyze the networks.

Considering that the nodes that act as hubs or bridging nodes in a network might play critical roles in drug actions [11, 20], we performed degree and betweenness analyses to determine the hubs and bridge nodes. In the network, a node with a higher degree (number of edges linked to the node) is defined as a hub. Hubs play important roles in biological networks because they tend to be encoded by essential genes [21]. In this study, we determine hubs by plotting the degree distribution, adopting the methods described by Yu et al. [22]. We defined the degree value as the cutoff where the distribution begins to straighten out. For bridging nodes, we calculated the betweenness centrality using algorithms implemented in the Cytoscape plugin NetworkAnalyzer [23] and then drew the betweenness distribution to define the point that the distribution began to reach its asymptote.

*2.5. Literature Search.* To evaluate the prediction of non-addictive drugs' associations with addiction, we adopted the NCBI PubMed automatic term mapping strategy to examine whether a drug and an addiction-related keyword cooccur in the same PubMed document [24]. The addiction-related keywords included "addiction," "addictive," "abuse," and "abused." The total number of abstract records in the 2012 PubMed was 21,508,439 (http:// www.nlm.nih.gov/bsd/authors1.html). For each drug, we obtained three numbers corresponding to three subsets of PubMed abstracts: the number of abstracts with the given drug name, the number of abstracts with at least one addiction-related keyword, and the unique number of abstracts with a co-occurrence of the drug name and at least one of the addiction-related keywords. Then, we performed the Fisher's exact test based on these numbers for each drug. To identify and determine the predicted non-addictive drugs that were more significantly associated with addiction study than expected, we required that the drugs have a $P$ value of less than 0.05 after Bonferroni multiple testing correction [17].

## 3. Results

*3.1. Addictive Drugs and Their Targets.* This study included 44 compounds listed as addictive drugs by NIDA. We extracted their target information from the DrugBank database. Among them, 39 belonged to the approved drugs category in at least one country, 22 were illicit drugs that were scheduled in at least one country, three were withdrawn drugs, and three were experimental drugs. According to similarities regarding how they function and what effects they produce in the human body and brain, as annotated by the NIDA, these drugs could be grouped into six categories: depressants (12), dissociative anesthetics (2), hallucinogens (1), opioids and morphine derivatives (10), stimulants (6), and other compounds (13). Table 1 summarizes the detailed information for each drug.

According to ATC system classification, 32 drugs belonged to "nervous system," four to "respiratory system," three to "alimentary tract and metabolism," and two to "sensory organs." This observation confirmed that almost all of the addictive drugs perform their actions by affecting brain function.

Among the 44 addictive drugs, 41 had at least one target gene. After deleting redundancy and mapping gene names to NCBI gene annotations (http://www.ncbi.nlm.nih.gov/gene), we obtained 91 target genes (additional file, Table S1 in supplementary material available online at http://10.1155/2014/258784). To examine the pathways in which those target genes involve, we conducted a KEGG pathway enrichment analysis using the online tool WebGestalt. Nine pathways were significantly enriched with the 91 addictive drug target genes (adjusted $P$-value $< 0.01$) (Table 2). Among them, the most significant one is "neuroactive ligand-receptor interaction," which includes more than half of the addictive drug target genes (61.54%). This pathway finding is consistent with the molecular mechanisms underlying the addiction [25].

*3.2. Addictive Drug-Target Network.* According to the relationship between addictive drugs and their targets, we first generated an addictive drug-target interaction network, which provided general insights into the organization and association between addictive drugs and their targets. Through finding interesting features from this network, we aimed to prove the value of the network application concept when investigating drug repurposing. In this network, an addictive drug connects to a target (i.e., an edge) if the target is a known target of the drug. The addictive drug-target network contained 132 nodes (41 addictive drugs and 91 target genes) and 297 edges. After superimposing the drug categories onto the network, five clusters were observed, which corresponded to five major drug categories: depressants, stimulants, dissociative anesthetics, opioids and morphine derivatives, and other compounds (Figure 1). Interestingly, there are several bridging nodes that link the major subnetworks together. These bridging nodes are GABRA1, SLC6A4, GRIN3A, CHRNA2, CHRNA4, and CHRNA7.

*3.3. Expanded Addictive Drug-Target Interaction Network.* Drugs sharing the same targets might participate in the same pathways and have similar actions. Thus, an investigation of the drugs that share the same targets with addictive drugs might provide information for further addiction treatment. Here, we added these non-addictive drugs to the addictive drug-target network to construct an expanded addictive drug-target interaction network. The network contained 705 nodes and 1797 edges. These 705 nodes included 41 addictive

Table 1: Summary of addictive drugs, their targets, and classification.

| DrugBank ID | Drug name | Number of targets | DrugBank group | NIDA category[a] |
|---|---|---|---|---|
| DB00316 | Acetaminophen | 2 | Approved | Opioids and morphine derivatives |
| DB00404 | Alprazolam | 20 | Approved, illicit | Depressants |
| DB01351 | Amobarbital | 10 | Approved, illicit | Depressants |
| DB00182 | Amphetamine | 4 | Approved, illicit | Stimulants |
| DB01541 | Boldenone | 1 | Experimental, illicit | Other compounds |
| DB00475 | Chlordiazepoxide | 19 | Approved, illicit | Depressants |
| DB00907 | Cocaine | 8 | Approved, illicit | Stimulants |
| DB00318 | Codeine | 3 | Approved, illicit | Opioids and morphine derivatives |
| DB01189 | Desflurane | 7 | Approved | Other compounds |
| DB00514 | Dextromethorphan | 4 | Approved | Other compounds |
| DB00829 | Diazepam | 18 | Approved, illicit | Depressants |
| DB00228 | Enflurane | 8 | Approved | Other compounds |
| DB00898 | Ethanol | 7 | Approved | Other compounds |
| DB00813 | Fentanyl | 3 | Approved, illicit | Opioids and morphine derivatives |
| DB01544 | Flunitrazepam | 6 | Approved, illicit | Depressants |
| DB01440 | Gamma hydroxybutyric acid | 1 | Approved, illicit | Depressants |
| DB01159 | Halothane | 17 | Approved | Other compounds |
| DB01452 | Heroin | 3 | Approved, illicit | Opioids and morphine derivatives |
| DB00956 | Hydrocodone | 2 | Approved, illicit | Opioids and morphine derivatives |
| DB00327 | Hydromorphone | 3 | Approved, illicit | Opioids and morphine derivatives |
| DB00753 | Isoflurane | 7 | Approved | Other compounds |
| DB01221 | Ketamine | 3 | Approved | Dissociative anesthetics |
| DB00186 | Lorazepam | 20 | Approved | Depressants |
| DB04829 | Lysergic acid diethylamide | 0 | Illicit, withdrawn | Hallucinogens |
| DB00454 | Meperidine | 6 | Approved | Opioids and morphine derivatives |
| DB01577 | Methamphetamine | 11 | Approved, illicit | Stimulants |
| DB04833 | Methaqualone | 0 | Illicit, withdrawn | Depressants |
| DB01028 | Methoxyflurane | 7 | Approved | Other compounds |
| DB00422 | Methylphenidate | 3 | Approved, investigational | Stimulants |
| DB01442 | MMDA | 8 | Experimental, illicit | Stimulants |
| DB00295 | Morphine | 3 | Approved | Opioids and morphine derivatives |
| DB00486 | Nabilone | 2 | Approved | Other compounds |
| DB00984 | Nandrolone phenpropionate | 0 | Approved, illicit | Other compounds |
| DB00184 | Nicotine | 11 | Approved | Stimulants |
| DB00621 | Oxandrolone | 1 | Approved | Other compounds |
| DB00497 | Oxycodone | 3 | Approved, illicit | Opioids and morphine derivatives |
| DB00312 | Pentobarbital | 10 | Approved | Depressants |
| DB03575 | Phencyclidine | 2 | Experimental, illicit | Dissociative anesthetics |
| DB01174 | Phenobarbital | 10 | Approved | Depressants |
| DB00647 | Propoxyphene | 3 | Approved, illicit | Opioids and morphine derivatives |
| DB00418 | Secobarbital | 10 | Approved | Depressants |
| DB01236 | Sevoflurane | 7 | Approved | Other compounds |
| DB00624 | Testosterone | 1 | Approved | Other compounds |
| DB00897 | Triazolam | 20 | Approved, illicit, withdrawn | Depressants |

[a]Drug category is defined based on the similarities regarding how drugs function and what effects they produce in the human body, including the brain, as annotated by NIDA.

drugs, 573 non-addictive drugs, and 91 targets. The edges contained 297 interactions between addictive drugs and their targets and 1500 interactions between non-addictive drugs and addictive drug targets.

Among these 573 non-addictive drugs, 407 had at least one ATC classification distributed among all 14 categories. Among them, the percentage of addictive and non-addictive drugs was significantly different in the category of "nervous

Table 2: KEGG pathways significantly enriched with the target genes of addictive drugs.

| Pathway name | Number of target genes (%) | Nominal $P$ value[a] | Adjusted $P$ value[b] |
|---|---|---|---|
| Neuroactive ligand-receptor interaction | 56 (61.54) | $8.65 \times 10^{-102}$ | $7.78 \times 10^{-101}$ |
| Long-term potentiation | 10 (10.99) | $3.35 \times 10^{-16}$ | $1.51 \times 10^{-15}$ |
| Calcium signaling pathway | 12 (13.19) | $3.25 \times 10^{-15}$ | $9.75 \times 10^{-15}$ |
| Amyotrophic lateral sclerosis (ALS) | 7 (7.69) | $1.94 \times 10^{-11}$ | $4.36 \times 10^{-11}$ |
| Alzheimer's disease | 9 (9.89) | $8.50 \times 10^{-11}$ | $1.53 \times 10^{-10}$ |
| Tyrosine metabolism | 5 (5.49) | $2.50 \times 10^{-8}$ | $3.75 \times 10^{-8}$ |
| Drug metabolism-cytochrome P450 | 5 (5.49) | $4.75 \times 10^{-7}$ | $6.11 \times 10^{-7}$ |
| Salivary secretion | 5 (5.49) | $1.28 \times 10^{-6}$ | $1.44 \times 10^{-6}$ |
| Metabolic pathways | 8 (8.79) | $2.50 \times 10^{-3}$ | $2.50 \times 10^{-3}$ |

[a]Nominal $P$ values were calculated using the hypergeometric test.
[b]Adjusted $P$ values were estimated by Benjamini-Hochberg (1995) multiple testing corrections [17].

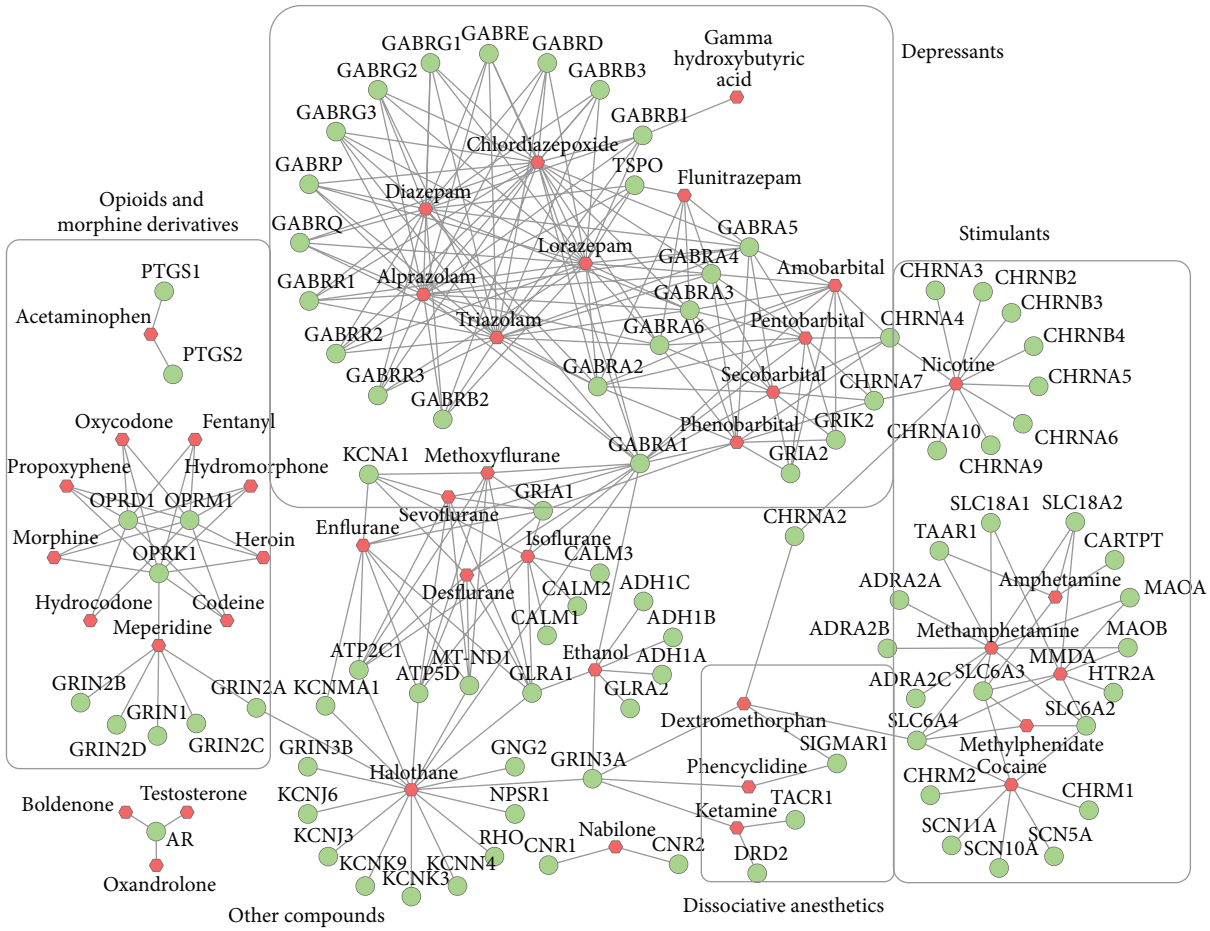



Figure 1: The addictive drug-target network. The red nodes denote addictive drugs and the green nodes denote their targets. The edges indicate the relationship between each drug and its targets. Subnetworks are highlighted to differentiate five drug categories: depressants, stimulants, opioids and morphine derivatives, dissociative anesthetics, and other compounds.

system" (N) (Fisher's exact test, $P$-value: $2.56 \times 10^{-5}$). Though almost half of the non-addictive drugs (203/407, 49.88%) belong to "nervous system," this proportion is significantly lower than that of addictive drugs involved in the expanded network (32/38, 84.21%; $P$-value: $3.00 \times 10^{-5}$). The difference in the category "nervous system" was expected since almost all addictive drugs function through the brain.

In the network, the average drug degree (number of targets) was 2.9 with a range between 1 and 20, while the average target degree (number of drugs) was 19.5 with a range

between 1 and 73. In the network, the target degree was oversaturated compared to the drug degree, which was mainly caused by the approach used to generate this network. The distribution of drug degrees followed a power law, but the distribution for target degrees did not have this feature (Figure 2). Thus, to identify the drugs related to addiction, we calculated the drug degree distribution to determine the drug hubs in the network. As shown in Figure 2, the nodes with degrees greater than three were defined as hubs. Similarly, we defined each node with a betweenness centrality greater than 0.04 as a bridging node (data not shown). After retaining the hub nodes and bridging nodes, a subnetwork was extracted from the expanded addictive drug-target network. The subnetwork contained 193 nodes (25 addictive drugs, 94 non-addictive drugs, and 74 targets) and 1002 edges (Figure 3). As a result, we identified 94 drugs that either have a high potential for having addictive effects or could be used as a potential treatment for addiction. The degree and betweenness values of these 94 drugs were provided in additional file, Table S2.

*3.4. Evaluation of Predicted Non-Addictive Drugs for Addiction.* To evaluate the association between these 94 non-addictive drugs and addiction, we examined the co-occurrence of each drug and addiction-related keywords including "addiction," "addictive," "abuse," and "abused" in PubMed abstracts. Among the 94 drugs, 51 drugs (54.26%) (yellow nodes in Figure 3) had statistically significant $P$-values after Bonferroni correction of multiple testing (Fisher's exact test, $P$-value: $0 \sim 0.0004$) (additional file, Table S2). For example, the drug temazepam, which is a hub with 20 targets shared with addictive drugs, is a highly addictive benzodiazepine medication [26–30]. The drug dronabinol, which is the strongest drug bridge node in the network, has the potential for addiction [31]. It is also a promising medication for the treatment of cannabis dependence [32]. The drug methadone is the fourteenth strongest bridge node and the top one based on the ratio of the observed versus expected number of documents in PubMed. We added more discussion below.

Methadone is the most widely available pharmacotherapy for opioid addiction and it has been shown to be an effective and safe treatment for many years [33, 34]. To illustrate the molecular mechanism of this drug, we generated a methadone-specific network (Figure 4). This network included 74 nodes and 94 edges. The nodes included the drug methadone, its four targets and 12 enzymes from DrugBank, and 67 proteins directly interacting with the four targets and 12 enzymes (Figure 4). The edges included 4 interactions between the drug and four targets, 12 relationship between the drug and 12 enzymes, and 78 protein-protein interactions between targets/enzymes and other proteins, which were extracted from the protein interaction network analysis (PINA) database [35]. According to the KEGG pathway annotations, all four targets (OPRM1, GRIN3A, CHRNA10, and OPRD1) are neuroactive ligand receptors. Among the 12 enzymes, ten are directly involved in drug metabolism. There are 20 KEGG pathways that were significantly enriched in the 67 proteins. Among them, seven pathways are directly involved in the neurodevelopment, including "Long-term potentiation" (7 proteins: CALM2, CALM3, GRIN2A, PRKCA, CALM1, GRIN2B, and GRIN1; $P$-value: $1.75\times10^{-11}$), "Long-term depression" (5 proteins: PRKCA, GNAI2, GNAZ, GNAO1, and GNAI1; $P$-value: $8.70\times10^{-8}$), and "Neuroactive ligand-receptor interaction" (5 proteins: ADRB2, OPRK1, GRIN2A, GRIN2B, and GRIN1; $P$-value: $6.75\times10^{-5}$). These observations confirmed that methadone directly acts with neurotransmitters and further regulates the other molecular components in neurodevelopment.

Put together, the drug pool through our network analyses might provide a list of candidate drugs for further investigation of their potential for addiction treatment or addiction risk.

## 4. Discussion

In this study, we investigated the relationships between addictive drugs, their targets, and non-addictive drugs that have targets in common with addictive drugs in the context of drug-target networks. Most of the addictive drugs with similar functions could cluster together in their drug-target network (Figure 1), indicating that network-assisted approaches could effectively capture drug classification characteristics. After studying the network topological characteristics, we predicted some drugs that might have the potential leading to addictive effects or to addiction treatment. These results illustrate that the network pharmacology approach is promising for drug repositioning [36, 37]. Therefore, the strategy employed for building the basic and the expanded networks in this study is effective and straightforward, offering a promising computational method to predict potential drugs for a given disease. Furthermore, this study proves the concept that such a network approach can be implemented in predicting drug-target relationships and uncovering novel drugs/targets for both basic and clinical research.

We mainly extracted the drugs and their targets from DrugBank. Though the study provides some promising results, future improvement is needed. One limitation of this study is that the current data is neither complete nor bias-free. In future, we will include more drug-target information from multiple data sources such as the binding database (binding DB) [38], therapeutic targets database (TTD) [39], and other drug-target centered databases. We also expect that data quality and annotations of drug-target interactions will be substantially improved in the near future due to numerous ongoing efforts in this research area.

In our previous study, we explored the relationship between illicit drugs and their targets [15]. Illicit drugs are those drugs that are annotated as illicit in at least one country according to DrugBank annotation. Some illicit drugs could lead to addiction once they are abused by humans. However, only the drugs that could lead to addiction are referred to as abused drugs by NIDA. In this study, we mainly focused on the 44 drugs that lead to addiction, of which only 20 belong to the illicit drugs category.

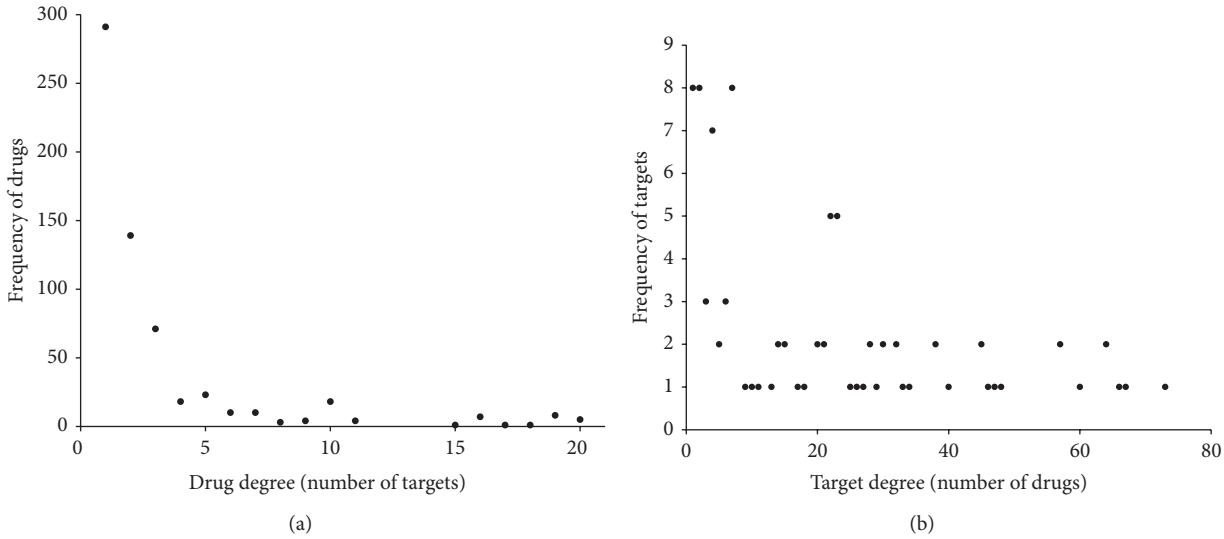



Figure 2: Degree distribution of drugs (a) and targets (b) in the expanded addictive drug-target network.

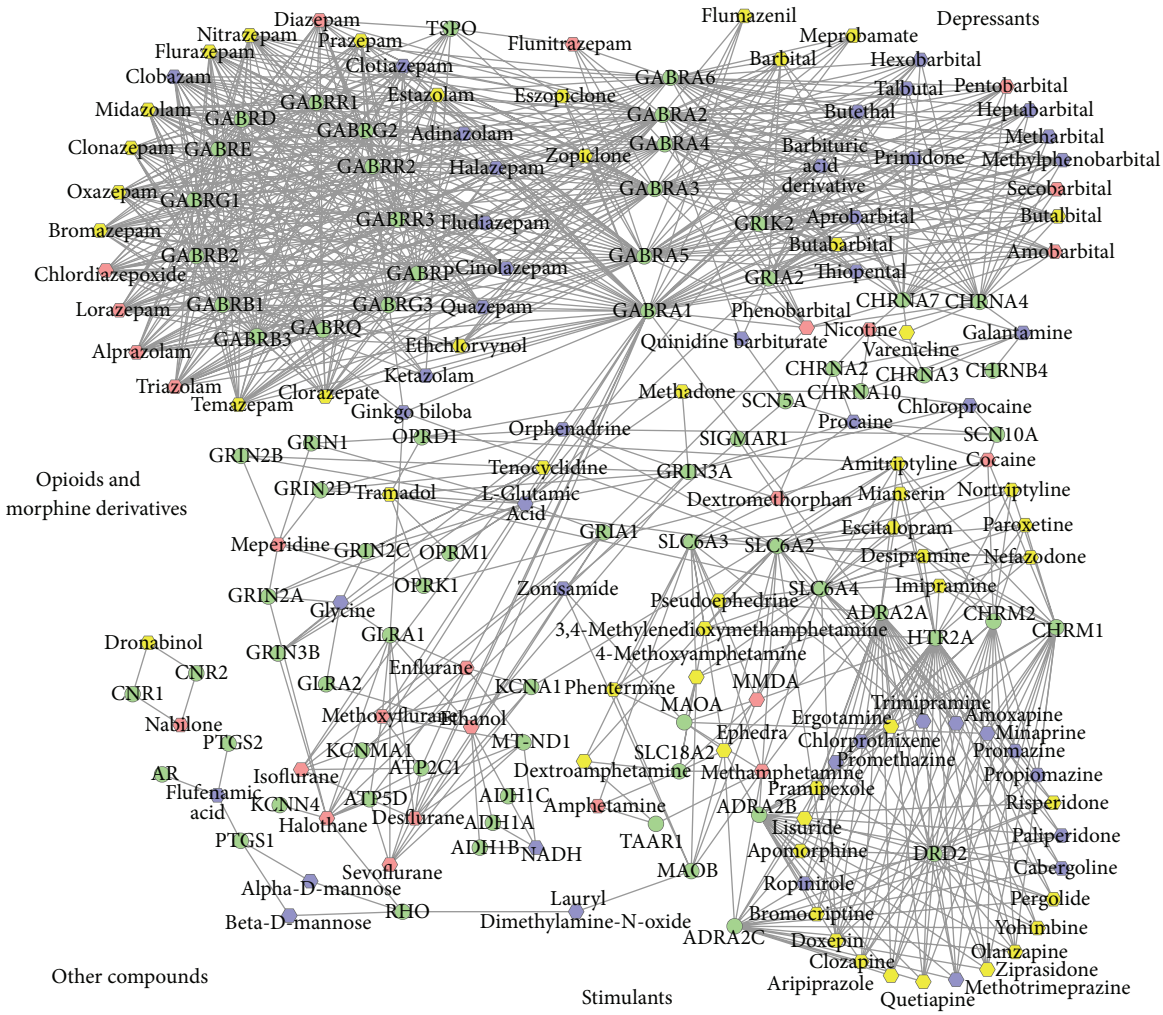



Figure 3: The expanded addictive drug-target network after filtering by hubs and bridge nodes. The red nodes denote addictive drugs, the green nodes denote targets, the yellow nodes denote nonaddictive drugs that have a significantly higher cooccurrence than expected with addiction-related keywords in the literature from PubMed, and the blue nodes denote non-addictive drugs that have a co-occurrence with addiction keywords but are not significantly higher than expected in the literature data.

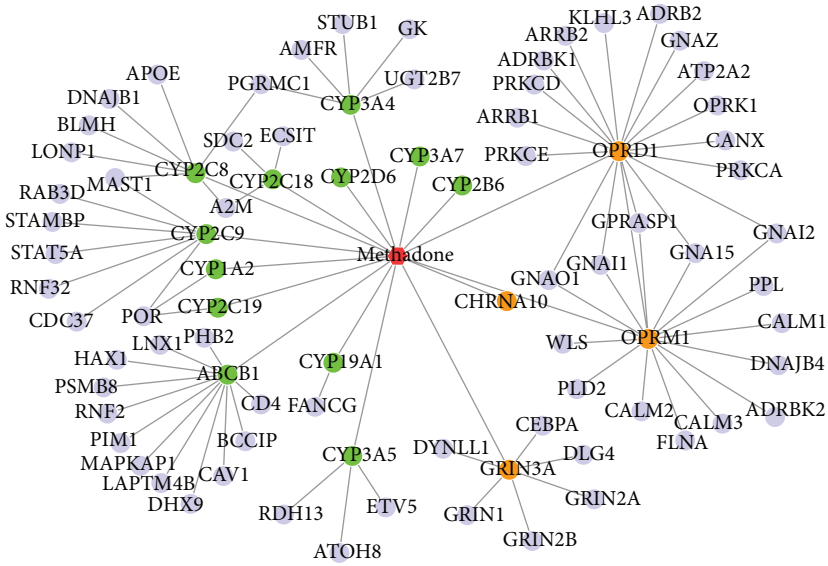



FIGURE 4: The methadone drug-target network. The red nodes denote the methadone (drug), the orange nodes denote its targets, the green nodes denote its enzymes, and the blue nodes denote the directly interacting proteins of targets and enzymes from protein-protein interaction network.

In this study, we predicted 94 non-addictive drugs that might have associations with addiction. To explore if some of these drugs have been studied with addiction, we used a keyword-based literature search followed by a co-occurrence analysis. The literature search approach we utilized largely relied on the co-occurrence of addictive drugs and addiction-related keywords in the PubMed database. The high throughput literature search revealed that more than half (54.26%) of non-addictive drugs have been previously investigated or reported as linked to addiction. However, the current literature survey method did not allow us to examine the logical relationship between these drugs and addiction. Thus, we could not filter those negative studies based on negative logical relationship information in abstracts. In the future, we may improve our strategy for searching the co-occurrence of drugs and keywords by creating a more efficient algorithm using natural language processing techniques.

## Conflict of Interests

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

## Acknowledgments

The authors thank Ms. Rebecca Hiller Posey for critically reading an early version of this paper. This work was partially supported by The Brain and Behavior Research Foundation under a 2010 NARSAD Young Investigator Award to JS.

## References

[1] M. J. Kreek, "Drug addictions molecular and cellular endpoints," *Annals of the New York Academy of Sciences*, vol. 937, pp. 27–49, 2001.

[2] J. C. Crabbe, "Genetic contributions to addiction," *Annual Review of Psychology*, vol. 53, pp. 435–462, 2002.

[3] G. R. Uhl, "Molecular genetic underpinnings of human substance abuse vulnerability: likely contributions to understanding addiction as a mnemonic process," *Neuropharmacology*, vol. 47, supplement 1, pp. 140–147, 2004.

[4] C.-Y. Li, W.-Z. Zhou, P.-W. Zhang, C. Johnson, L. Wei, and G. R. Uhl, "Meta-analysis and genome-wide interpretation of genetic susceptibility to drug addiction," *BMC Genomics*, vol. 12, article 508, 2011.

[5] A. J. Robison and E. J. Nestler, "Transcriptional and epigenetic mechanisms of addiction," *Nature Reviews Neuroscience*, vol. 12, no. 11, pp. 623–637, 2011.

[6] A. I. Leshner, "Science-based views of drug addiction and its treatment," *Journal of the American Medical Association*, vol. 282, no. 14, pp. 1314–1316, 1999.

[7] J. Knowles and G. Gromo, "A guide to drug discovery: target selection in drug discovery," *Nature Reviews Drug Discovery*, vol. 2, no. 1, pp. 63–69, 2003.

[8] C. Knox, V. Law, T. Jewison et al., "DrugBank 3.0: a comprehensive resource for "Omics" research on drugs," *Nucleic Acids Research*, vol. 39, no. 1, pp. D1035–D1041, 2011.

[9] J. Sun, Y. Wu, H. Xu, and Z. Zhao, "DTome: a web-based tool for drug-target interactome construction," *BMC Bioinformatics*, vol. 13, supplement 9, 2012.

[10] H. Li, Z. Gao, L. Kang et al., "TarFisDock: a web server for identifying drug targets with docking approach," *Nucleic Acids Research*, vol. 34, pp. W219–W224, 2006.

[11] A. L. Hopkins, "Network pharmacology: the next paradigm in drug discovery," *Nature Chemical Biology*, vol. 4, no. 11, pp. 682–690, 2008.

[12] D. K. Arrell and A. Terzic, "Network systems biology for drug discovery," *Clinical Pharmacology and Therapeutics*, vol. 88, no. 1, pp. 120–125, 2010.

[13] J. Sun, H. Xu, and Z. Zhao, "Network-assisted investigation of antipsychotic drugs and their targets," *Chemistry Abd Biodiversity*, vol. 9, pp. 900–910, 2012.

[14] J. Sun, M. Zhao, A. H. Fanous, and Z. Zhao, "Characterization of schizophrenia adverse drug interactions through a network approach and drug classification," *BioMed Research International*, vol. 2013, Article ID 458989, 10 pages, 2013.

[15] R. Atreya, J. Sun, and Z. Zhao, "Exploring drug-target interaction networks of illicit drugs," *BMC Genomics*, vol. 14, article S1, supplement 4, 2013.

[16] J. Wang, D. Duncan, Z. Shi, and B. Zhang, "WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013," *Nucleic Acids Research*, vol. 41, pp. W77–W83, 2013.

[17] 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*, vol. 57, pp. 289–300, 1995.

[18] J. Xu, J. Sun, J. Chen et al., "RNA-Seq analysis implicates dysregulation of the immune system in schizophrenia," *BMC Genomics*, vol. 13, article S2, supplement 8, 2012.

[19] 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.

[20] W.-C. Hwang, A. Zhang, and M. Ramanathan, "Identification of information flow-modulating drug targets: a novel bridging paradigm for drug discovery," *Clinical Pharmacology and Therapeutics*, vol. 84, no. 5, pp. 563–572, 2008.

[21] C. Boone, H. Bussey, and B. J. Andrews, "Exploring genetic interactions and networks with yeast," *Nature Reviews Genetics*, vol. 8, no. 6, pp. 437–449, 2007.

[22] H. Yu, D. Greenbaum, H. X. Lu, X. Zhu, and M. Gerstein, "Genomic analysis of essentiality within protein networks," *Trends in Genetics*, vol. 20, no. 6, pp. 227–231, 2004.

[23] Y. Assenov, F. Ramírez, S.-E. S.-E. Schelhorn, T. Lengauer, and M. Albrecht, "Computing topological parameters of biological networks," *Bioinformatics*, vol. 24, no. 2, pp. 282–284, 2008.

[24] J. Sun, P. Jia, A. H. Fanous et al., "A multi-dimensional evidence-based candidate gene prioritization approach for complex diseases-schizophrenia as a case," *Bioinformatics*, vol. 25, no. 19, pp. 2595–2602, 2009.

[25] A. Agrawal, K. J. Verweij, N. A. Gillespie et al., "The genetics of addiction-a translational perspective," *Transl Psychiatry*, vol. 2, Article ID e140, 2012.

[26] D. T. Mansie, "Abuse of temazepam capsules," *Australian Family Physician*, vol. 24, no. 11, p. 2146, 1995.

[27] G. E. Ralston and J. A. Taylor, "Temazepam abuse," *Addiction*, vol. 88, no. 3, p. 423, 1993.

[28] R. Hammersley, T. Lavelle, and A. Forsyth, "Buprenorphine and temazepam: abuse," *British Journal of Addiction*, vol. 85, no. 2, pp. 301–303, 1990.

[29] M. S. Sakol, C. Stark, and R. Sykes, "Buprenorphine and temazepam abuse by drug takers in Glasgow: an increase," *British Journal of Addiction*, vol. 84, no. 4, pp. 439–411, 1989.

[30] C. Stark, R. Sykes, and P. Mullin, "Temazepam abuse," *The Lancet*, vol. 2, no. 8562, pp. 802–803, 1987.

[31] S. R. Calhoun, "Abuse potential of dronabinol (Marinol)," *Journal of Psychoactive Drugs*, vol. 30, no. 2, pp. 187–195, 1998.

[32] F. R. Levin, J. J. Mariani, D. J. Brooks, M. Pavlicova, W. Cheng, and E. V. Nunes, "Dronabinol for the treatment of cannabis dependence: a randomized, double-blind, placebo-controlled trial," *Drug and Alcohol Dependence*, vol. 116, no. 1–3, pp. 142–150, 2011.

[33] M. J. Kreek, L. Borg, E. Ducat, and B. Ray, "Pharmacotherapy in the treatment of addiction: methadone," *Journal of Addictive Diseases*, vol. 29, no. 2, pp. 209–216, 2010.

[34] P. A. M. Meulenbeek, "Addiction problems and methadone treatment," *Journal of Substance Abuse Treatment*, vol. 19, no. 2, pp. 171–174, 2000.

[35] M. J. Cowley, M. Pinese, K. S. Kassahn et al., "PINA v2. 0: mining interactome modules," *Nucleic Acids Research*, vol. 40, pp. D862–D865, 2012.

[36] D. Emig, A. Ivliev, O. Pustovalova et al., "Drug target prediction and repositioning using an integrated network-based approach," *PLoS ONE*, vol. 8, Article ID e60618, 2013.

[37] Z. Wu, Y. Wang, and L. Chen, "Network-based drug repositioning," *Molecular BioSystems*, vol. 9, pp. 1268–1281, 2013.

[38] T. Liu, Y. Lin, X. Wen, R. N. Jorissen, and M. K. Gilson, "BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities," *Nucleic Acids Research*, vol. 35, no. 1, pp. D198–D201, 2007.

[39] F. Zhu, Z. Shi, C. Qin et al., "Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery," *Nucleic Acids Research*, vol. 40, pp. D1128–D1136, 2012.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 725780, 10 pages
http://dx.doi.org/10.1155/2014/725780

*Research Article*

# Computational Analysis of Transcriptional Circuitries in Human Embryonic Stem Cells Reveals Multiple and Independent Networks

**Xiaosheng Wang[1] and Chittibabu Guda[1,2]**

[1] *Department of Genetics, Cell Biology and Anatomy, University of Nebraska Medical Center, Omaha, NE 68198-5805, USA*
[2] *Bioinformatics and Systems Biology Core, University of Nebraska Medical Center, Omaha, NE 68198-5805, USA*

Correspondence should be addressed to Chittibabu Guda; babu.guda@unmc.edu

Received 22 September 2013; Revised 12 November 2013; Accepted 17 November 2013; Published 9 January 2014

Academic Editor: Zhongming Zhao



It has been known that three core transcription factors (TFs), NANOG, OCT4, and SOX2, collaborate to form a transcriptional circuitry to regulate pluripotency and self-renewal of human embryonic stem (ES) cells. Similarly, MYC also plays an important role in regulating pluripotency and self-renewal of human ES cells. However, the precise mechanism by which the transcriptional regulatory networks control the activity of ES cells remains unclear. In this study, we reanalyzed an extended core network, which includes the set of genes that are cobound by the three core TFs and additional TFs that also bind to these cobound genes. Our results show that beyond the core transcriptional network, additional transcriptional networks are potentially important in the regulation of the fate of human ES cells. Several gene families that encode TFs play a key role in the transcriptional circuitry of ES cells. We also demonstrate that MYC acts independently of the core module in the regulation of the fate of human ES cells, consistent with the established argument. We find that TP53 is a key connecting molecule between the core-centered and MYC-centered modules. This study provides additional insights into the underlying regulatory mechanisms involved in the fate determination of human ES cells.

## 1. Introduction

Pluripotency and self-renewal are two defining properties of embryonic stem (ES) cells. Pluripotency is the capacity to generate all cell types, while self-renewal is the capacity to maintain ES cells in a proliferative state for prolonged periods [1]. It has been of great interest to know how the ES cells balance the two statuses of pluripotency and self-renewal. It has been found that the three core transcription factors (TFs) NANOG, OCT4, and SOX2 collaborate to regulate pluripotency and self-renewal of human ES cells in the form of a regulatory circuitry [2]. NANOG is a gene expressed in ES cells, which plays a key role in maintaining the pluripotency of ES cells. Downregulation of NANOG will result in differentiation, while expression will block differentiation of ES cells. OCT4, also known as POU5F1, is a gene encoding the protein that is critically involved in the self-renewal of undifferentiated ES cells. OCT4 expression level must be within a certain range to maintain the undifferentiated status of ES cells. SOX2 gene encodes a member of the SRY-related HMG-box (SOX) family of TFs involved in the regulation of embryonic development and in the determination of cell fate. It plays a critical role in the maintenance of embryonic and neural stem cells. SOX2 has been shown to interact with PAX6 [3], NPM1 [4], and OCT4 [5] and cooperatively regulate REX1 with OCT3/4 [6].

Boyer et al. have identified the bound genes of the three core TFs *in vivo* by genome-scale location analysis [2]. They found that OCT4 is associated with 623 (3%) promoter regions of the known protein-coding genes in human ES cells, while SOX2 and NANOG are associated with 1271 (7%) and 1687 (9%) genes, respectively. Further, they identified a set of 353 genes (Table S1; see Supplementary Material available online at http://dx.doi.org/10.1155/2014/725780) that are cobound by all the three TFs in human ES cells and found that this set includes a substantial number of genes that encode

homeodomain TFs, which are important in developmental regulation of ES cells. These discoveries suggested that the three TFs function together to control pluripotency and self-renewal of human ES cells. Hereafter, we refer to the set of 353 genes as the core-bound genes.

MYC is another important transcriptional regulator in ES cells, which is involved in somatic cell reprogramming and cancer [7]. Takahashi and Yamanaka generated induced pluripotent stem cells (iPSCs) by forced expression of four transcriptional factors (OCT3/4, SOX2, KLF4, and MYC) in mouse embryonic and adult fibroblast cultures [8] and later in adult human dermal fibroblasts [9]. These studies indicate that MYC also plays a key role in controlling pluripotency and self-renewal of ES cells, although it may act in a distinct way from the core module [1, 7, 10]. However, the precise mechanism by which the transcriptional regulatory networks control the activity of ES cells remains unclear. It is likely that the transcriptional circuitry of ES cells is regulated by multiple core TFs using independent networks, to regulate self-renewal and differentiation of human ES cells.

In this study, we reanalyzed the core-bound genes using Ingenuity Pathway Analysis tool (IPA, Ingenuity Systems, http://www.ingenuity.com/) and the gene set enrichment analysis (GSEA) software [11]. Important networks, biological functions, and pathways associated with the gene sets were annotated. We induced the TFs that bind to the subsets of the core-bound genes with DAVID tool [12, 13] and analyzed the transcriptional network based on the induced TFs. In addition, we compared the regulatory targets of MYC with the core-bound genes and also the MYC-centered and core-centered regulatory modules to determine if these regulatory circuits operate independently or collaboratively.

## 2. Materials and Methods

We obtained the set of 353 genes that are cobound by NANOG, OCT4, and SOX2 in human ES cells from Boyer et al. [2]. We downloaded the 189 TFs which have been experimentally verified to contribute to transcriptional regulation in human ES cells from the literature [14]. The MYC targeted gene lists in human ES cells were obtained from the literature [15]. The gene lists for the core module and the MYC module in ES cells were downloaded from Kim et al. [7].

We inferred significant networks, biological functions, and pathways associated with gene sets using the core analysis tool in IPA (Ingenuity Systems, http://www.ingenuity.com/). IPA is a system that yields a set of networks relevant to a list of genes based on the preserved records contained in the Ingenuity Pathways Knowledge Base (IPKB). For the input of a gene set into IPA, its core analysis tool will map the gene list to the IPKB and then algorithmically generate molecular networks, biological functions, and canonical pathways that are most likely relevant to the input gene list. IPA is the primary tool used by us to produce visualized gene regulatory networks for analysis of transcriptional regulatory circuits in human ES cells.

We classified genes into different gene families using the "Investigate Gene Sets Tool" in the molecular signatures database (MSigDB) of the gene set enrichment analysis (GSEA) software [11]. We induced the TFs that bind to subsets of a given gene list using the "Functional Annotation Tool" in DAVID [12, 13]. DAVID provides a category called "UCSC_TFBS" in the "Protein_Interactions" option of the functional annotation tool. For an input gene list, DAVID analysis will output a list of TFs that bind subsets of the given gene set. For each identified TF, its binding genes and corresponding $P$ values are provided.

## 3. Results and Discussion

*3.1. Functional Analysis of the Core-Bound Genes.* We first classified the core-bound genes into different gene families using the gene set enrichment analysis (GSEA) software [11]. Table 1 shows that a significant proportion of genes are TF genes (90 of 353), suggesting that the core TFs in turn bind and regulate a large number of other TF genes in the ES cells [2]. The genes encoding homeodomain proteins also have a large proportion in the core-bound genes (34 of 353), all of which encode homeodomain TFs. The homeodomain TFs have been shown to play key roles in fate-determination of ES cells by contributing to the core regulatory networks. It should be noted that there are 11 oncogenes in the core-bound genes, which is indicative of certain similarities between ES and cancer cell transcription programs [7, 14].

Network analysis of the 353 core-bound genes using IPA (Ingenuity Systems, http://www.ingenuity.com/) shows that the top network involves 32 genes among which the three core TFs, NANOG, OCT4, and SOX2, were hub nodes in the network, and formed interconnected autoregulatory and feedforward circuitry (Figure 1). Biological function analysis shows that the core-bound genes are mostly relevant to regulation of gene expression and developmental processes. The developmental processes include nervous system development and function, embryonic development, and organ, organismal, tissue, and cellular development. The six most significant pathways associated with the core-bound genes include transcriptional regulatory network in embryonic stem cells ($P$ value $\approx 10^{-47}$), role of OCT4 in mammalian embryonic stem cell pluripotency ($P$ value $\approx 10^{-8}$), human embryonic stem cell pluripotency ($P$ value $\approx 10^{-7}$), embryonic stem cell differentiation into cardiac lineages ($P$ value $\approx 10^{-5}$), Wnt/$\beta$-catenin signaling ($P$ value $\approx 10^{-4}$), and role of NANOG in mammalian embryonic stem cell pluripotency ($P$ value $\approx 10^{-4}$). These results corroborate the previous findings that the core TFs and the core TF-bound genes are essential for maintaining the pluripotency of ES cells.

*3.2. Identification of Other TFs That Target the Core-Bound Genes.* In addition to the three core TFs, many other TFs

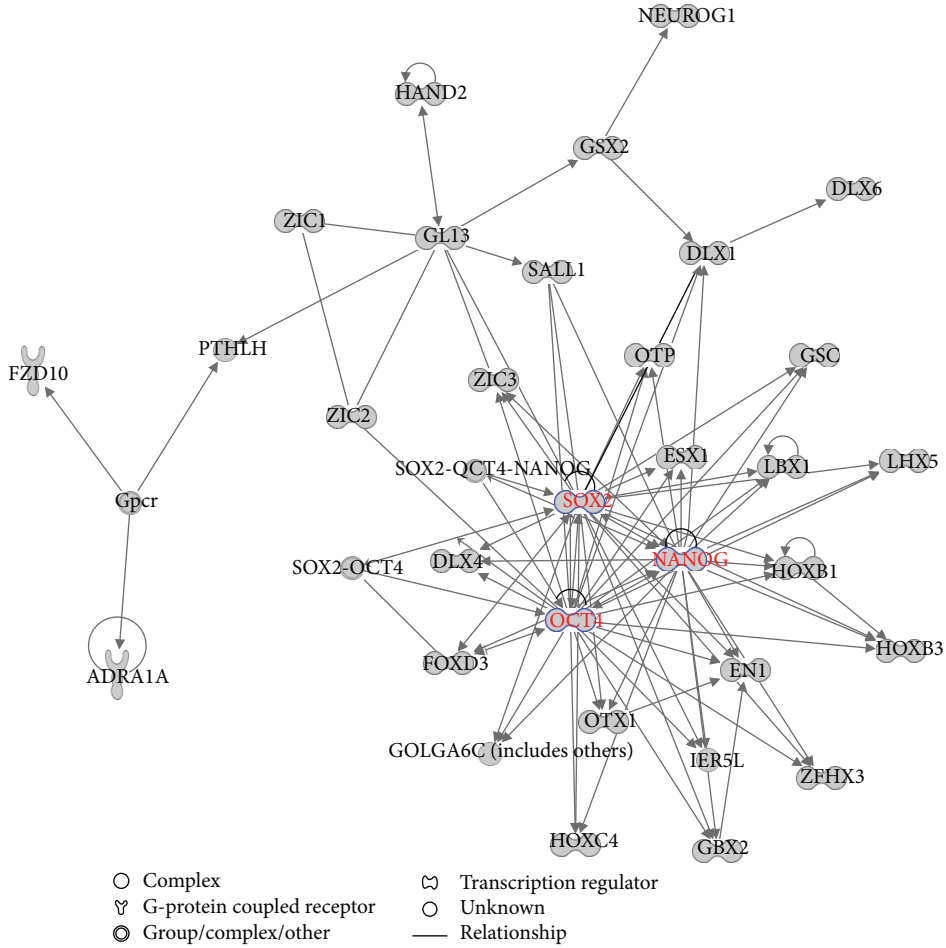



Figure 1: Top network related to the core-bound genes. The three core TFs form hub nodes in the network as highlighted in red color.

Table 1: Category of the core-bound genes.

| | Cytokines and growth factors | Transcription factors | Homeodomain proteins | Cell differentiation markers | Protein kinases | Translocated cancer genes | Oncogenes |
|---|---|---|---|---|---|---|---|
| Cytokines and growth factors | 14 | | | | | | |
| Transcription factors | 0 | 90 | | | | | |
| Homeodomain proteins | 0 | 34 | 34 | | | | |
| Cell differentiation markers | 0 | 0 | 0 | 8 | | | |
| Protein kinases | 0 | 1 | 0 | 3 | 11 | | |
| Translocated cancer genes | 0 | 6 | 1 | 1 | 2 | 9 | |
| Oncogenes | 0 | 6 | 1 | 3 | 4 | 9 | 11 |

*Some genes are not present in any gene family above.

also bind to the same set of core-bound genes. Using DAVID tool [12, 13], we identified 145 TFs, where each TF bound at least 30 genes in the core-bound gene set (Table S2). We referred to the 145 TFs as the computationally predicted TFs associated with transcriptional regulation in human ES cells because these TFs are regulating the same genes that are also transcriptionally regulated by the core TFs. We carried out a network analysis for the 145 TFs using IPA. Our goal is to see if these extended TFs are part of the original core TF circuitry or if they use independent circuitries to regulate the core-bound genes. Figure 2 presents a significant regulatory network related to the 145 TF gene set. The network involved

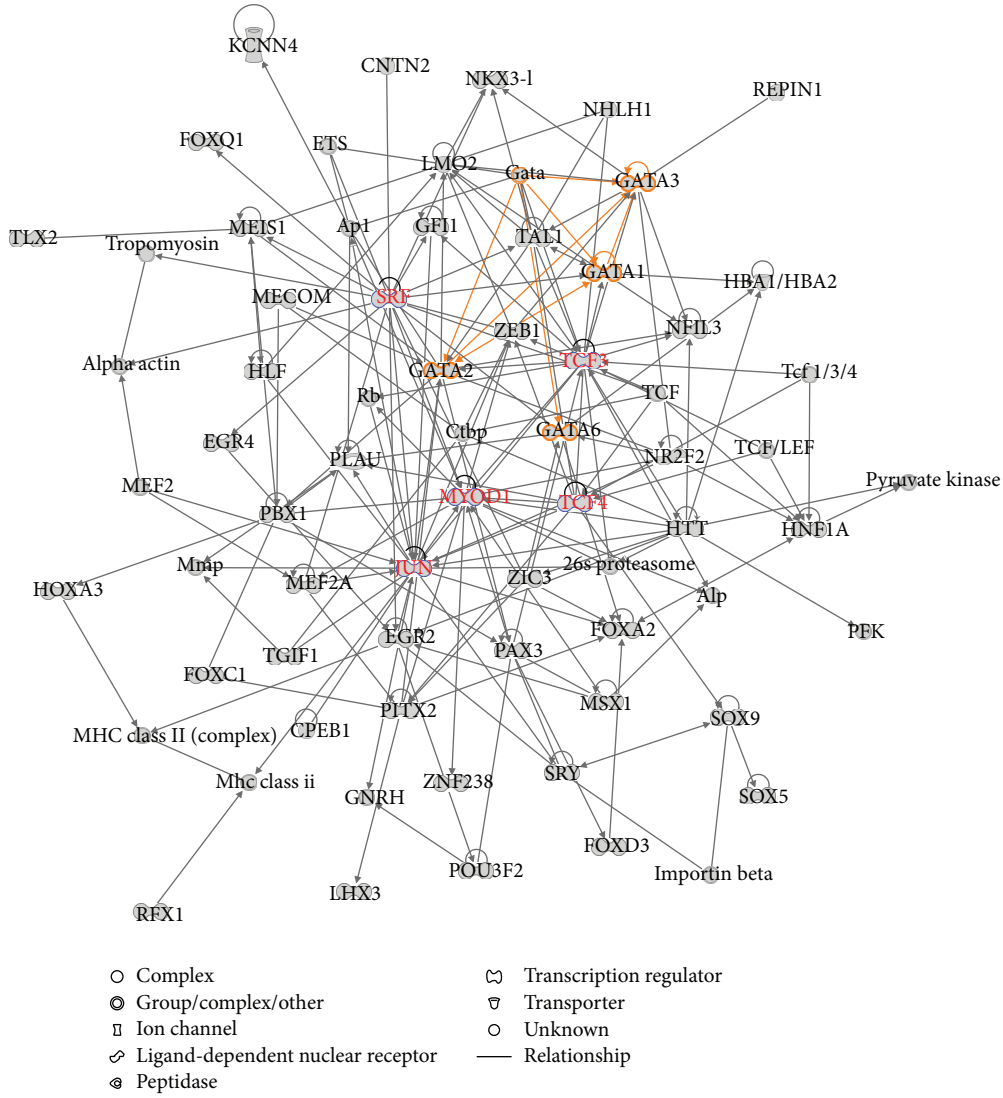



Figure 2: A significant regulatory network related to the 145 computationally predicted human ES cell related gene set. The GATA family of TFs and other important TFs are highlighted.

70 nodes among which the GATA transcription factor family members (GATA1, GATA2, GATA3, and GATA6) form interconnected autoregulatory and feedforward circuitry (in yellow), suggesting that GATA TFs are active in transcriptional regulation in human ES cells. The network also shows that several TF genes such as TCF3, TCF4, SRF, MYOD1, and JUN form hub nodes (in red), suggesting their significance in the same circuitry. Biological function analysis indicated that the TFs were significant in regulation of cell and organ development (Figure S1). Pathway analysis indicated that the TFs were mostly involved in the transcriptional regulatory network in embryonic stem cells pathway ($P$ value $\approx 10^{-12}$) (Figure S2), the same result as that shown in the core-bound gene analysis.

In a recent study [14], we have collected 189 TFs that have been experimentally verified to contribute to transcriptional regulation in human ES cells. We found that there were 41 overlaps between the 189 TFs set and the computationally predicted 145 TF set from DAVID program as shown in Table S3.

*3.3. Extension of Transcriptional Network in Human ES Cells.* Boyer et al. have identified the core transcriptional regulatory network in human ES cells in which the three core TFs collaborate to regulate a substantial number of their target genes [2]. We tried to extend the core transcriptional regulatory network based on the combination of the core-bound genes and the TFs that bind to subsets of the core-bound genes. The combined gene set was composed of the 353 core-bound genes and the aforementioned 145 TF genes. The five most significant networks associated with the combined gene set were summarized in Table S4. Note that 4 of the 5 networks were associated with embryonic development.

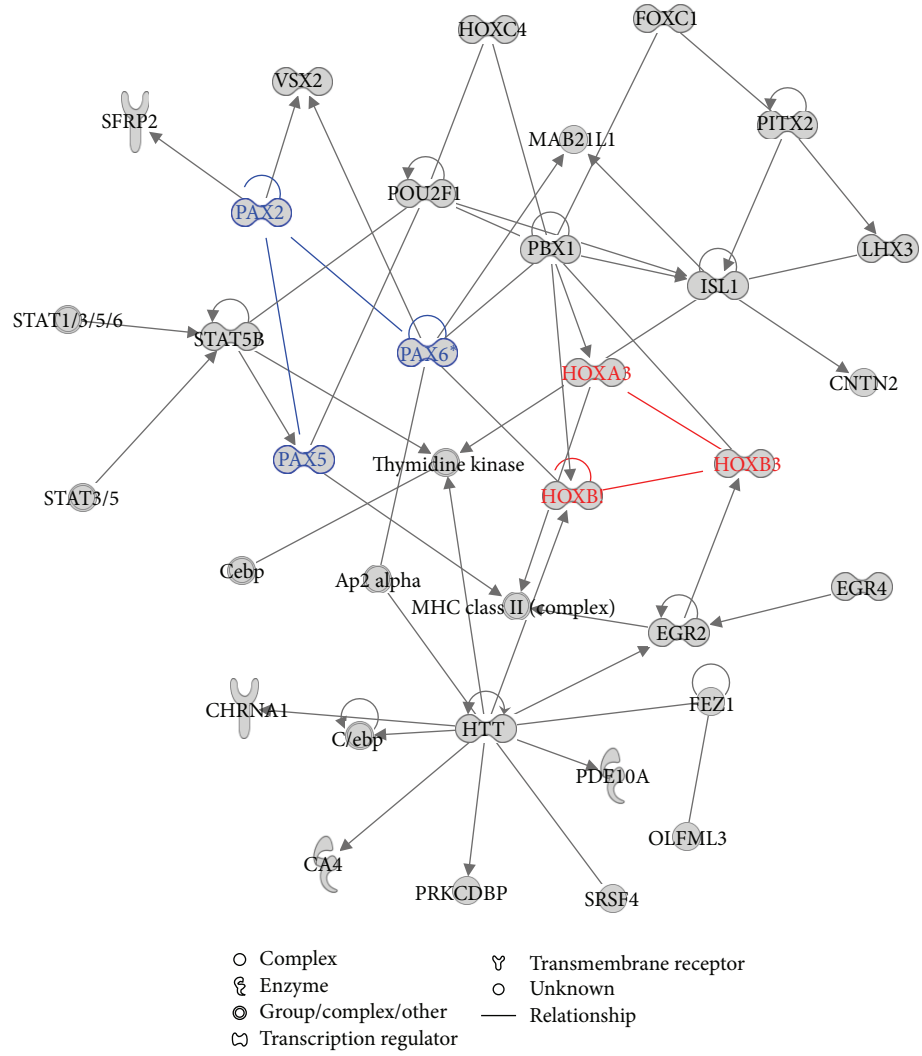



FIGURE 3: The network related to the combined gene set with the fourth highest score. The combined gene set is made up of the 353 core-bound genes and the 145 computationally predicted human ES cell related TF genes. The HOX and PAX families of TF genes are highlighted in red and blue color, respectively.

Below, we describe the regulatory circuits of important TFs or families of TFs in each network.

Figure S3 shows that the three core TFs (NANOG/OCT4/SOX2) act as the hub genes in the regulatory network, which is anticipated; Figure S4 shows that TP53 is the center of the regulatory network with the third highest score, indicating that TP53 plays an active role in the transcriptional circuit for human ES cells. In fact, many experimental lines of evidence have revealed that TP53 plays a key role in determining the fate of ES cells [16–20]. Silencing of the tumor suppressor gene TP53 significantly increased the reprogramming efficiency of human somatic cells [17]. Some studies have shown that the p53 pathway can maintain the homeostasis of self-renewal and differentiation of human ES cells [21–23].

Figure 3 shows three important gene families, HOX, PAX, and STAT, that are highly active in the regulatory network. The members of PAX and HOX gene family form autoregulatory loop and also regulate members of other gene families in the network. Interestingly, within individual autoregulatory loop, PAX2 and PAX6 self-regulate and show bidirectional regulation on each other but with contrary effect: PAX2 positively regulated PAX6, while PAX6 has inhibitory effect on PAX2. Based on the regulatory circuitry shown in Figure 3, we infer that HOX, PAX, and STAT gene families play a very important role in controlling the fate of human ES cells by forming a specific regulatory motif. In fact, these three gene families have been experimentally verified to be important in the regulation of developmental processes of human ES cells. HOX genes encode TF proteins which are master regulators of embryonic development [24]. They are important targets of OCT4, SOX2, and NANOG and often transcriptionally inactive when bound by the core regulators to inhibit differentiation. Our results show that except for

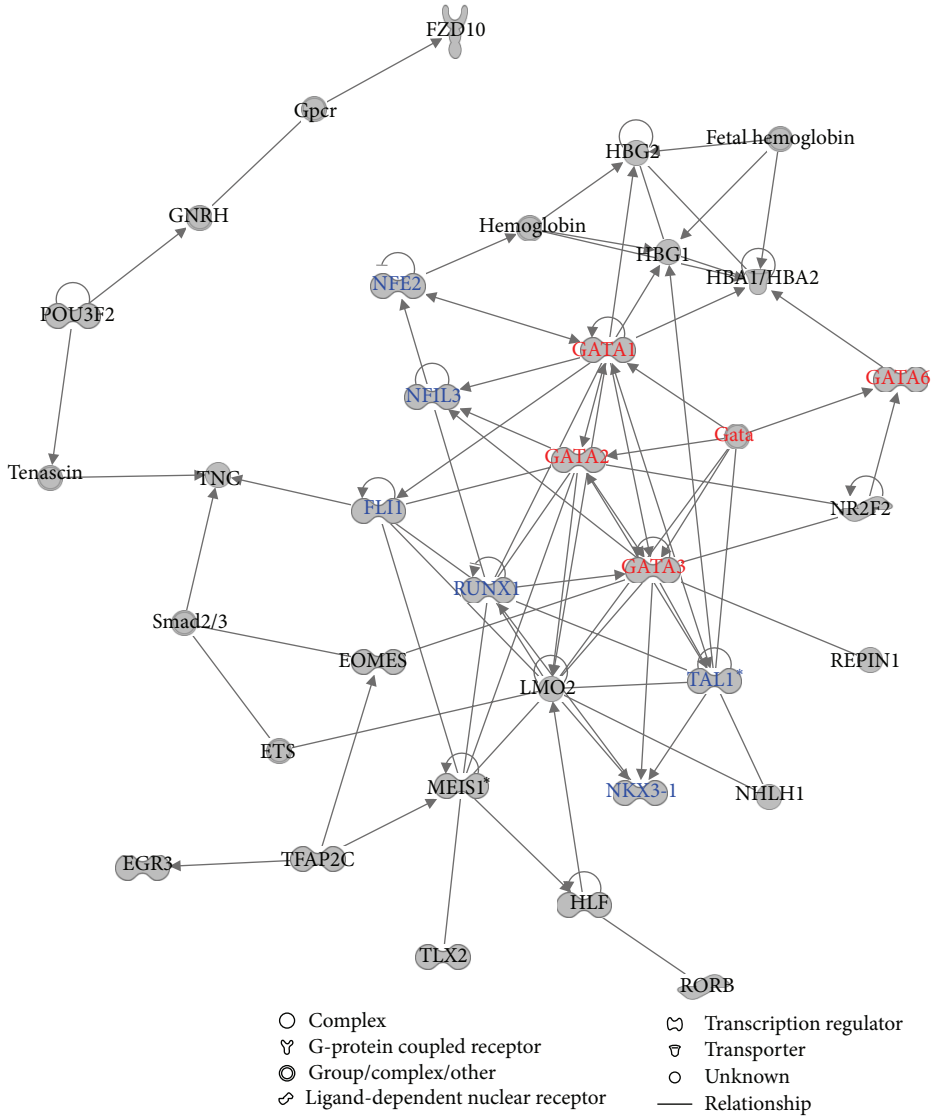



Figure 4: The network related to the combined gene set with the fifth highest score. The combined gene set is made up of the 353 core-bound genes and the 145 computationally predicted human ES cell related TF genes. Red indicates the GATA TF family member and blue indicates the other TFs that interact with the GATA TF family member.

regulated by the core regulators (Figure 1), HOX family of genes could form their own internal and autoregulatory loop to control the developmental processes of human ES cells. On the other hand, PAX is a family of tissue-specific TFs containing a paired domain and usually with a partial or complete homeodomain. PAX regulates cell proliferation and self-renewal, resistance to apoptosis, migration of embryonic precursor cells, and the coordination of specific differentiation programs during embryonic development. Therefore, PAX plays an essential role in regulation of the pluripotency and self-renewal of human ES cells [25]. Finally, STAT family of TFs regulate cell growth, survival, and differentiation via activation by JAK (Janus kinase). This pathway is critical for regulation of stem cell self-renewal and differentiation [26].

Another network (Figure 4) shows that the GATA family of TFs interconnects and forms regulatory circuit with the other six TFs including NFE2, NFIL3, RUNX1, NKX3-1, TAL1, and FLI1. Therefore we infer that GATA is also important in regulation of pluripotency and self-renewal of human ES cells. Previous studies have revealed that GATA was active in transcriptional regulation in human ES cells through transcriptional coexpression with many other key regulators [25, 27, 28].

Therefore, in addition to the core transcriptional network, we infer that some other transcriptional networks are potentially important in regulation of pluripotency and self-renewal of human ES cells.

Pathway analysis shows that the most significant pathways associated with the combined gene set (353 core-bound genes and 145 TFs) include transcriptional regulatory network in embryonic stem cells ($P$ value = $3.73 \times 10^{-49}$), role of OCT4 in mammalian embryonic stem

cell pluripotency ($P$ value = $4.76 \times 10^{-11}$), Wnt/$\beta$-catenin signaling ($P$ value = $1.08 \times 10^{-8}$), and human embryonic stem cell pluripotency ($P$ value = $4.03 \times 10^{-8}$). Apparently, these pathways are strongly associated with the function of regulating the fate of human ES cells.

*3.4. MYC Transcriptional Network in Human ES Cells.* Similar to the core TFs, MYC is a very important TF in the ES cells. A set of 369 genes was identified as MYC targeted genes in human ES cells [26], which are listed in Table S5. We explored the regulatory network involving MYC and MYC target genes using IPA software. As expected, MYC and TP53 turned out as the hub genes in the two most important networks, respectively (Figure S5 and Figure S6). The most significant biological functions associated with this gene set are involved in embryonic, organismal, tissue, and cell development, cell cycle, gene expression, cancer, and so forth. The most significant pathways associated with the gene set included Wnt/$\beta$-catenin ($P$ value = $9.37 \times 10^{-5}$), human ES cell pluripotency ($P$ value = $2.14 \times 10^{-4}$), MYC mediated apoptosis signaling role ($P$ value = $7.59 \times 10^{-4}$), and so forth. Notably, MYC regulated a cluster of genes that were involved in the human ES cell pluripotency pathway.

We were also curious to see if MYC and core TFs regulate the same transcriptional circuitry or operate individually. Hence, we carried out a combined network analysis of the core-bound genes and the MYC target genes. There are only 17 overlapping genes between the core-bound gene set and the MYC target gene set that corresponds to only 5% of each target gene set. In fact, the number of overlapping genes between the MYC target gene set and each of the three core TF's target gene sets is also small (50, 19, and 37 for NANOG, OCT4, and SOX2, resp.). The lower overlapping rate supports the previous argument that the MYC-centered regulatory network belonged to a different module from the core transcriptional module in ES cells [1, 7, 10]. Our network analysis clearly shows that there are two separable modules, the core-centered module and the MYC-centered module, which form the transcriptional circuity in the ES cells (Figure 5).

It has been shown that the core TFs and MYC play key roles in the regulation of ES cells' fate by regulating many TF genes which in turn regulate a large number of other genes [1, 2, 10]. We found that there are 90 TF genes in the 353 core-bound genes and 38 TF genes in the 369 MYC target genes. We carried out an analysis of the regulatory network based on these TF genes only. Figure 6 shows that the core TFs and MYC form center of the two distinct modules. An interesting finding is that MYC has no connection with any of the three core TFs but interconnects with TP53, which in turn regulates NANOG and is regulated by OCT4. This finding suggests that TP53 has stronger link with the core TFs than MYC and also indicates that TP53 might play a key role in bridging the core-centered and MYC-centered modules.

To further investigate the differences in the regulatory modules of the core- and MYC-centered networks, we obtained two gene sets: a gene set in the core regulatory module and a gene set in the MYC-centered regulatory module, both from the mouse ES cells [7]. We used the human orthologs of the mouse genes in both modules, which contained 75 and 356 genes, respectively (Table S6). There were only three overlapping genes between both modules, again showing that both modules were functionally separate. Similarly, we inferred the significant networks associated with the core module and the MYC module, respectively (Figure S7 and Figure S8). The top 5 pathways associated with both modules were present in Table S7. There are no overlapping pathways between both modules, suggesting that the MYC module and the core module are indeed involved in very different pathway patterns in regulating pluripotency and self-renewal of ES cells.

## 4. Conclusions

It has been found that transcriptional networks were essentially responsible for regulation of pluripotency and self-renewal of human ES cells. Some key TFs like NANOG, OCT4, and SOX2 have been identified to collaboratively control pluripotency and self-renewal by forming interactive regulatory circuits [2, 29]. However, it is presently unclear how the transcriptional networks precisely control the activity of ES cells. It is likely that additional TFs may also regulate the key downstream TFs or form additional regulatory circuits that are involved in the regulation of pluripotency and self-renewal of human ES cells. We have explored an extension of the core transcriptional regulatory network by adding additional TFs into the core transcriptional networks.

Evidence shows that many TFs are involved in both ES cell fate determination and cancerous pathogenesis. For example, oncogene MYC and tumor suppressor gene TP53 have been shown to significantly contribute to the formation of the transcriptional networks that determine the self-renewal or differentiation fate of human ES cells. Several families of human ES cell associated TFs like MYB, E2F, PAX, SMAD, STAT, POU, SP, and GLI are related to cancer [14]. This evidence suggests that ES cell and cancer cells may share essential regulatory mechanisms. Therefore, understanding of how the regulatory network regulates self-renewal or differentiation fate of human ES cells may pave the way for understanding of cancer, and further conquering cancer.

In addition, based on the comparisons of the MYC-centered regulatory module and the core regulatory module in human ES cells, our results suggest that MYC acts independently of the core module in the regulation of pluripotency of human ES cells. In addition, we also showed that TP53 is a key connecting molecule between the core-centered and MYC-centered modules.

Our computational network-based approach supplements the experimental methods to unravel the transcriptional regulatory mechanisms that control pluripotency and self-renewal in the ES cells, although the reliability of our results needs further experimental verification. However, it

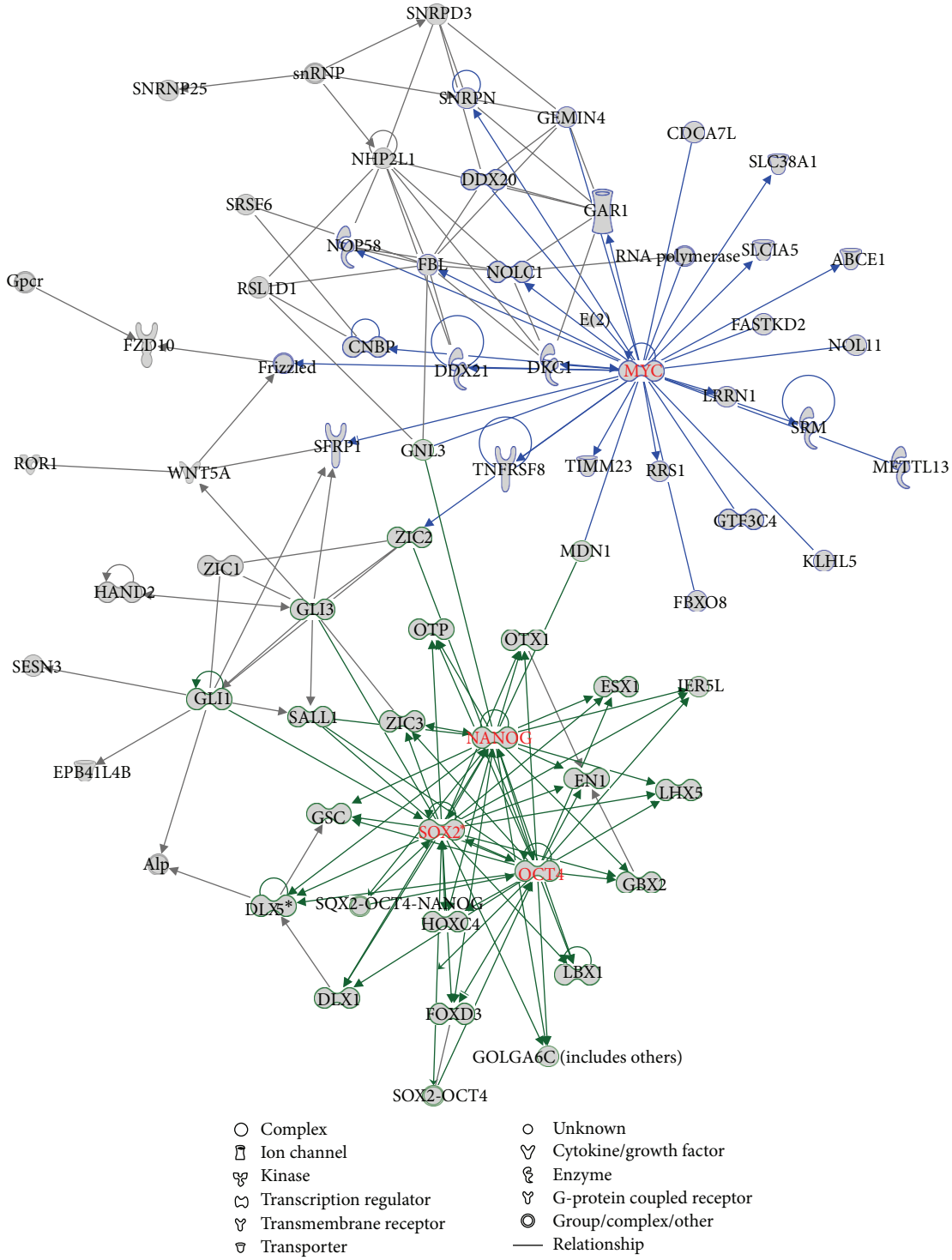



Figure 5: The network based on the combination of the core-bound genes and the MYC targeted genes. The core-centered module and the MYC-centered module are highlighted.

should be noted that there exist certain limitations in the present methods. First, the information collected by IPA and DAVID databases is from many different studies that are not necessarily specific to human ES cells; hence, the extrapolation of such data to ES cells may lead to false positive information in certain cases. Secondly, as the new findings presented in this study lack experimental verification, it is difficult to assess the sensitivity and specificity of this approach. We plan to collaborate with experimental investigators to validate some of these findings in the future.

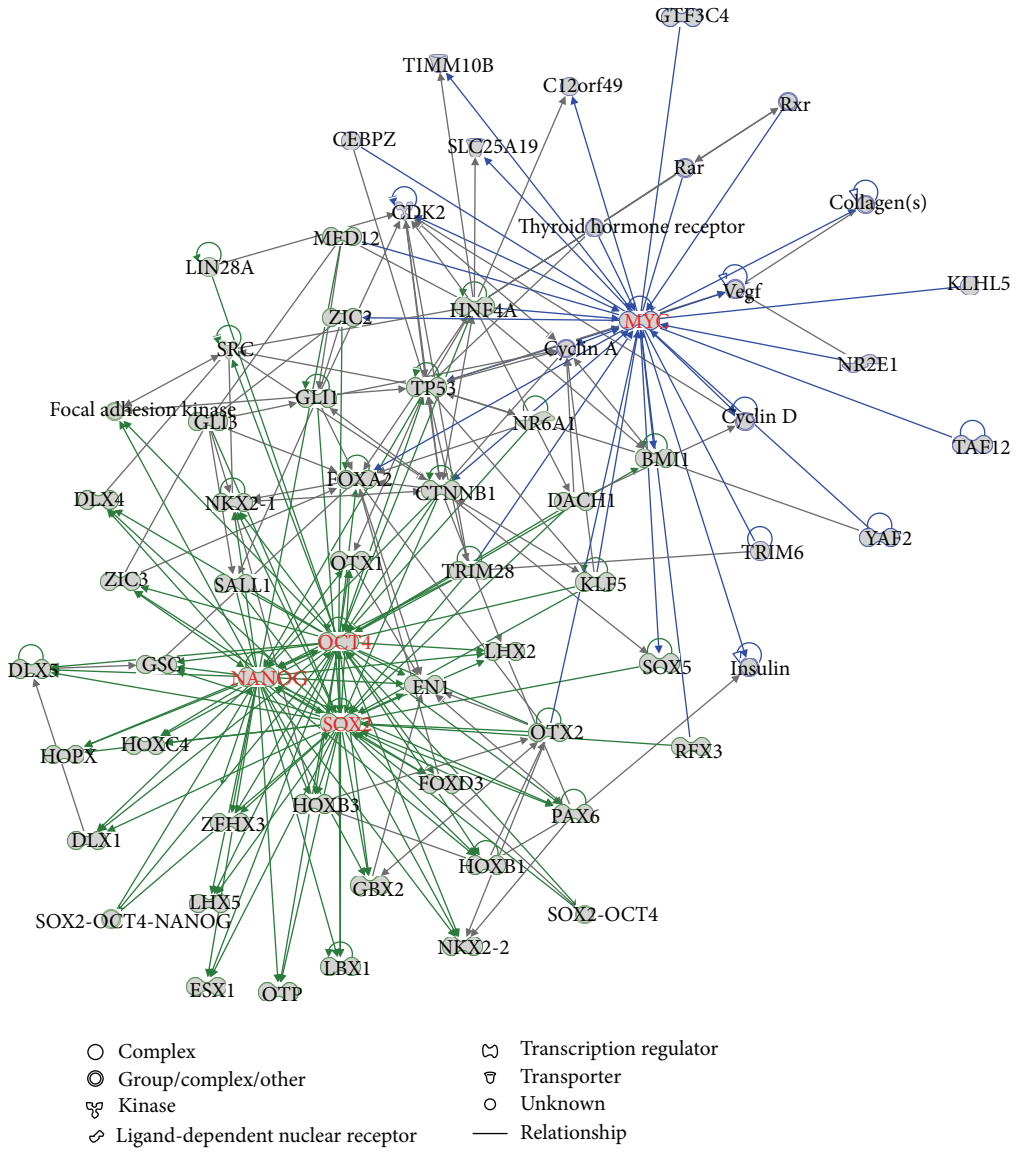



Figure 6: The network based on the combination of the core-bound TF genes and the MYC targeted TF genes. The core-centered module and the MYC-centered module are highlighted.

## Conflict of Interests

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

## Acknowledgment

This work was fully supported by the start-up funds to Chittibabu Guda from University of Nebraska Medical Center.

## References

[1] J. Kim, J. Chu, X. Shen, J. Wang, and S. H. Orkin, "An extended transcriptional network for pluripotency of embryonic stem cells," *Cell*, vol. 132, no. 6, pp. 1049–1061, 2008.

[2] L. A. Boyer, T. I. Lee, M. F. Cole et al., "Core transcriptional regulatory circuitry in human embryonic stem cells," *Cell*, vol. 122, no. 6, pp. 947–956, 2005.

[3] S.-I. Aota, N. Nakajima, R. Sakamoto, S. Watanabe, N. Ibaraki, and K. Okazaki, "*Pax6* autoregulation mediated by direct interaction of Pax6 protein with the head surface ectoderm-specific enhancer of the mouse *Pax6* gene," *Developmental Biology*, vol. 257, no. 1, pp. 1–13, 2003.

[4] H. Niwa, K. Ogawa, D. Shimosato, and K. Adachi, "A parallel circuit of LIF signalling pathways maintains pluripotency of mouse ES cells," *Nature*, vol. 460, no. 7251, pp. 118–122, 2009.

[5] I. Chambers and S. R. Tomlinson, "The transcriptional foundation of pluripotency," *Development*, vol. 136, no. 14, pp. 2311–2322, 2009.

[6] W. Shi, H. Wang, G. Pan, Y. Geng, Y. Guo, and D. Pei, "Regulation of the pluripotency marker *Rex-1* by nanog and

Sox2," *Journal of Biological Chemistry*, vol. 281, no. 33, pp. 23319–23325, 2006.

[7] J. Kim, A. J. Woo, J. Chu et al., "A Myc network accounts for similarities between embryonic stem and cancer cell transcription programs," *Cell*, vol. 143, no. 2, pp. 313–324, 2010.

[8] K. Takahashi and S. Yamanaka, "Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors," *Cell*, vol. 126, no. 4, pp. 663–676, 2006.

[9] K. Takahashi, K. Tanabe, M. Ohnuki et al., "Induction of pluripotent stem cells from adult human fibroblasts by defined factors," *Cell*, vol. 131, no. 5, pp. 861–872, 2007.

[10] X. Chen, H. Xu, P. Yuan et al., "Integration of external signaling pathways with the core transcriptional network in embryonic stem cells," *Cell*, vol. 133, no. 6, pp. 1106–1117, 2008.

[11] 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.

[12] 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.

[13] 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.

[14] X. Wang, "Computational analysis of expression of human embryonic stem cell-associated signatures in tumors," *BMC Research Notes*, vol. 4, article 471, 2011.

[15] H. Ji, G. Wu, X. Zhan et al., "Cell-type independent MYC target genes reveal a primordial signature involved in biomass accumulation," *PLoS ONE*, vol. 6, no. 10, Article ID e26057, 2011.

[16] H. Hong, K. Takahashi, T. Ichisaka et al., "Suppression of induced pluripotent stem cell generation by the p53–p21 pathway," *Nature*, vol. 460, no. 7259, pp. 1132–1135, 2009.

[17] T. Kawamura, J. Suzuki, Y. V. Wang et al., "Linking the p53 tumour suppressor pathway to somatic cell reprogramming," *Nature*, vol. 460, no. 7259, pp. 1140–1144, 2009.

[18] R. M. Marión, K. Strati, H. Li et al., "A p53-mediated DNA damage response limits reprogramming to ensure iPS cell genomic integrity," *Nature*, vol. 460, no. 7259, pp. 1149–1153, 2009.

[19] T. Lin, C. Chao, S. Saito et al., "p53 induces differentiation of mouse embryonic stem cells by suppressing *Nanog* expression," *Nature Cell Biology*, vol. 7, no. 2, pp. 165–171, 2005.

[20] A. Cicalese, G. Bonizzi, C. E. Pasi et al., "The tumor suppressor p53 regulates polarity of self-renewing divisions in mammary stem cells," *Cell*, vol. 138, no. 6, pp. 1083–1095, 2009.

[21] H. Qin, T. Yu, T. Qing et al., "Regulation of apoptosis and differentiation by p53 in human embryonic stem cells," *Journal of Biological Chemistry*, vol. 282, no. 8, pp. 5842–5852, 2007.

[22] C. Grandela, M. F. Pera, S. M. Grimmond, G. Kolle, and E. J. Wolvetang, "p53 is required for etoposide-induced apoptosis of human embryonic stem cells," *Stem Cell Research*, vol. 1, no. 2, pp. 116–128, 2007.

[23] K. Sabapathy, M. Klemm, R. Jaenisch, and E. F.wagner, "Regulation of ES cell differentiation by functional and conformational modulation of p53," *The EMBO Journal*, vol. 16, no. 20, pp. 6217–6229, 1997.

[24] T. R. J. Lappin, D. G. Grier, A. Thompson, and H. L. Halliday, "HOX genes: seductive science, mysterious mechanisms," *The Ulster Medical Journal*, vol. 75, no. 1, pp. 23–31, 2006.

[25] Y. Sun, H. Li, Y. Liu, M. P. Mattson, M. S. Rao, and M. Zhan, "Evolutionarily conserved transcriptional co-expression guiding embryonic stem cell differentiation," *PLoS ONE*, vol. 3, no. 10, Article ID e3406, 2008.

[26] J.-Y. Rho, K. Yu, J.-S. Han et al., "Transcriptional profiling of the developmentally important signalling pathways in human embryonic stem cells," *Human Reproduction*, vol. 21, no. 2, pp. 405–412, 2006.

[27] M. Jung, H. Peterson, L. Chavez et al., "A data integration approach to mapping OCT4 gene regulatory networks operative in embryonic stem cells and embryonal carcinoma cells," *PLoS ONE*, vol. 5, no. 5, Article ID e10709, 2010.

[28] L. Chavez, A. S. Bais, M. Vingron, H. Lehrach, J. Adjaye, and R. Herwig, "*In silico* identification of a core regulatory network of OCT4 in human embryonic stem cells using an integrated approach," *BMC Genomics*, vol. 10, article 314, 2009.

[29] A. Rizzino, "The Sox2-Oct4 connection: critical players in a much larger interdependent network integrated at multiple levels," *Stem Cells*, vol. 31, no. 6, pp. 1033–1039, 2013.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 470418, 9 pages
http://dx.doi.org/10.1155/2013/470418

*Research Article*

# HGF Accelerates Wound Healing by Promoting the Dedifferentiation of Epidermal Cells through $\beta_1$-Integrin/ILK Pathway

**Jin-Feng Li,[1] Hai-Feng Duan,[2] Chu-Tse Wu,[2] Da-Jin Zhang,[3] Youping Deng,[4] Hong-Lei Yin,[1] Bing Han,[1] Hui-Cui Gong,[1] Hong-Wei Wang,[5] and Yun-Liang Wang[1]**

[1] *The Neurology Department of the 148th Hospital, 20 Zhanbei Road, Zibo 255300, China*
[2] *Department of Experimental Hematology, Beijing Institute of Radiation Medicine, 27 Taiping Road, Beijing 100850, China*
[3] *Medical Research Center of Naval General Hospital, 6 Fucheng Road, Beijing 10037, China*
[4] *Wuhan University of Science and Technology, Wuhan, Hubei 430081, China*
[5] *Department of Medicine, The University of Chicago, Chicago, IL 60637, USA*

Correspondence should be addressed to Hong-Wei Wang; hongweiwang2012@gmail.com
and Yun-Liang Wang; wangyunliang81@163.com

Received 2 September 2013; Revised 15 November 2013; Accepted 2 December 2013

Academic Editor: Jason E. Mcdermott



Skin wound healing is a critical and complex biological process after trauma. This process is activated by signaling pathways of both epithelial and nonepithelial cells, which release a myriad of different cytokines and growth factors. Hepatocyte growth factor (HGF) is a cytokine known to play multiple roles during the various stages of wound healing. This study evaluated the benefits of HGF on reepithelialization during wound healing and investigated its mechanisms of action. Gross and histological results showed that HGF significantly accelerated reepithelialization in diabetic (DB) rats. HGF increased the expressions of the cell adhesion molecules $\beta_1$-integrin and the cytoskeleton remodeling protein integrin-linked kinase (ILK) in epidermal cells *in vivo* and *in vitro*. Silencing of ILK gene expression by RNA interference reduced expression of $\beta_1$-integrin, ILK, and c-met in epidermal cells, concomitantly decreasing the proliferation and migration ability of epidermal cells. $\beta_1$-Integrin can be an important maker of poorly differentiated epidermal cells. Therefore, these data demonstrate that epidermal cells become poorly differentiated state and regained some characteristics of epidermal stem cells under the role of HGF after wound. Taken together, the results provide evidence that HGF can accelerate reepithelialization in skin wound healing by dedifferentiation of epidermal cells in a manner related to the $\beta_1$-integrin/ILK pathway.

## 1. Introduction

Skin wound healing is a multifaceted process of reepithelialization that requires epidermal cell proliferation and migration, collagen fiber rearrangement, and cutaneous adnexa repair [1–3]. These epidermal cells are terminally-differentiated, but the molecular mechanisms involved in their proliferation and migration remain incompletely understood. According to Jones' report, if the differentiated epidermal cells high expressed $\beta_1$-integrin, they would have a stronger ability to form clones and passage and generate a complete epithelium after moving them to skin wounds [4, 5]. Therefore, high expressed $\beta_1$-integrins can prompt epidermal cells into a high proliferative and dedifferentiated state. It is well known that hepatocyte growth factor (HGF) regulates cell growth, cell motility, and morphogenesis in various types of cells, including epithelial and endothelial cells, supporting the hypothesis that it promotes epithelial repair and neovascularization during wound healing [6–9]. However, there are few reports that HGF was associated with $\beta_1$-integrin in the process of promoting wound healing. In the present study, a plasmid carrying the HGF gene (PUDKH) was locally injected into the injured skin of diabetic rats, causing the plasmid-treated group to exhibit accelerated wound healing and increased expression of $\beta_1$-integrin in epidermal lays and the molecular surface

marker of epidermal stem cells. Overexpression of HGF in scratched primary rat epidermal cells also increased the expression of $\beta_1$-integrin, confirming the result in a more isolated experimental system *in vitro*. $\beta_1$-Integrin is thought to be one of specific markers for epidermal stem cells [10]. ILK was first discovered as a $\beta_1$-integrin subunit binding protein. It localizes at the focal adhesions as well as sites of invasion and migration and is involved in cytoskeleton remodeling. Although the function of ILK has been intimately associated with integrin function, little association with HGF/c-met has been reported [11]. In this study, silencing ILK gene expression by RNA interference decreased the expressions of $\beta_1$-integrin and c-met, concomitantly reducing the proliferation and migration ability of epidermal cells. These data demonstrate that HGF can accelerate skin wound healing by promoting the dedifferentiation of epidermal cells, while this process is closely related to ILK, an intracellular effector of cell-matrix interactions.

## 2. Materials and Methods

*2.1. Reagents, Antibodies, and the Expression Vectors.* Male Wistar rats were purchased from the Animal Center of the Academy of Military Medical Sciences (9-week old and weighing 270–300 g); Ham's F12 nutrient medium (Ham's F12), recombinant human HGF protein, and MTT were purchased from Gibco (Grand Island, New York, USA). Dimethyl sulfoxide and streptozotocin (STZ) were bought from Sigma (Austin, Texas, USA). All antibodies were obtained from R&D systems (Minneapolis, Minnesota, USA). The plasmid carrying human HGF cDNA was constructed by a colleague of the authors. The siRNAs against ILK and a nontarget siRNA were from Shanghai Gene Chemical Company (Songjiang, Shanghai, China).

*2.2. Preparation of Animal Model [12].* The rate of wound healing in animal is usually more faster than that of human beings. So the effect of HGF will not be easy to be observed in the rapid wound healing process, while chronic wound healing is a troublesome and common complication of diabetes. Therefore rats were made diabetic by a single intraperitoneal injection of streptozotocin (STZ, Sigma Company) in this study. According to previous report, the diabetic rats actually showed significantly delayed wound healing than nondiabetic rats [13]. Rats with STZ-induced diabetes were fed a high fat diet (HFD) during the whole experimental period, whereas control rats were fed with basal diet (BD) served at the same time. HFD was prepared by adding sucrose (20%, w/w) and lard (20%, w/w) into BD. After 5 weeks, a single intraperitoneal injection of STZ (40 mg/kg dissolved in 100 mM citrate buffer pH 4.5) was administered to rats fed with HFD. Control rats received an equivalent volume of citrate buffer by intraperitoneal injection. Blood glucose levels were measured 72 h after STZ injection by tail vein puncture blood sampling using a hand-held glucometer (Changsha Sinocare Inc., China). Serum triglyceride (TG) and total cholesterol (TC) were determined by an autobiochemical analysis system (AU2700, Olympus, Japan), and body weight was recorded every week. Rats with blood sugar values at least 11.6 mmol/L were used for this study. During this period DB rats showed clinical signs of diabetes mellitus, for example, polyuria, polyphagia, and weight loss.

All DB rats were randomly divided into three groups (PUDKH group, PUDK group, and PBS group) ($N = 20$). DB rats were anesthetized for wounding with an intraperitoneal injection of sodium pentobarbital (0.5 mL/kg), and the hair on their back was clipped and the skin was cleaned. A round full-thickness wound measuring 2 cm in diameter was then made on the back of each animal using a 2 cm round scalpel. Wounds on rats in the PUDKH group were dressed with 50 $\mu$g PUDKH per square centimeter by high-pressure syringe, while PUDK and PBS control rats were, respectively, treated with the same amount of empty plasmid and PBS. The wound areas of six DB rats were measured on 1, 3, 5, 7, and 14 d after gene transfer. Measurements were made with the aid of an image analyzer and a VEV image analysis software package (both from Di Meide Science and Technology Co., Ltd., Beijing). Two rats were killed at each time point, and skin samples were fixed in 4% formalin solution, embedded in paraffin, and sectioned (4 $\mu$m) for histopathological evaluation.

*2.3. Immunofluorescence and Immunohistochemistry.* In the first stage of skin wound healing, inflammatory exudate and blood crust formation are dominant. The obvious reepithelialization phenomenon usually starts after 5 days. We therefore chose to detect the expressions of $\beta_1$-integrin and ILK in rat skin tissue 7 d after gene transfer. Immunofluorescence (IF) staining was used to detect $\beta_1$-integrin in epidermis during wound healing. Skin tissues were fixed in ice-cold methanol for 10 min and washed in PBS. After being blocked with 2% BSA in PBS, primary antibodies were applied overnight in a moist chamber set at 4°C and then rinsed with PBS three times for 5 min each rinse. Samples were then incubated with the 1 : 200 diluted tetraethyl rhodamine isothiocyanate- (TRITC-) conjugated-goat and anti-mouse immunoglobulin G secondary antibodies for 45 min in a dark incubation chamber at 37°C. After the skin tissues were washed in PBS, DAPI (1 : 1000) was used for nuclear staining. All specimens were examined under a fluorescence microscope (IX71-A12FL/PH, Olympus, Japan). Negative controls were prepared by incubation with the secondary antibody alone.

Immunohistochemistry was used to detect the expression of ILK. Four-micrometer paraffin sections were subjected to antigen retrieval using a pressure cooker, in sodium citrate (pH 6.0), for 4 minutes. Endogenous peroxidase was blocked with 3% hydrogen peroxide ($H_2O_2$) in PBS followed by nonspecific blocking with 2% PBS + bovine serum albumin (BSA) for 15 minutes. The sections were incubated with the primary antibody overnight at 4°C. After washing with PBS, slides treated with biotin-labeled secondary antibodies (1 : 500, R&D, USA) were incubated at RT for 1 h. The chromogenic reagent DAB was used to show the antibody conjugation. The intensity of the reaction observed on the slides was qualitatively analyzed.

*2.4. Isolation and Culture of Epidermal Cells.* Briefly, ultrathin epidermal sheets (grafted or ungrafted) were cut into pieces, digested with 0.25% trypsin for 20 min at 37°C, and made into single cell suspensions. After centrifugation, the epidermal cells were gently resuspended in Epilife medium supplemented with 1% human epidermal cells growth supplement and seeded on collagen IV coated culture flasks at a density of $5 \times 10^5$ epidermal cells/cm$^2$. After 24 hrs, nonadherent cells were gently removed.

*2.5. In Vitro Studies Using Rat Epidermal Cells.* For wounding experiments *in vitro*, epidermal cells were cultured to 100% confluency. The monolayer was scratched with a sterile needle to give a 0.8 mm wide wound, washed twice, and cultured in culture media with rat HGF (50 ng/mL). Inhibition of ILK expression in epidermal cells with ILK-specific siRNA reagents was performed as described previously [14]. Small interfering RNAs (siRNA) for rat ILK were synthesized by Shanghai Gene Chemical Company. siRNAs were transfected into the cells using Lipofectamine 2000 reagent (Invitrogen Life Technologies).

Images of wound areas were captured with Moticam (Motic Microscopes). Total scratch wound area was measured using Image Plus Software, and the percentage of wound closure at each time point was derived by following the formula: (1-[current wound size/initial wound size]) × 100 [15]. Following treatment, cells were washed in ice-cold PBS, and total cell lysates were prepared by scraping the cells in lysis buffer. Lysates were rotated at 4°C for 1 h and the insoluble material was removed by centrifugation at 12,000 ×g for 10 min. Equal amounts of denatured proteins were separated by 12% SDS-PAGE and transferred to PVDF membranes (Pharmarcia). The membranes were blocked by incubation in Tris-buffered saline nonfat dry milk for 2 h, followed by incubation at room temperature with indicated antibodies (against ILK, c-met, and $\beta$-Actin) at room temperature for 2 h. After extensively washing in Tris-buffered saline containing 0.1% Tween-20, the membranes were incubated for 1 h with horseradish peroxidase-conjugated secondary antibody. Membranes were then washed and developed using enhanced chemiluminescence substrate (ECL, Amersham Pharmacia Biotech).

After being scratched, the cells were cultured with HGF (50 ng/mL) for 48 h. The cells were then fixed in 4% paraformaldehyde in 0.1 M phosphate buffer (PBS, pH 7.4) for 15 min at room temperature and washed three times with phosphate buffered saline (PBS; 0.01 M phosphate, pH 7.3, 0.15 M NaCl). Next, cells were incubated in blocking buffer (PBS with 0.02% sodium azide, 0.2% Triton X-100, and 10% serum) for 1 h at room temperature. For immunoperoxidase labeling, cells were incubated in primary antibody ($\beta_1$-integrin, CK19, and CK10) overnight at 4°C. After washing three times with PBS, cells were incubated in secondary biotinylated antibody for 2 h at room temperature. The chromogen was diaminobenzidine (DAB; 0.5 mg/mL in PBS) with 0.12% $H_2O_2$. After immunostaining, cells on coverslips were mounted and analyzed by an image analysis system. In the negative control, the antibodies were replaced with PBS.

*2.6. Cell Migration Assay.* Assessment of cell migration was performed as recently described [16] with minor modifications. Epidermal cells were dislodged after brief trypsinization and dispersed into homogeneous single cell suspensions that were washed extensively with DMEM/0.1% acid-free bovine serum albumin (migration medium) and resuspended in the same medium. Cells ($1 \times 10^5$) were dispersed onto collagen-coated chemotaxis filters that partition transwell inserts into upper and lower chambers. Migration medium (600 $\mu$L) was placed in the lower chambers and the cells were allowed to adhere onto transwells for 1 h at 37°C. The medium in the lower chambers was then removed and cells were challenged by adding 600 $\mu$L of fresh migration medium containing 0 or 50 ng/mL HGF into the lower chamber. Migration was allowed to proceed for 2 h at 37°C. Cells remaining attached to the upper surface of the filters were carefully removed with cotton swabs. The numbers of migrating cells in at least 10 consecutive fields were enumerated and their average was calculated after crystal violet staining. Data were expressed as the number of migrating cells per field.

*2.7. Statistical Analysis.* All data were expressed as mean ± standard deviation (X ± SD). Comparisons between groups were made using one-way analysis of variance. A $P$ value of less than 0.05 was considered to be statistically significant. The results of immunofluorescence and immunocytochemistry were analyzed by an image analysis system. The integrated optical density (IOD) values of 10 fields were randomly determined in each sample under a microscope with the resulting IOD values used to do statistical analysis. The IOD and gray values were assayed by Image-Pro Plus 5.0 image analyzer (Media Cybernetics, USA).

## 3. Results

*3.1. Wound Lesion Size and Histopathological Observation.* Gross observation of dorsal wound revealed that a reduction of the wound area in the PUDKH group was qualitatively visible (Figure 1(a)). The wound areas of six DB rats in PBS, PUDK, and PUDKH groups were measured at 1, 3, 5, 7, and 14 d after HGF gene transfer (specific wound area values are not shown). Measurements were made with the aid of an image analyzer and an image analysis software package. As shown in Figure 1(b), it is obvious that the degree of reepithelialization of the wounds in HGF gene-transfer rats (PUDKH group) was significantly increased after 5, 7, 10, and 14 days compared to the PBS and PUDK groups following gene transfer ($^*P < 0.05$). At 14 d, the PUDKH group rat wounds were almost healed, while it was 20 days until the wounds in the PBS and PUDK rats had healed.

In Figure 1(c), tissue sections from normal rats (hematoxylin/eosin staining) showed a regularly-stratified epithelium with ordinary developed hair follicles. In experimental animals, the first stage of skin wound healing is dedicated to hemostasis and the formation of a provisional wound matrix, initiating the inflammatory process. Next, neovascularization and angiogenesis are activated, marked by the immigration of local fibroblasts along the fibrin network and the beginning

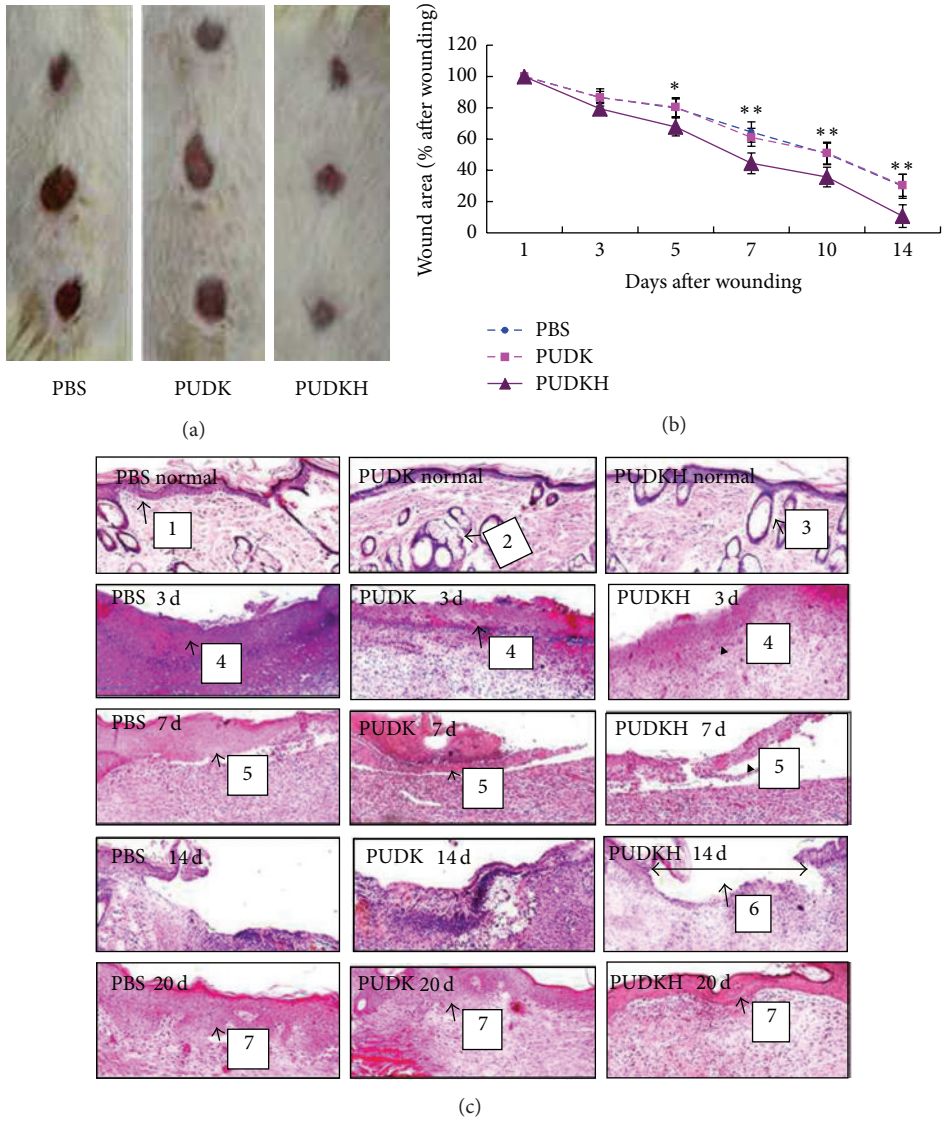



Figure 1: The gross and histological observation. (a) Photographs of full-thickness excisional punch wounds created in the skin of DB rats using a 2 cm biopsy tool 14 d prior to photography. After 14 d, the PUDKH group rat wounds are almost healed, while it took 20 days for wounds on rats in PBS and PUDK groups to heal. (b) Photographs of wounds were captured on 1, 3, 5, 7, 10, and 14 days after wounding to determine the degree of wound closure in DB rats. Graph represents the percentage of wound area at different times after wounding. The $P$ value means that there are differences between PUDKH group versus the PBS and PUDK groups animals at matched time point ($^{**}P < 0.01$). (c) Hematoxylin/eosin staining showed a regularly stratified epithelium with ordinary developed hair follicles in normal rats. Reepithelialization of wounds in HGF gene-transfer rats was significantly increased on 5, 7, 10, and 14 days than that of the PBS and PUDK groups after gene transfer ($^{*}P < 0.05$). On 14 d, regenerated epidermal cells layers recovered skin wounds of DB rats in PUDKH group; until 20 days, the rats in PBS and PUDK groups wound healed. 1: The epidermis of normal rat skin, 2: the sebaceous gland of normal rat skin, 3: the hair follicle of normal rat skin, 4: granuloma in rat skin wounds on 3 d, 5: reepithelialization in the wound edges on 7 d, 6: the regenerated epidermal cells layers recovered skin wounds on 14 d in PUDKH group, and 7: the newly-formed epithelium on 20 d.

of reepithelialization of the wound edges. In the process of wound healing, rapid reepithelialization may prevent pathological scar formation to some extent. Interestingly, the newly-formed epithelium in PBS and PUDK groups was very thick, with more epidermis nipple in the base and coarse/disordered collagen fibers in the dermal layer. This seems to confirm the above view. These results indicated that HGF gene transfer into skin wound may have aided reepithelialization in wound healing. To further explore the specific mechanisms, we measured $\beta_1$-integrin and ILK expression in DB rats wounds.

*3.2. HGF Promotes the Regeneration of Epidermal Cells Expressing $\beta_1$-Integrin and ILK.* Frozen sections of skin epidermis from the three groups were made 7 d after gene-transfer. The expression of $\beta_1$-integrin, the epidermal stem cell molecular surface marker, was detected by immunofluorescence staining. As shown in Figure 2(a), all the four groups

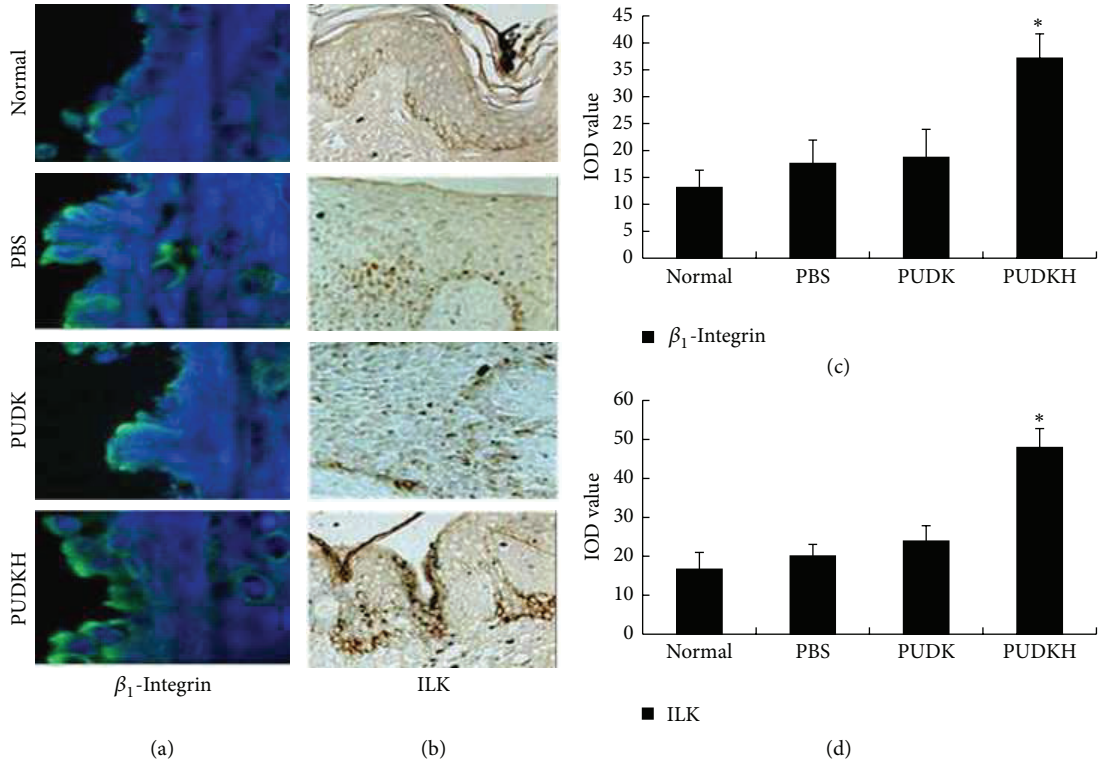



Figure 2: HGF promotes $\beta_1$-integrin and ILK expression in rat epidermal cells. (a) Rat skin tissues transfected with HGF gene on the PUDK plasmid (PUDKH group) and control rats (empty vector (PUDK) group and vehicle (PBS) group) skin tissues were fixed and stained with anti-$\beta_1$-integrin antibody and fluorescein isothiocyanate (FITC) conjugated secondary antibody (green) and nuclei visualized with Hoechst 33342 (blue). (c) A representative image is shown. All four groups expressed blue fluorescent protein in epidermal prickle cell or in the granule cell layer in the process of wound regeneration. But, after analysis by Image-Pro Plus 5.0 image analyzer, we found that blue fluorescent proteins expression was greatest in PUDKH group, yielding a statistically significant difference ($^*P < 0.05$). (b) ILK expression and activity were evaluated in the injured skin treated with HGF gene transfection by immunohistochemical staining and compared to healthy skin from the control animals. ILK expression was observed in healthy skin and was especially abundant in the basal epidermis. After injury, an increase in ILK staining was observed in the wounded area of PUDKH group than in PBS, PUDK, and normal group. (d) A representative image is shown. The results of immunohistochemistry were analysed by Image-Pro Plus 5.0 image analyzer; we found that IOD value was greatest in PUDKH group than in the PBS, PUDK, and normal group, yielding a statistically significant difference ($^*P < 0.05$).

expressed blue fluorescent protein in epidermal prickle cell or in the granule cell layer during wound healing. Quantitative analysis using Image-Pro Plus 5.0 image analyzer revealed that blue fluorescent protein expression was significantly higher in PUDKH group than that of the three groups ($^*P < 0.05$) (Figure 2(c)). This indicated that $\beta_1$-integrin in epidermis on the edge of the round wound was significantly increased in HGF gene-transfer rats 7 days after gene transfer, compared with those in control rats. After 7 days, ILK expression and activity were evaluated in the injured skin by immunohistochemical staining and compared to healthy skin from the control animals. ILK expression was observed in healthy skin and was especially abundant in the basal epidermis. After injury, an increase in ILK staining was observed in the wounded area of the PUDKH group compared to the PBS, PUDK, and normal groups (Figures 2(b) and 2(d)).

*3.3. In Vitro Studies Using Primary Rat Epidermal Cells.* The possible role of HGF in accelerating the $\beta_1$-integrin and ILK expression was assayed by RNA interference *in vitro*. Primary rat epidermal cells were isolated from newborn rat by dispase/trypsin treatment and cultured as previously described [17]. Once 100% confluent, cells were scratched with a sterile needle and HGF protein (50 ng/mL) was added to the conditions medium for 48 h. As an alternative method to deplete ILK expression, small interference RNA (siRNA) knockdown experiments targeting ILK were also performed. After isolating the total cell lysates, western blotting was used to semiquantitatively analyze the ILK and c-met expression. An increase of ILK and c-met protein was observed in epidermal cells treated with HGF (50 ng/mL) for 48 h, while, after ILK siRNA, the expression of ILK and c-met in epidermal cells was downregulated despite the presence of HGF ($^*P < 0.05$, $^{**}P < 0.01$). And the expression of c-met (the receptor of HGF [16]) was also detected by immunofluorescence; results were similar with the above conclusion. $\beta_1$-Integrin and CK10 in scratched epidermal cells were measured by immunocytochemistry. Graph that represents densitometric analysis showed that HGF increased the expression of $\beta_1$-integrin and downregulated CK10 ($^*P < 0.05$, $^{**}P < 0.01$). In summary, HGF

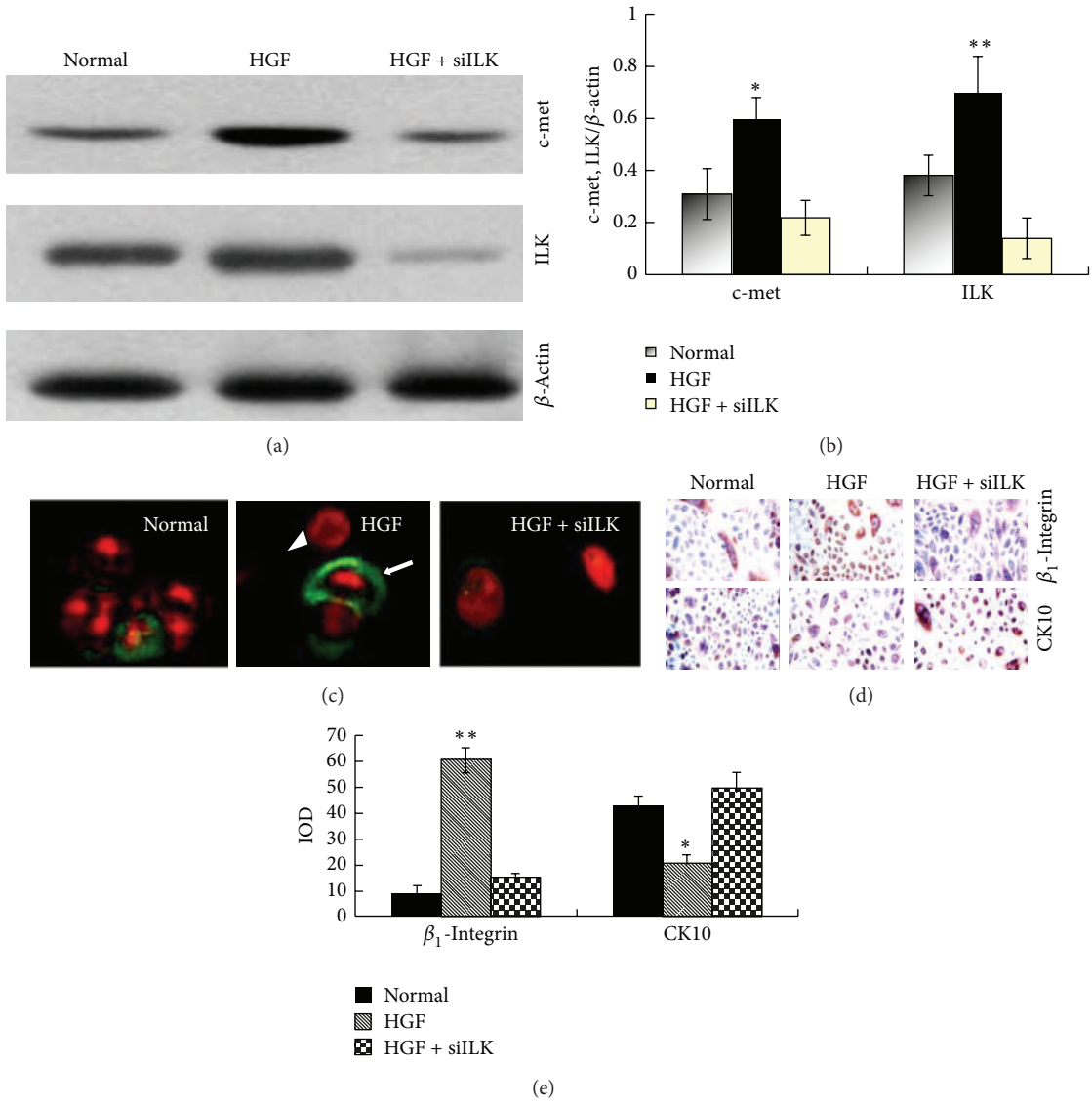



FIGURE 3: HGF promotes ILK and c-met expression in scratched epidermal cells. (a) ILK and c-met expression in cultured epidermal cells was analyzed by western blot 48 h after scratch wounding. An increase of ILK and c-met expression was observed in epidermal cells treated with HGF (50 ng/mL) for 48 h. After epidermal cells were transfected with ILK siRNA, the expression of ILK and c-met was downregulated, despite the presence of HGF protein (50 ng/mL). (b) Graph represents densitometric analysis of western blots described in (a) ($^{*}P < 0.05$, $^{**}P < 0.01$). (c) The distribution of c-met, the receptor of HGF, was detected by IF. The c-met protein at the surface of the epidermal cells is green (white arrow) and in the nuclei is red (white triangle). (d) $\beta_1$-Integrin and CK10 expression in scratched epidermal cells were tested by immunocytochemistry. HGF increased the expression of $\beta_1$-integrin and downregulated CK10. However, ILK siRNA-transfected cells showed a significant knockdown of $\beta_1$-integrin expression, with an increase in CK10 expression, compared to nontransfected or control cells. (e) Graph represents densitometric analysis of western blot described for (d) ($^{*}P < 0.05$, $^{**}P < 0.01$). The results revealed that there was a statistically significant difference between HGF group versus normal and HGF + siILK groups. All these suggested that HGF could activate the $\beta_1$-integrin and ILK signaling pathway; inversely, $\beta_1$-integrin and ILK might regulate the expression and function of HGF/c-met.

increased the expressions of c-met, ILK, and $\beta_1$-integrin but lowered the expression of CK10. However, ILK siRNA-transfected cells showed a significant decrease in ILK and $\beta_1$-integrin expression, with no increase in c-met expression observed after wounding, compared to nontransfected or control cells (Figure 3). All these suggested that HGF could activate the $\beta_1$-integrin and ILK signaling pathway. Inversely, the $\beta_1$-integrin/ILK pathways might regulate the expression and function of HGF/c-met. In next study, we used wound healing and cell migration assay to examine that whether HGF functions were affected after ILK gene depletion.

*3.4. ILK Depletion Inhibits Epidermal Cells Proliferation and Migration.* To investigate whether the observed defects in wound closure were due to deficiencies in cellular migration or proliferation, scratch wound assays were repeated.

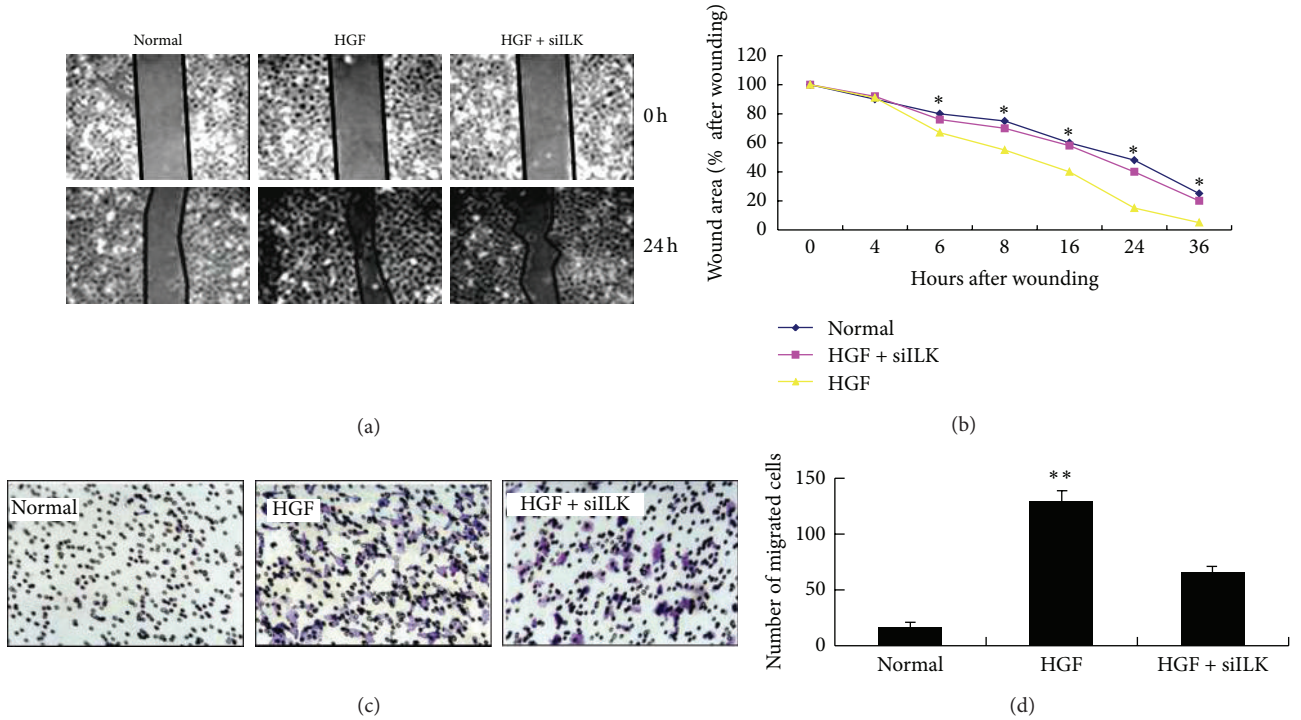



FIGURE 4: ILK knockdown epidermal cells, cell migration and invasion. (a) Epidermal cells treated with HGF (50 ng/mL) and ILK knockdown were grown to confluence and a wound was created. Photographs of wounds were captured at 0 or 36 h after wounding to determine the degree of wound closure. A representative experiment is shown. (b) Graphs represent the percentage of $t = 0$ h wound area at different times after wounding ($^{*}P < 0.05$ HGF group versus normal and HGF + siILK groups at matched time point). (c) ILK knockdown epidermal cells were seeded onto migration chambers in triplicate and were allowed to migrate for 2 hours. Cells that migrated through the membrane were fixed and visualized by crystal violet staining. (d) Graphs represent the amount of cells in normal, HGF, and HGF + siILK groups. After ILK knockdown, the migration and invasion ability of epidermal cells were obviously declined ($^{**}P < 0.01$ was regarded as statistically significant between HGF group versus normal and HGF + siILK groups).

The migratory ability of ILK knockdown epidermal cells during wound healing was measured using transwell migration chambers. Confluent cell monolayer wound healing requires cell migration from the leading edge and cellular proliferation to replace lost cells. Results showed differences between monolayer closure in the presence or absence of HGF protein and between epidermal cells treated with or without ILK siRNA-transfection (Figures 4(a) and 4(b)), indicating that the retarded closure is due to reduced proliferation of ILK-deficient cells. Cell invasiveness of ILK knockdown epidermal cells was assessed by seeding the cells onto matrigel-coated invasion chamber. In presence of HGF protein, more invaded cells were observed in invasion chamber, while significantly fewer cells were able to invade through the matrigel in ILK knockdown epidermal cells (Figures 4(c) and 4(d)). Taken together, these data demonstrate that ILK expression is necessary for HGF induction during wound healing *in vitro*.

## 4. Discussion

Skin is the biggest organ of the human being and has many functions, including covering the whole body and protecting other tissues and organs. Therefore, the healing of a skin wound involves an extraordinary mechanism of cascading cellular functions that is unique in nature [1, 2, 6]. The process requires the regulation of a variety of cell types and growth factors, including epidermal cells, stromal cells, epidermal growth factor (EGF), fibroblast growth factor (FGF), and HGF, each playing a critical role by mediating reepithelialization, epidermal cell differentiation, fibrosis, and angiogenesis [6]. HGF is a multifunctional mediator of these processes. The half-life of exogenous HGF protein is relatively short, and even repeated infusions of HGF fail to maintain high levels of HGF *in vivo* [7]. To better understand the effects of HGF on wound healing, the plasmids carrying the HGF gene (PUDKH) were locally injected into the trauma skin of diabetic rat. The degree of reepithelialization of wounds in HGF gene-transfer rats was significantly increased on 5, 7, 10, and 14 days than that of PBS and PUDK groups.

During wound-induced cell proliferation, the main focus of the healing process is to recover the wound surface. The proliferation and migration of epidermal cells are crucial for the rapid closure of the epidermis [1, 2]. The reepithelialization process is mediated by local epidermal cells at the wound edges and by epithelial stem cells from hair follicles or sweat glands [18–21], while, according to previous report, if the differentiated epidermal cells high expressed $\beta_1$-integrin, they would have a stronger ability to form clones and passage and generate a complete epithelium after

moving them to skin wounds [4, 5]. Therefore, high expressed $\beta_1$-integrins can prompt epidermal cells into a high proliferative and dedifferentiated state. In subsequent investigations, the model of aged epidermal cell dedifferentiation *in vivo* was constructed [22] and the dedifferentiation-derived stem cell-like cells were isolated [9].

In this study, we also found a significant increase in epidermal $\beta_1$-integrin and ILK expressions on the edge of the round wound in HGF gene-transfer rats 7 d after gene transfer relative to control rats. $\beta_1$-Integrins are essential for tissue development and maintenance [1, 21–26]. They are highly expressed on stem and progenitor cell populations, which orchestrate organogenesis and represent a cellular reservoir to maintain organ homeostasis [26, 27]. ILK, an integrin $\beta_1$-subunit binding protein, is an intracellular effector of cell-matrix interactions and regulates many cellular processes, including growth, proliferation, survival, differentiation, migration, invasion, and angiogenesis [7, 28, 29]. We believe that HGF promoted the transformation of epidermal cells into dedifferentiated stem-like cells, characteristic of the powerful ability of proliferation and migration. In fact, epidermal cell proliferation and migration are very important in facilitating epithelial wound repair. HGF appeared to be involved in the dedifferentiation of epidermal cells, accelerating wound healing in a $\beta_1$-integrin/ILK signaling pathway-related manner. Experiments *in vitro* verified these findings. The scratched epidermal cells showed ectopic expression of $\beta_1$-integrin and ILK under the role of HGF. However, under the same conditions, as in 50 ng/mL HGF, the expression of $\beta_1$-integrin and ILK was downregulated after ILK gene silencing by RNA interference, and c-met, the receptor of HGF, decreased within the epidermal cells plasma membrane. At the same time, wound healing and cell migration assays showed that the proliferation and migration ability of the epidermal cells were partly suppressed after ILK gene silencing. It must be noted that ILK deficiency leads to retarded wound closure in skin, while it is related to HGF expression after knockdown of ILK in Isabel Serrano's research report. And after exogenous administration of human HGF, alterations in cell proliferation and wound closure in ILK-deficient mouse embryonic fibroblast or mice could be observed [28, 29]. However, in our study, HGF can promote the expression of ILK in the process of wound healing, and the capacity of HGF to promote epidermal cell proliferation and migration will be affected after blocking ILK expression.

Our data demonstrate that HGF can accelerate wound healing by promoting the dedifferentiation of epidermal cells in a process closely related to the $\beta_1$-integrin and ILK signaling pathways. Notably, dedifferentiation is observed in a variety of processes such as cancer, organ regeneration, and stem cell renewal, but it has been difficult to study because there are few experimentally tractable systems to identify it. Particularly in skin, dedifferentiation of the epidermal cells is still a matter of some debate [8, 9]. Although the finding of the differentiated cell dedifferentiation was just preliminary, it was confirmed that HGF increased the expression of $\beta_1$-integrin in epidermal cells. Most importantly, HGF improved the proliferation and migration of epidermal cells. It is therefore possible that dedifferentiation-derived cells will be another source of epidermal stem cells for wound repair and regeneration in the future. Simultaneously, the potential powerful ability of HGF to promote skin wound healing should attract more attention.

## Conflict of Interests

There is no conflict of interests for any authors.

## Authors' Contribution

Jin-Feng Li and Hai-Feng Duan have contributed equally to this work.

## Acknowledgment

All experiments were reviewed by the Ethics Committee of the 148th Hospital.

## References

[1] K. P. Krafts, "Tissue repair: the hidden drama," *Organogenesis*, vol. 6, no. 4, pp. 225–233, 2010.

[2] J. Richards, "Stiffness in healing fractures," *Critical Reviews in Biomedical Engineering*, vol. 15, no. 2, pp. 145–185, 1987.

[3] S. Werner and R. Grose, "Regulation of wound healing by growth factors and cytokines," *Physiological Reviews*, vol. 83, no. 3, pp. 835–870, 2003.

[4] P. H. Jones, S. Harper, and F. M. Watt, "Stem cell patterning and fate in human epidermis," *Cell*, vol. 80, no. 1, pp. 83–93, 1995.

[5] G. E. Hannigan, C. Leung-Hagesteijn, L. Fitz-Gibbon et al., "Regulation of cell adhesion and anchorage-dependent growth by a new $\beta$1-integrin-linked protein kinase," *Nature*, vol. 379, no. 6560, pp. 91–96, 1996.

[6] K. Conway, P. Price, K. G. Harding, and W. G. Jiang, "The molecular and clinical impact of hepatocyte growth factor, its receptor, activators, and inhibitors in wound healing," *Wound Repair and Regeneration*, vol. 14, no. 1, pp. 2–10, 2006.

[7] K. Matsumoto and T. Nakamura, "Hepatocyte growth factor (HGF) as a tissue organizer for organogenesis and regeneration," *Biochemical and Biophysical Research Communications*, vol. 239, no. 3, pp. 639–644, 1997.

[8] K. Nakanishi, M. Uenoyama, N. Tomita et al., "Gene transfer of human hepatocyte growth factor into rat skin wounds mediated by liposomes coated with the Sendai virus (hemagglutinating virus of Japan)," *The American Journal of Pathology*, vol. 161, no. 5, pp. 1761–1772, 2002.

[9] M. S. Islam and H. Zhou, "Isolation and characterization of putative epidermal stem cells derived from Cashmere goat fetus," *European Journal of Dermatology*, vol. 17, no. 4, pp. 302–308, 2007.

[10] D. W. Tan, K. B. Jensen, M. W. Trotter, J. T. Connelly, S. Broad, and F. M. Watt, "Single-cell gene expression profiling reveals functional heterogeneity of undifferentiated human epidermal cells," *Development*, vol. 140, no. 7, pp. 1433–1444, 2013.

[11] C. Zhang, X. Fu, P. Chen et al., "Dedifferentiation derived cells exhibit phenotypic and functional characteristics of epidermal stem cells," *Journal of Cellular and Molecular Medicine*, vol. 14, no. 5, pp. 1135–1145, 2010.

[12] A. K. Sharma, S. Bharti, S. Ojha et al., "Up-regulation of PPAR$\gamma$, heat shock protein-27 and -72 by naringin attenuates insulin resistance, $\beta$-cell dysfunction, hepatic steatosis and kidney damage in a rat model of type 2 diabetes," *British Journal of Nutrition*, vol. 106, no. 11, pp. 1713–1723, 2011.

[13] Z. Zhang, M. Zhao, J. Wang, Y. Ding, X. Dai, and Y. Li, "Oral administration of skin gelatin isolated from chum salmon (*Oncorhynchus keta*) enhances wound healing in diabetic rats," *Marine Drugs*, vol. 9, no. 5, pp. 696–711, 2011.

[14] L. Guo, W. Yu, X. Li et al., "Targeting of integrin-linked kinase with a small interfering RNA suppresses progression of experimental proliferative vitreoretinopathy," *Experimental Eye Research*, vol. 87, no. 6, pp. 551–560, 2008.

[15] N. S. Tan, L. Michalik, N. Di-Poï, B. Desvergne, and W. Wahli, "Critical roles of the nuclear receptor PPAR$\beta$ (peroxisome-proliferator-activated receptor $\beta$) in skin wound healing," *Biochemical Society Transactions*, vol. 32, no. 1, pp. 97–102, 2004.

[16] T. Tarui, M. Majumdar, L. A. Miles, W. Ruf, and Y. Takada, "Plasmin-induced migration of endothelial cells: a potential target for the anti-angiogenic action of angiostatin," *The Journal of Biological Chemistry*, vol. 277, no. 37, pp. 33564–33570, 2002.

[17] M. Kan, G. Zhang, R. Zarnegar et al., "Hepatocyte growth factor/hepatopoietin A stimulates the growth of rat kidney proximal tubule epithelial cells (RPTE), rat nonparenchymal liver cells, human melanoma cells, mouse keratinocytes and stimulates anchorage-independent growth of SV-40 transformed RPTE," *Biochemical and Biophysical Research Communications*, vol. 174, no. 1, pp. 331–337, 1991.

[18] P. Martin, "Wound healing—aiming for perfect skin regeneration," *Science*, vol. 276, no. 5309, pp. 75–81, 1997.

[19] K. Lau, R. Paus, S. Tiede, P. Day, and A. Bayat, "Exploring the role of stem cells in cutaneous wound healing," *Experimental Dermatology*, vol. 18, no. 11, pp. 921–933, 2009.

[20] S. J. Miller, E. M. Burke, M. D. Rader, P. A. Coulombe, and R. M. Lavker, "Re-epithelialization of porcine skin by the sweat apparatus," *Journal of Investigative Dermatology*, vol. 110, no. 1, pp. 13–19, 1998.

[21] C. Roh and S. Lyle, "Cutaneous stem cells and wound healing," *Pediatric Research*, vol. 59, no. 4, part 2, pp. 100R–103R, 2006.

[22] H. Li, X. Fu, L. Zhang, T. Sun, and J. Wang, "*In vivo* dedifferentiation of human epidermal cells," *Cell Biology International*, vol. 31, no. 11, pp. 1436–1441, 2007.

[23] W. E. Lowry, L. Richter, R. Yachechko et al., "Generation of human induced pluripotent stem cells from dermal fibroblasts," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 8, pp. 2883–2888, 2008.

[24] J. Yu, M. A. Vodyanik, K. Smuga-Otto et al., "Induced pluripotent stem cell lines derived from human somatic cells," *Science*, vol. 318, no. 5858, pp. 1917–1920, 2007.

[25] A. Pivarcsi, M. Széll, L. Kemény, A. Dobozy, and Z. Bata-Csörgo, "Serum factors regulate the expression of the proliferation-related genes $\alpha$5 integrin and keratin 1, but not keratin 10, in HaCat keratinocytes," *Archives of Dermatological Research*, vol. 293, no. 4, pp. 206–213, 2001.

[26] X. Fu, X. Sun, and T. Sun, "Epidermal growth factor induce the epithelial stem cell island formation in the regenerated epidermis," *Zhonghua Yi Xue Za Zhi*, vol. 81, no. 12, pp. 733–736, 2001.

[27] P. C. McDonald, A. B. Fielding, and S. Dedhar, "Integrin-linked kinase—essential roles in physiology and cancer biology," *Journal of Cell Science*, vol. 121, no. 19, pp. 3121–3132, 2008.

[28] W. Xie, F. Li, J. E. Kudlow, and C. Wu, "Expression of the integrin-linked kinase (ILK) in mouse skin: loss of expression in suprabasal layers of the epidermis and up-regulation by erbB-2," *The American Journal of Pathology*, vol. 153, no. 2, pp. 367–372, 1998.

[29] I. Serrano, M. L. Díez-Marqués, M. Rodríguez-Puyol et al., "Integrin-linked kinase (ILK) modulates wound healing through regulation of hepatocyte growth factor (HGF)," *Experimental Cell Research*, vol. 318, no. 19, pp. 2470–2481, 2012.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 781618, 7 pages
http://dx.doi.org/10.1155/2013/781618

*Research Article*

# Multiple Biomarker Panels for Early Detection of Breast Cancer in Peripheral Blood

**Fan Zhang,[1,2] Youping Deng,[3] and Renee Drabier[1]**

[1] *Department of Academic and Institutional Resources and Technology, University of North Texas Health Science Center, Fort Worth, 76107, USA*
[2] *Department of Forensic and Investigative Genetics, University of North Texas Health Science Center, Fort Worth, 76107, USA*
[3] *Department of Internal Medicine Kidston House, Rush University Medical Center, 630 S. Hermitage Avenue, Room 408, Chicago, IL 60612, USA*

Correspondence should be addressed to Renee Drabier; renee.drabier@unthsc.edu

Received 12 September 2013; Accepted 8 November 2013

Academic Editor: Zhongming Zhao



Detecting breast cancer at early stages can be challenging. Traditional mammography and tissue microarray that have been studied for early breast cancer detection and prediction have many drawbacks. Therefore, there is a need for more reliable diagnostic tools for early detection of breast cancer due to a number of factors and challenges. In the paper, we presented a five-marker panel approach based on SVM for early detection of breast cancer in peripheral blood and show how to use SVM to model the classification and prediction problem of early detection of breast cancer in peripheral blood. We found that the five-marker panel can improve the prediction performance (area under curve) in the testing data set from 0.5826 to 0.7879. Further pathway analysis showed that the top four five-marker panels are associated with signaling, steroid hormones, metabolism, immune system, and hemostasis, which are consistent with previous findings. Our prediction model can serve as a general model for multibiomarker panel discovery in early detection of other cancers.

## 1. Introduction

Traditional methods mostly used for early detection have been regular and periodic self-examination and annual or biannual checkups using mammography and analysis of tissue biopsies. But mammography as a screening tool for early detection has many drawbacks. For example, mammography may not detect small tumors and is often unsatisfactory for younger women, who typically have dense breast tissue. And if a patient does have a suspicious mammogram, a biopsy will probably be done to make the diagnosis. Obtaining tissue biopsies can be difficult for several reasons, including small size of lump, lack of available medical facilities, and patients' reluctance to undergo invasive procedures due to scaring and costs.

In recent years, functional genomics studies using DNA microarrays have been shown effective in differentiating between breast cancer tissues and normal tissues by measuring thousands of differentially expressed genes simultaneously [1–3]. However, early cancer detection and treatment are still challenging. One reason is that obtaining tissue samples for microarray analysis can be still difficult. Another reason is that the signatures of gene expression difference between normal and cancer obtained in different studies are not sufficiently reproducible or informative to be prognostically useful, although they do give valuable insights into the pathogenesis and biology of human tumor metastasis [4]. Moreover, the fact that breast cancer is not a single homogeneous disease but consists of multiple disease status, each arising from a distinct molecular mechanism and having a distinct clinical progression path [5, 6], makes the disease difficult to detect in early stages.

To address these issues, a novel and minimally invasive test that uses easily obtained peripheral blood for breast

cancer detection has been developed [7, 8]. For example, Sharma et al. used microarrays and nearest-shrunken-centroid method to analyze the expression pattern of 1,368 genes in peripheral blood cells of 24 women with breast cancer and 32 women with no sign of this disease. The study found that a blood-based gene expression test can be developed to detect breast cancer early in asymptomatic patients [8]. Aarøe et al. collected peripheral blood from 67 breast cancer samples and 63 normal samples, identified a set of 738 differentially expressed probes, and achieved an estimated prediction accuracy of 79.5% with a sensitivity of 80.6% and a specificity of 78.3% [7].

There is a need for more reliable diagnostic tools for early detection of breast cancer in peripheral blood which can achieve high prediction accuracy with as few genes as possible and to reduce the required examination of a large number of genes which increases the dimensionality, computational complexity, and clinical cost of diagnosis [8]. Li estimated that five or six genes rather than 37 or 738 would be sufficient for the early detection of breast cancer, based on colon cancer, leukemia, and breast cancer [8]. Therefore, it is desirable to adopt a "multimarker panel" concept and nontrivial computational methods that can integrate microarray measurement of multiple differential gene expression levels between disease and controls to achieve good performance for clinical genomic development of biomarkers [9].

Support vector machine (SVM) has several unique characteristics as a research tool for prediction in cancer classification applications. One unique characteristic as a specific type of learning algorithm is that it is characterized by the capacity control of the decision function, the use of the kernel functions, and the sparsity of the solution [10]. The second unique characteristic of SVM is that it is established on the unique theory of the structural risk minimization principle to estimate a function by minimizing an upper bound of the generalization error and therefore very resistant to the overfitting problem, eventually achieving a high generalization performance. The third unique characteristic of SVM is that training SVM is equivalent to solving a linearly constrained quadratic programming problem so that the solution of SVM is always unique and globally optimal, unlike neural networks training, which requires nonlinear optimization with the danger of getting stuck at local minima.

For classification and prediction of breast cancer samples, these unique characteristics make SVM appealing as compared with regression-based models or neural network as seen in [11–13]. For example, Liu et al. used SVM to predict the state of breast cancer and found that SVM outperformed K-means cluster and artificial neural network [11]. Henneges et al. applied oscillating search algorithm for feature selection (OSAF) to iteratively improve features for training of Support vector machines (SVM) to better predict breast cancer [12]. They selected 35 out of 51 nucleosides/ribosylated metabolites in the urine of breast cancer women and controls by LC-ITMS coupling for subsequent computational analyses, and they identified 44 pairwise ratios of metabolite features by iterative optimization of SVM. Liu et al. combined genetic algorithm (GA) and all paired (AP) support vector machine (SVM) methods to determine the predictive features for multiclass breast cancer categorization [13].

Table 1: Statistics of samples.

| | #health | #cancer | #total |
|---|---|---|---|
| Training group | 32 | 34 | 66 |
| Testing group | 31 | 33 | 64 |
| Total | 63 | 67 | 130 |

There has not been any report until this study that applied SVM to the development of multimarker panels for early detection of breast cancer based on peripheral blood. Based on a neural network approach to multibiomarker panel development for LC/MS/MS proteomics profiles we developed [14], we propose for the first time a multimarker panel development solution for early detection of breast cancer in peripheral blood by using a SVM and show how to use SVM to model the classification and prediction problem of early detection of breast cancer in peripheral blood.

## 2. Methods and Materials

*2.1. Peripheral Blood Data Collection.* The peripheral blood data are publicly available through the GEO database with the accession number GSE16443 [7] and were collected with the purpose of determining the potential of gene expression profiling of peripheral blood cells for early detection of blood cancer. It consists of 130 samples with 67 cases and 63 controls. We downloaded the 130 samples which contain 32,879 probes. Then we randomly divided the 130 samples into two groups: group A as a training group and group B as a testing group (Table 1).

*2.2. Normalization.* Per sample normalization was performed to normalize for staining intensity variations among samples. All expression data on a sample were normalized to the 50th percentile of log base 2 of all values on that sample. First, log ratio base 2 transformation was used to transform the data. And then for each probe the median of the log summarized values from all the samples was calculated and subtracted from each of the samples.

*2.3. Linear Mixed Model.* We used the ABI Human Genome Survey Microarray Version 2 to manage and map probe IDs. A full factorial model was used to represent the fixed effect and the random effect which are used to account for group and probe. The expression log ratios value is the final quantity that is fit by a separate analysis of the variance (ANOVA) statistical model for each probe as $y_{ij}$ using the following:

$$y_{ij} = \mu + T_i + S_j + \varepsilon_{ij}, \quad (1)$$

where $S_j \in N(0, \sigma_1^2)$, $\varepsilon_{ij} \in N(0, \sigma^2)$. Here, $\mu$ is the mean expression value, $T_i$ is the fixed group effect (caused by the experimental conditions or treatments being evaluated), $S_j$ is the random sample effect (random effects from either individual biological samples or sample preparations), and $\varepsilon_{ij}$ is the within-groups errors. All random effects are assumed

independent of each other and independent of the within-groups errors $\varepsilon_{ij}$.

*2.4. Statistics.* Statistical significance was measured by a three-step method. First, we conducted the above linear mixed model to obtain the $P$ value of the significance for the group effect. Then we calculated the FDR adjusted $P$ value. Last, we calculated the FDR $q$ value using the Storey-Tibshirani method [15]. We chose a significance screening filter ($q < 0.01$) to select genes of which we estimated significant differences in the health and breast cancer samples.

*2.5. Support Vector Machine Analysis.* The classification problem of breast cancer can be restricted to consideration of the two-class problem without loss of generality (breast cancer and normal). We used a support vector machine- (SVM-) based methods [16] to develop the classifier for breast cancer from peripheral blood. And then we applied the classifier to predict blind dataset of breast cancer from peripheral blood.

For the use of the support vector machine as an appropriate tool for prediction of the breast cancer, a three-way data split is applied for training, validation, and testing. The training set is used for learning to fit the parameters of the classifier. The validation set is used to tune the parameters of the classifier. And the testing set is used only to assess the performance of the fully-trained classifier. We first randomly split the data into two groups: group A (training group) and group B (testing group), with roughly equal size. Then we use the $k$-fold cross validation on the training group to find the "optimal" parameters for the classifier. Group A is randomly partitioned into $k$ subsamples. For each subsample, a cross section of the data is flagged for use as the *validation set*, and a new model is created by training on the remaining data which are the *training set* and not in the subsample. The cross validation process is then repeated $k$ times (the folds), with each of the $k$ subsamples used exactly once as the validation data. The $k$ results from the folds then can be averaged to produce a single estimation. The testing group is used as testing set.

We chose each combination of $N$ ($N = 5$ for five-marker panel) out of all the 42 genes differentially expressed in the training group as inputs to the SVM. In order to find the optimal classifier, we presented an optimization method that measures the area under the curve (AUC) for receiver operating characteristics (ROC). In this scheme, we first train SVM for each combination in the training set with 5-fold cross validation. Then, we measured the AUC for each combination in the validation set. Lastly, the optimal combination $C^*$ was determined by

$$C^* = \underset{C}{\operatorname{argmax}} \operatorname{AUC}\left(\mathrm{SVM}_C, V\right), \tag{2}$$

where AUC is the area under the ROC curve of SVM prediction, SVM is the support vector machine, $C$ is combination of picking five out of the 42 genes, and $V$ is the validation set of training group.

Fivefold cross validation was used to increase the number of estimates and improve the accuracy of the prediction model by avoiding the overfitting. In 5-fold cross validation, the original sample is randomly partitioned into 5 subsamples. Of the 5 subsamples, a single subsample is retained as the validation data for testing the model, and the remaining 4 subsamples are used as training data. The cross validation process is then repeated 5 times, with each of the 5 subsamples used exactly once as the validation data. The 5 results from the folds then can be averaged to produce a single estimation. The advantage of this method over repeated random subsampling is that all observations are used for both training and validation, and each observation is used for validation only once.

*2.6. Pathway Analysis.* The Integrated Pathway Analysis Database (IPAD) (http://bioinfo.hsc.unt.edu/ipad/) [17] is used for pathway analysis.

## 3. Results

We downloaded from the Gene Expression Omnibus (accession number GSE16443) [7] the 130 samples with 67 breast cancer and 63 healthy women. After we randomly divided the 130 samples into two groups, group A as training group and group B as testing group (Table 1), we obtained 32 healthy samples and 34 cancer samples in the training set and validation set and 31 healthy samples and 33 cancer samples in the testing set.

We obtained 42 markers in the training group with $q$ value <0.01. No data from the testing set were utilized in (1) identification of peripheral blood markers or (2) development of the SVM model.

An SVM model with 5-fold cross validation was built on all 42 markers in the training group. We obtained a high performance (AUC = 1.0, precision = 94.4%, accuracy = 97.0%, sensitivity = 100.0%, and specificity = 93.8%) for the training group but a low performance (AUC = 0.58, precision = 58.3%, accuracy = 57.8%, sensitivity = 63.6%, and specificity = 51.6%) for the testing group (Figure 1). The result shows that using all markers as a predictor can improve the prediction accuracy only for training group but not for the testing group. Therefore, we constructed an SVM with 5-fold cross validation for each combination of five out of 42 markers and trained with breast cancer cells in peripheral blood derived from 34 women diagnosed with breast cancer and 32 control women in the training group. The three-way data split was applied for training, validation, and testing. The optimal combinations were obtained by our optimization model based on the training set and validation set in the training group.

Training of the SVM was performed using radius basis function (RBF) kernels function and five-fold cross validation. Receiver operating characteristic (ROC) curve and area under curve (AUC) were calculated to help evaluating the predictive performance of the SVM. We choose $N = 5$ for five-marker panel because (1) our pilot study shows that five markers can be enough to achieve a satisfied performance for prediction and classification of cancer [14], (2) previous papers from other labs estimated that five or six genes would be sufficient for the early detection of breast cancer [18], and

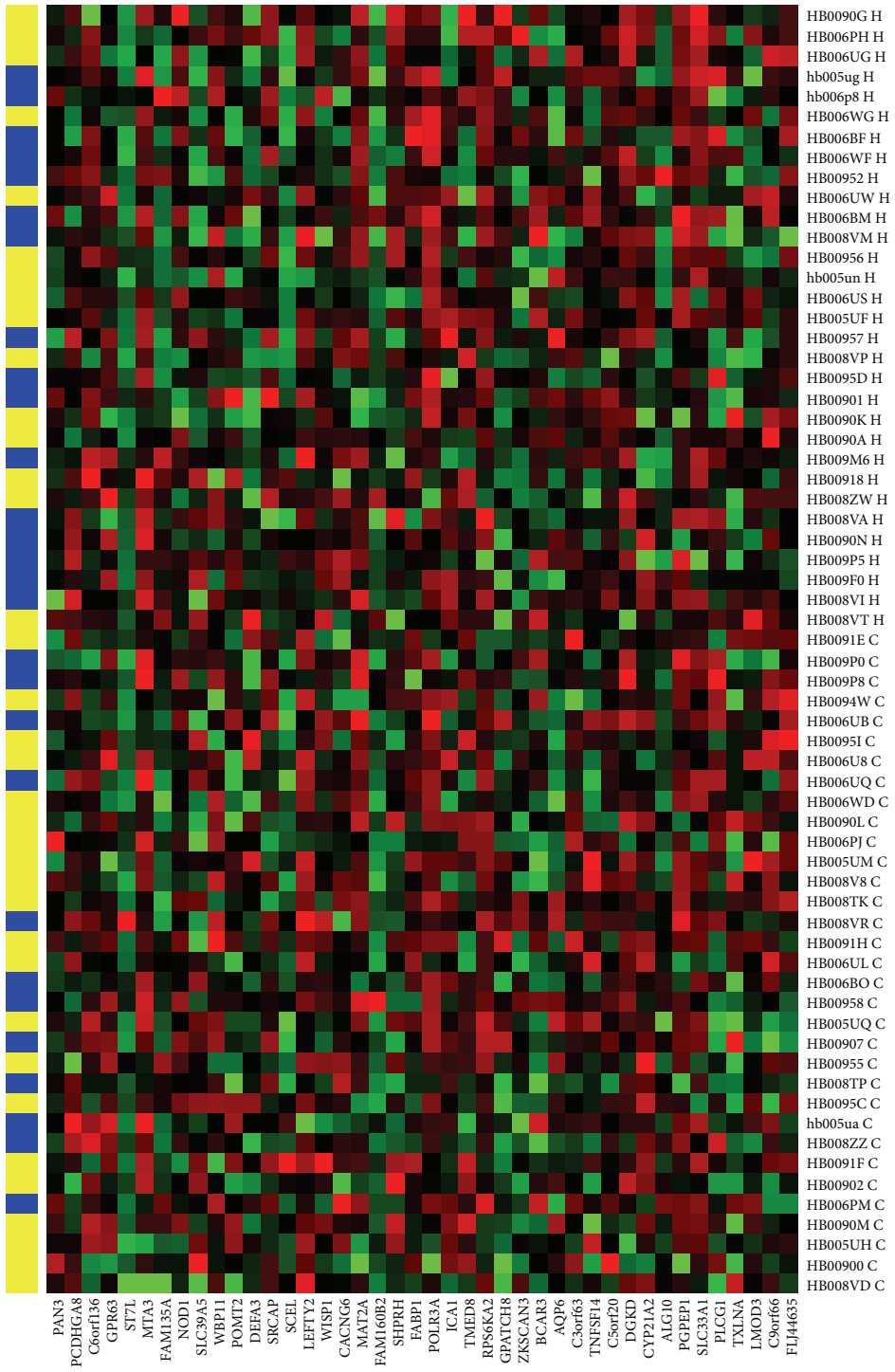



FIGURE 1: 42 biomarkers predicting the healthy and breast cancer samples in testing set. $X$-axis is the 42 biomarkers. $Y$-axis shows the 33 breast cancer and 31 healthy samples (H: healthy, blue; C: cancer, yellow).

(3) we expect to achieve high prediction accuracy for breast cancer with as few genes/proteins as possible.

In order to validate our prediction method, we compared the ROCs for the best four 5-marker panel predictions determined by our method with the ROCs for four randomly chosen 5-marker panels from 42 candidate biomarkers (Figure 2). As shown in the Figure 2, the top four best predictions determined by our method (solid lines) have better sensitivity-specificity-tradeoff performance than those chosen randomly from 42 candidate biomarkers.

In Table 2, we show the best four five-marker panels identified, using the SVM. Two genes, BCAR3 and LEFTY2,

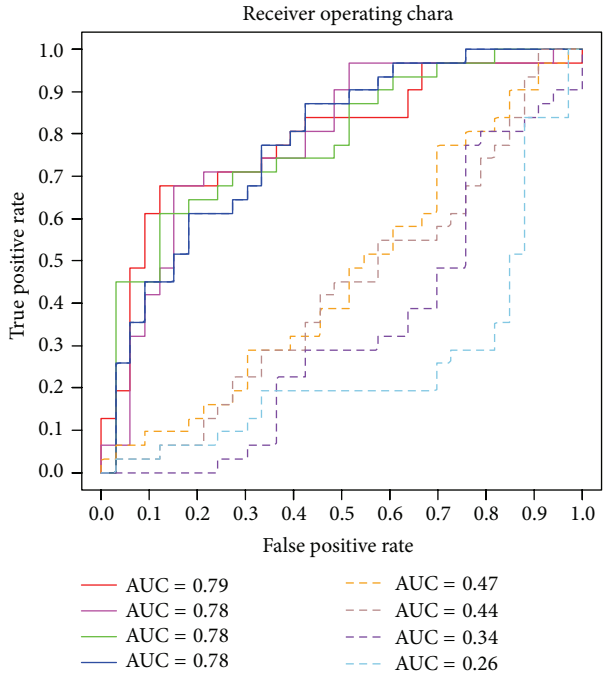



Figure 2: A comparison of best four 5-marker panel ROCs (solid lines) and randomly chosen four (out of 42 candidates) 5-marker ROCs (dotted lines).

Table 2: Best four five-marker panels identified.

| Panel | Training group AUC | Testing group AUC |
|---|---|---|
| PCDHGA8; LEFTY2; CACNG6; BCAR3; CYP21A2 | 0.9053 | 0.7879 |
| PCDHGA8; DEFA3; SCEL; LEFTY2; BCAR3 | 0.9127 | 0.7830 |
| DEFA3; SCEL; LEFTY2; CACNG6; BCAR3 | 0.9154 | 0.7801 |
| DEFA3; LEFTY2; CACNG6; BCAR3; DGKD | 0.8897 | 0.7801 |

are in common between the best four five-marker panels. Two genes, CACNG6 and DEFA3, are shown three times, and two genes, PCDHGA8 and SCEL, are shown twice.

Pathway analysis shows that the pathways linked with the best four five-marker panels are signaling, steroid hormones, metabolism, immune system, and hemostasis (Table 3), which are consistent with previous findings [7].

The confusion matrix and common performance metrics for both the training group and testing group for the best five-marker panel are shown in the Table 4. Although the final accuracy is 68.75%, it can be considered as an improvement if compared to the original accuracy 58.81%. In addition, the AUC, a comprehensive measurement of sensitivity and specificity, is improved markedly from 0.5826 to 0.7879 (Figure 2 and Table 4).

We further evaluated our multimarker panel prediction performance by comparing our results with prediction performance in previously published findings. Sharma et al.

Table 3: Pathway analysis for the best four five-marker panels.

| Pathway ID | Pathway name | Molecule |
|---|---|---|
| 200071 | Regulation of CDC42 activity | BCAR3 |
| hsa04260 | Cardiac muscle contraction | CACNG6 |
| hsa05412 | Arrhythmogenic right ventricular cardiomyopathy (ARVC) | CACNG6 |
| hsa05410 | Hypertrophic cardiomyopathy (HCM) | CACNG6 |
| hsa05414 | Dilated cardiomyopathy | CACNG6 |
| hsa04010 | MAPK signaling pathway | CACNG6 |
| 194002 | Glucocorticoid biosynthesis | CYP21A2 |
| 193993 | Mineralocorticoid biosynthesis | CYP21A2 |
| 211976 | Endogenous sterols | CYP21A2 |
| 209943 | Steroid hormones | CYP21A2 |
| 196071 | Metabolism of steroid hormones and vitamins A and D | CYP21A2 |
| 211897 | Cytochrome P450, arranged by substrate type | CYP21A2 |
| 211945 | Phase 1, functionalization of compounds | CYP21A2 |
| 211859 | Biological oxidations | CYP21A2 |
| hsa00140 | Steroid hormone biosynthesis | CYP21A2 |
| 556833 | Metabolism of lipids and lipoproteins | CYP21A2 |
| 1430728 | Metabolism | CYP21A2 |
| 1462054 | Alpha-defensins | DEFA3 |
| 1461973 | Defensins | DEFA3 |
| hsa05202 | Transcriptional misregulation in cancer | DEFA3 |
| 168249 | Innate immune system | DEFA3 |
| 168256 | Immune system | DEFA3 |
| 114508 | Effects of PIP2 hydrolysis | DGKD |
| hsa00561 | Glycerolipid metabolism | DGKD |
| hsa04070 | Phosphatidylinositol signaling system | DGKD |
| hsa00564 | Glycerophospholipid metabolism | DGKD |
| 416476 | G alpha (q) signalling events | DGKD |
| 388396 | GPCR downstream signaling | DGKD |
| 372790 | Signaling by GPCR | DGKD |
| 162582 | Signal transduction | DGKD |
| 76002 | Platelet activation, signaling, and aggregation | DGKD; LEFTY2 |
| 109582 | Hemostasis | DGKD; LEFTY2 |
| 1433617 | Regulation of signaling by NODAL | LEFTY2 |
| 1181150 | Signaling by NODAL | LEFTY2 |
| 114608 | Platelet degranulation | LEFTY2 |
| 76005 | Response to elevated platelet cytosolic Ca2+ | LEFTY2 |
| hsa04350 | TGF-beta signaling pathway | LEFTY2 |
| 1266738 | Developmental biology | LEFTY2 |

identified a panel of 37 genes that permitted early detection with the classification accuracy of 82% [8], and Aarøe et al. identified a set of 738 differentially expressed probes that achieved an estimated prediction accuracy of 79.5% with a

Table 4: Prediction result for the best 5-marker panel.

| Predicted | Training group | | Testing group | |
|---|---|---|---|---|
| | Cancer | Normal | Cancer | Normal |
| Cancer | 29 | 6 | 21 | 8 |
| Normal | 5 | 26 | 12 | 23 |
| Precision | | 82.86% | | 72.41% |
| Accuracy | | 83.33% | | 68.75% |
| Sensitivity | | 85.29% | | 63.64% |
| Specificity | | 81.25% | | 74.19% |

sensitivity of 80.6% and a specificity of 78.3% [7]. Considering that their methods were not applied to independent testing group randomly separated from training group but used $k$-fold cross validation where the original sample was randomly partitioned into $k$ subsamples and of the $k$ subsamples, a single subsample was retained as the validation data for testing the model, and the remaining $k-1$ subsamples were used as training data, our prediction performance actually outperformed them. When we applied an SVM and 5-fold cross validation with our best 5-marker panel to the training group of 34 women with breast cancer and 32 healthy women controls, we obtained a higher performance than these previously published findings (precision = 82.86%, accuracy = 83.33%, sensitivity = 85.29%, and specificity = 81.25%, Table 4). We believe that our approach is a significant success, considering that we only used five gene markers in a panel to achieve the prediction performance (AUC = 0.7879, precision = 72.41%, accuracy = 68.75%, sensitivity = 63.64%, and specificity = 74.19%).

## 4. Discussions

In this study, we incorporate the use of a three-way data split in combination with an enumeration method based on SVM. It is a reasonably straightforward application of existing methods and achieves substantially higher prediction performance. In our three-way data split, the testing set is used for the purpose of independent testing only and the validation set is used for tuning the parameters in the SVM training. Splitting the data three ways to get training, validation, and testing sets actually makes our approach very close to real applications. We cannot always select markers based on testing data because in most real applications the testing data are blind or unknown pending for prediction. The prediction performance of the testing set in a three-way data split can actually reflect the outcome in a real application. The best model selected from the training group may not produce the best prediction performance in the testing data due to the inconsistence between the training data and testing data. However, our results show that the selected top models will produce acceptable performance in the testing set, although not best performance.

Although some other researches achieved higher performance, for example, 82% by Sharma et al. [8] and 79.5% by Aarøe et al. [7], our prediction result outperforms theirs if we use training group only (precision = 82.86%, accuracy = 83.33%, sensitivity = 85.29%, and specificity = 81.25%, Table 4) as they did. Our prediction performance which is more close to a real application is actually based on the testing set which is totally blind to the training group (precision = 72.41%, accuracy = 68.75%, sensitivity = 63.64%, and specificity = 74.19%, Table 4).

One limitation of the three-way data split is the sample size. If we split a small size sample into three ways, we would end up with so little data in each set that our analysis would lack any power. If we identify the inconsistence of prediction performance between the validation set and testing set, we can increase the size of training group (training set and validation set) and decrease the size of testing set by simply moving some samples in the testing set to the training group.

Since our approach enumerates all possible combinations of 5 out of $N$ markers, there is a limitation for the size of $N$ due to current computational capability. In our talon supercomputer, it would take about 1 hour to calculate all combinations of 5 out of 32 markers and about two weeks to finish the computation of picking 5 out of 100 markers. It is acceptable for us to set the maximum of $N$ to be 100 because in most cases the top 100 markers can be both specific and sensitive in understanding the treatment, diagnosis, and prognosis of cancer and can be limited by setting a reasonable $P$ value threshold.

An ANOVA statistical model is used for identifying differentially expressed genes between cancer and normal samples. For a simple two-group comparison, we would get the identical result if we were to compare the two groups using ANOVA, $t$-test, or SAM. However, ANOVA is a much more flexible and powerful technique that can be applied to much more complex research issues with multiple factors than the other two methods. For example, for the peripheral blood data, we should take into account two factors: (1) the fixed group effect (caused by the experimental conditions or treatments being evaluated) and (2) the random sample effect (random effects from either individual biological samples or sample preparations). In this case, ANOVA method is more efficient than multiple two-group studies analyzed via $t$-test or SAM, because with fewer observations we can gain more information.

In this work, we use the support vector machine (SVM) for classification, which is in general believed to outperform the other classification methods such as the logistic regression (LR) and the artificial neural networks (ANN) [19, 20], because the SVM prediction improves LR and ANN significantly along the specificity axis [21]. However, we understand that for special problems the ANN may still yield reasonable results and that the conclusion that SVM outperforms ANN is in general from a theoretical perspective and in particular for the considered case study [22]. Therefore, we strongly suggest that the tree-way data split method should be carried out for this kind of comparison before we reach any conclusions.

## 5. Conclusions

We developed an integrated computational approach that addressed a challenging multipanel biomarker development problem in the early detection of breast cancer in peripheral blood. The approach that we used combined simple statistical

filtering of ANOVA with an optimization model of SVM. The approach automatically learned nonlinear relationships between features and outcomes to generate predictive models, which achieved AUC = 0.7879 performance with a sensitivity of 63.64% and a specificity of 74.19% in the testing data set of 33 women with breast cancer and 31 healthy women controls. The SVM combined with the AUC optimization method is capable of identifying the optimal combination of multimarkers for performance comparable to that of conventional medical decision support systems. We believe that this computational approach works well with early detection of breast cancer in peripheral blood and can provide general guidance for future molecular medicine multimarker panel discovery applications in other diseases. In the future, we will follow up with biological experiments to validate these biomarkers with our collaborators.

## Conflict of Interests

All authors declare that there is no conflict of interests.

## Authors' Contribution

Renee Drabier conceived the initial work and designed the method. Fan Zhang generated the datasets, developed the statistics method, and performed the statistical analyses of the case studies. Youping Deng validated markers for early detection of breast cancer in peripheral blood. All authors are involved in the drafting and revisions of the paper.

## Acknowledgment

This work was supported by the bioinformatics program in the University of North Texas Health Science Center.

## References

[1] X. Hu, Y. Zhang, A. Zhang et al., "Comparative serum proteome analysis of human lymph node negative/positive invasive ductal carcinoma of the breast and benign breast disease controls via label-free semiquantitative shotgun technology," *OMICS: A Journal of Integrative Biology*, vol. 13, no. 4, pp. 291–300, 2009.

[2] B. A. Zeidan, R. I. Cutress, N. Murray et al., "Proteomic analysis of archival breast cancer serum," *Cancer Genomics and Proteomics*, vol. 6, no. 3, pp. 141–148, 2009.

[3] A. Lebrecht, D. Boehm, M. Schmidt, H. Koelbl, R. L. Schwirz, and F. H. Grus, "Diagnosis of breast cancer by tear proteomic pattern," *Cancer Genomics and Proteomics*, vol. 6, no. 3, pp. 177–182, 2009.

[4] M. Suzuki and D. Tarin, "Gene expression profiling of human lymph node metastases and matched primary breast carcinomas: clinical implications," *Molecular Oncology*, vol. 1, no. 2, pp. 172–180, 2007.

[5] K. Polyak, "Breast cancer: origins and evolution," *The Journal of Clinical Investigation*, vol. 117, no. 11, pp. 3155–3163, 2007.

[6] F. Zhang and J. Y. Chen, "Discovery of pathway biomarkers from coupled proteomics and systems biology methods," *BMC Genomics*, vol. 11, supplement 2, article S12, 2010.

[7] J. Aarøe, T. Lindahl, V. Dumeaux et al., "Gene expression profiling of peripheral blood cells for early detection of breast cancer," *Breast Cancer Research*, vol. 12, no. 1, article R7, 2010.

[8] P. Sharma, N. S. Sahni, R. Tibshirani et al., "Early detection of breast cancer based on gene-expression patterns in peripheral blood cells," *Breast Cancer Research*, vol. 7, no. 5, pp. R634–644, 2005.

[9] A. Vlahou, C. Laronga, L. Wilson et al., "A novel approach toward development of a rapid blood test for breast cancer," *Clinical Breast Cancer*, vol. 4, no. 3, pp. 203–209, 2003.

[10] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 988–999, 1999.

[11] H. X. Liu, R. S. Zhang, F. Luan et al., "Diagnosing breast cancer based on support vector machines," *Journal of Chemical Information and Computer Sciences*, vol. 43, no. 3, pp. 900–907, 2003.

[12] C. Henneges, D. Bullinger, R. Fux et al., "Prediction of breast cancer by profiling of urinary RNA metabolites using Support Vector Machine-based feature selection," *BMC Cancer*, vol. 9, article 104, 2009.

[13] J. J. Liu, G. Cutler, W. Li et al., "Multiclass cancer classification and biomarker discovery using GA-based algorithms," *Bioinformatics*, vol. 21, no. 11, pp. 2691–2697, 2005.

[14] F. Zhang and J. Y. Chen, "A Neural network approach to multi-biomarker panel development based on LC/MS/MS proteomics profiles: a case study in breast cancer," in *Proceedings of the 22nd IEEE International Symposium on Computer-Based Medical Systems (CBMS '09)*, pp. 1–6, August 2009.

[15] J. D. Storey and R. Tibshirani, "Statistical significance for genomewide studies," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, pp. 9440–9445, 2003.

[16] D. Meyer, F. Leisch, and K. Hornik, "The support vector machine under test," *Neurocomputing*, vol. 55, no. 1-2, pp. 169–186, 2003.

[17] F. Zhang and R. Drabier, "IPAD: the integrated pathway analysis database for systematic enrichment analysis," *BMC Bioinformatics*, vol. 13, supplement 15, article S7, 2012.

[18] W. Li, "How many genes are needed for early detection of breast cancer, based on gene expression patterns in peripheral blood cells?" *Breast Cancer Research*, vol. 7, no. 5, p. E5, 2005.

[19] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*, Cambridge University Press, Cambridge, Mass, USA, 2000.

[20] J. V. Tu, "Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes," *Journal of Clinical Epidemiology*, vol. 49, no. 11, pp. 1225–1231, 1996.

[21] F. Dal Moro, A. Abate, G. R. G. Lanckriet et al., "A novel approach for accurate prediction of spontaneous passage of ureteral stones: support vector machines," *Kidney International*, vol. 69, no. 1, pp. 157–160, 2006.

[22] A. Abate, F. Dal Moro, and G. R. G. Lanckriet, "Response to 'support vector machines versus artificial neural network: who is the winner?'," *Kidney International*, vol. 71, no. 1, pp. 84–85, 2007.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 637424, 8 pages
http://dx.doi.org/10.1155/2013/637424

*Research Article*

# Expression Sensitivity Analysis of Human Disease Related Genes

**Liang-Xiao Ma,[1] Ya-Jun Wang,[1] Jing-Fang Wang,[1] Xuan Li,[1,2] and Pei Hao[1,3,4]**

[1] *Shanghai Center for Bioinformation Technology, Shanghai 201203, China*
[2] *Key Laboratory of Synthetic Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China*
[3] *Key Laboratory of Systems Biomedicine (Ministry of Education), Shanghai Jiao Tong University, Shanghai 200240, China*
[4] *Pathogen Diagnostic Center, Institute Pasteur of Shanghai, Chinese Academy of Sciences, Shanghai 200031, China*

Correspondence should be addressed to Jing-Fang Wang; jfwang8113@gmail.com and Pei Hao; phao@sibs.ac.cn

Received 23 August 2013; Accepted 11 October 2013

Academic Editor: Zhongming Zhao



*Background*. Genome-wide association studies (GWAS) have shown its revolutionary power in seeking the influenced loci on complex diseases genetically. Thousands of replicated loci for common traits are helpful in diseases risk assessment. However it is still difficult to elucidate the variations in these loci that directly cause susceptibility to diseases by disrupting the expression or function of a protein currently. *Results*. We evaluate the expression features of disease related genes and find that different diseases related genes show different expression perturbation sensitivities in various conditions. It is worth noting that the expression of some robust disease-genes doesn't show significant change in their corresponding diseases, these genes might be easily ignored in the expression profile analysis. *Conclusion*. Gene ontology enrichment analysis indicates that robust disease-genes execute essential function in comparison with sensitive disease-genes. The diseases associated with robust genes seem to be relatively lethal like cancer and aging. On the other hand, the diseases associated with sensitive genes are apparently nonlethal like psych and chemical dependency diseases.

## 1. Introduction

To elucidate the etiology and pathogenesis of the diseases, scientists have made efforts to map human disease loci genetically and clone many diseases genes [1, 2]. Recently with the development of new sequencing technology and high throughput microarray technology, the searching for the genetic traits of the diseases and scanning of personal genomic variations are revolutionized. Genome-wide association studies (GWAS) [3] showed its strong abilities in detecting complicated genetic variations in genes and building genomic variation patterns compared with linkage analysis [4] and candidate gene studies [5]. GWAS have made contribution to establish plenty of disorder-gene association pairs [6]. As reported, over 4008 SNPs are associated with 819 common diseases [7]. The Genetic Association Database (GAD) [8] is a valuable resource of human genetic association studies on complex diseases and disorders, which facility us to rapidly establish the relationship of disorder-gene association pairs. The association studies explained the relationships between the diseases and genes on the genomic levels. Diseases-associated studies have identified functional genetic variations, but they didn't well make sure the variations could cause the diseases directly. Annotating the diseases associated variations on different levels are necessary to identify the outstanding risking genes. Here we wonder whether the genes associated the same diseases show similar expression features?

In order to probe the expression feature of disease genes which are detected by the associated studies, we referred to the method of gene expression sensitivity analysis [9]. We firstly investigated human global gene expression characters in response to the environmental perturbation. Gene expression patterns are different in various biological conditions, a lot of case-control expression patterns have been profiled using the high throughput microarray technology. Studies show a group of genes' expression that could be easily disturbed with various external stimulations [9, 10]; however, some genes are stably expressed in different environments, which indicate that the genes show different expression

sensitivity. For example, housekeeping genes which have been well investigated in maintaining the basal cellular functions have revealed its expression stability [11, 12]. Recently gene expression sensitivity to external stimulation have been studied in yeast and human [9, 10], which guide us to generate the idea to investigate the expression sensitivity of diseases genes.

It is worthy to obtain a global view of the intrinsic properties of human disease gene expression as a response to perturbations. Gene Expression Omnibus (GEO) database [13] and Genetic Association Database (GAD) [8] were used to analyze the human associated gene expression sensitivity. A meta-analysis method could be used to seek the sensitive genes and robust genes in the expression profiles globally. Based on our calculation of sensitive values of gene expression, we firstly categorized the genes into robust and sensitive groups. Furthermore we investigated the expression sensitivity of disease related genes in response to the perturbations and found some of genes were detected by the association studies previously, but the expression of these genes is relatively stable in their corresponding disease studies. The results also indicate some diseases related genes that show their expression robustness (DGR) like cancer and aging genes, and chemical dependency disease related genes seem to be relatively sensitive (DGS).

## 2. Materials and Computational Methods

*2.1. Data Collection and Preprocessing.* The Genetic Association Database (GAD) includes over 80,000 gene records of genetic association studies. Importantly, the database has a designation of whether the gene record was reported to be associated with disease phenotype. The option Y means that the gene of record was associated with the disease phenotype; otherwise, the option N was not associated. We collected the records in GAD that associated with the disease phenotype and got the records only annotated with the standard disease phenotype keywords from MeSH (http://www.nlm.nih.gov/mesh/) vocabulary.

After filtering, 13277 records were used for further investigation. In our study, 2588 disease related genes were acquired from 13275 records from the Genetic Association Database. Among these genes, 1804 genes were associated with more than one disease and 784 were associated with only one disease (Supplementary Material available online at http://dx.doi.org/10.1155/2013/637424, Table S2). These diseases related genes are mainly from human genetic association studies of complex diseases and disorders. 1464 kinds of diseases were extracted from 13275 records. To establish the relationships between genes and their corresponding diseases groups, the 1464 kinds of diseases were divided into 16 groups by paring database.

We downloaded the HGU133plus2.0 microarray datasets from the GEO database, each dataset had been normalized with MAS5 when the authors submitted them into the database as required (http://www.ncbi.nlm.nih.gov/geo/). To calculate the expression level of each gene, we referred to the methods from the previous work [9]. We discarded the data sets with less than 6 arrays and changed the expression values into 10 if the expression values are less than 10. Then the expression values of all probes were logarithmic transformed (base 2). We choose the maximum expression value as gene expression value if multiple probes illustrate the same gene expression.

*2.2. Calculate Sensitive Values (SV) of Each Gene.* In our study, 167 datasets (labeled as $M$ in the Formula below) in GEO were used to calculate the sensitive values of genes. In each dataset $j$, we calculate the standard deviation (SD) and mean of each gene ($g_i$), getting the coefficient of variation (CV) of each gene ($g_i$) with SD divided mean. The CV of each gene in each dataset is calculated as follows:

$$\mathrm{CV}_{g_{ij}} = \frac{\mathrm{SD}\left(g_{ij}\right)}{\mathrm{mean}\left(g_{ij}\right)}. \quad (1)$$

In order to merge the results of different datasets and minimum experimental variation during sensitivity analysis, we employed the large scale meta-analysis method reported in our previous work [9]. We ranked the CV of genes in each dataset and constructed a matrix of ranked CV to datasets. Sensitive values (SV) are calculated as mean of each ranked CVs of all datasets:

$$\mathrm{SV}_{g_i} = \frac{\sum_{j=1}^{M} \mathrm{rank}\left(\mathrm{CV}_{g_{ij}}\right)}{M}. \quad (2)$$

*2.3. Defining Sensitive Genes and Robust Genes.* In the current study, we tried to establish the relationships between the different kinds of disease and different sensitive genes. Firstly we selected a group of genes whose expression could be significantly disturbed and a group of genes whose expression significantly stable. After calculating the sensitive values of genes, we used the 5 percent as the cutoff value. The top 5 percent of genes are significantly sensitively expressed, we took five percent of genes with lowest sensitive values as robust genes groups, and five percent of genes with highest SV were considered as sensitive genes.

## 3. Results

The Affymetrix HGU 133a plus 2.0 microarray covers 31835 genes, which represents more genes than the HGU133a microarray does. The distribution of SV is skewed normal distribution (Figure 1). The figure suggested that there are more genes with a comparatively lower sensitive value than those with higher sensitive values indicating existing more robust genes than sensitive genes. Although the distribution of sensitive values are skewed normal, we took five percent of genes with lowest sensitive values as robust genes groups, and also five percent of genes with highest SV were considered as sensitive genes. Therefore, we got 1592 robust genes and sensitive genes individually. Therefore, we got 1592 robust genes and 1592 sensitive genes. After comparison with disease related genes, we got 131 diseases related robust genes and 467 disease related sensitive genes respectively (Supplementary Table S1).

Table 1: Enriched biological process. The table shows the main biological process that the robust genes and sensitive genes engaged. Most robust genes take part in the translational elongation and viral transcription; however, the sensitive genes prefer to respond to the progesterone receptor stimulation and regulation.

| Robust genes | | Sensitive genes | |
|---|---|---|---|
| Enriched biological process | $P$ value | Enriched biological process | $P$ value |
| Translational elongation | $3.12E-41$ | Progesterone receptor signaling pathway | $1.74E-4$ |
| Viral transcription | $5.42E-39$ | Negative regulation of osteoclast differentiation | $7.11E-4$ |
| Translational termination | $8.26E-38$ | Cell maturation | $7.14E-4$ |
| Protein complex disassembly | $8.26E-38$ | Golgi vesicle transport | $9.89E-4$ |

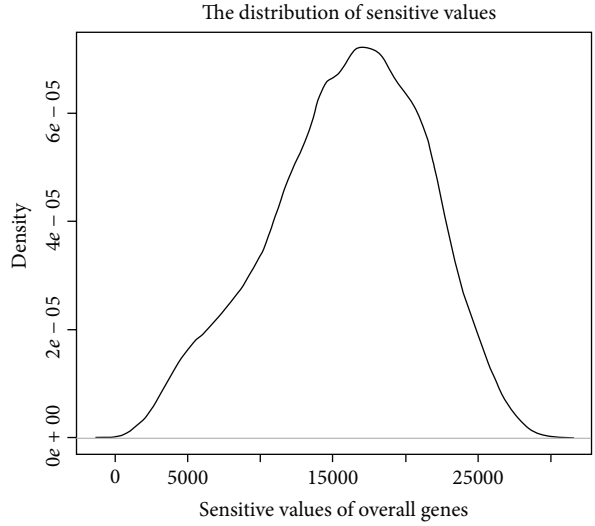



Figure 1: This figure demonstrates the distribution of the average rank order of gene expression standard deviations. The distribution of SV is skewed normal distribution.

*3.1. Functional Annotations of Expression Robust and Sensitive Genes.* In order to identify the biological functions in the cell, we conducted GO enrichment analysis [14, 15]. The results show that robust genes are strongly engaged in the cellular component organization (GO: 0071842), reproductive process (GO: 0022414), and viral reproduction (GO: 0016032) (Figure 2 and Table 1). However the sensitive genes play important roles in progesterone receptor signaling pathway (GO: 0050847) and negative regulation of osteoclast differentiation (GO: 0045671) (Figure 3 and Table 1). Furthermore enrichments analysis for cell components modules indicates that the robust gene are strongly enriched in the intracellular (GO: 0005622) and ribosome (GO: 0005840) (Figure 3), but the sensitive genes didn't show enrichment in the cell components. Based on the above observation, we conclude that the robust genes are engaged in the basic biological process.

*3.2. A Case about Robust Genes Consistently Expressing in Their Corresponding Disease.* We have found the expression of robust genes are stable in various conditions and inferred that the some disease related robust genes may express consistently in their corresponding diseases conditions. To verify our assumption, we took colorectal cancer and its associated robust genes HIF1A and MLH as an example.

HIF1A associated with colorectal cancer is one of robust genes. Hypoxia-inducible factor-1 (HIF1) is a heterodimer composed with HIF1A and HIF1B. HIF1 is functionally important in cellular and systemic homeostatic responses to hypoxia and initially found as transcription factor in mammalian cells cultured under reduced oxygen tension [16]. Fransén et al. found that polymorphic alleles in the gene of HIF1A show significant higher risk for the development of ulcerative colorectal cancer, which indicates that the HIF1A polymorphisms display their importance in the development of ulcerative intestinal tumors [17]. To view the gene expression variations of colorectal cancer, we took a case from a work [18] that analyzed expression changes in early onset colorectal cancer (GDS2609). The expression of HIF1A didn't show significant change in the study (Figure 4(a)). The robust gene HIF1A might be ignored in the colorectal cancer expression analysis.

Another robust gene MLH1 involved in DNA mismatch repair is also associated with colorectal cancer [19]. Liu et al. revealed that colorectal cancer is associated with 2 missense mutations in exon 16 of the MLH1 [20]. Chan et al. described a novel germline 1.8-kb deletion involving of the MLH1 gene associated with hereditary nonpolyposis colorectal cancer in a Hong Kong family [21]. Recently Nejda et al. suggests that gender should be considered in colorectal

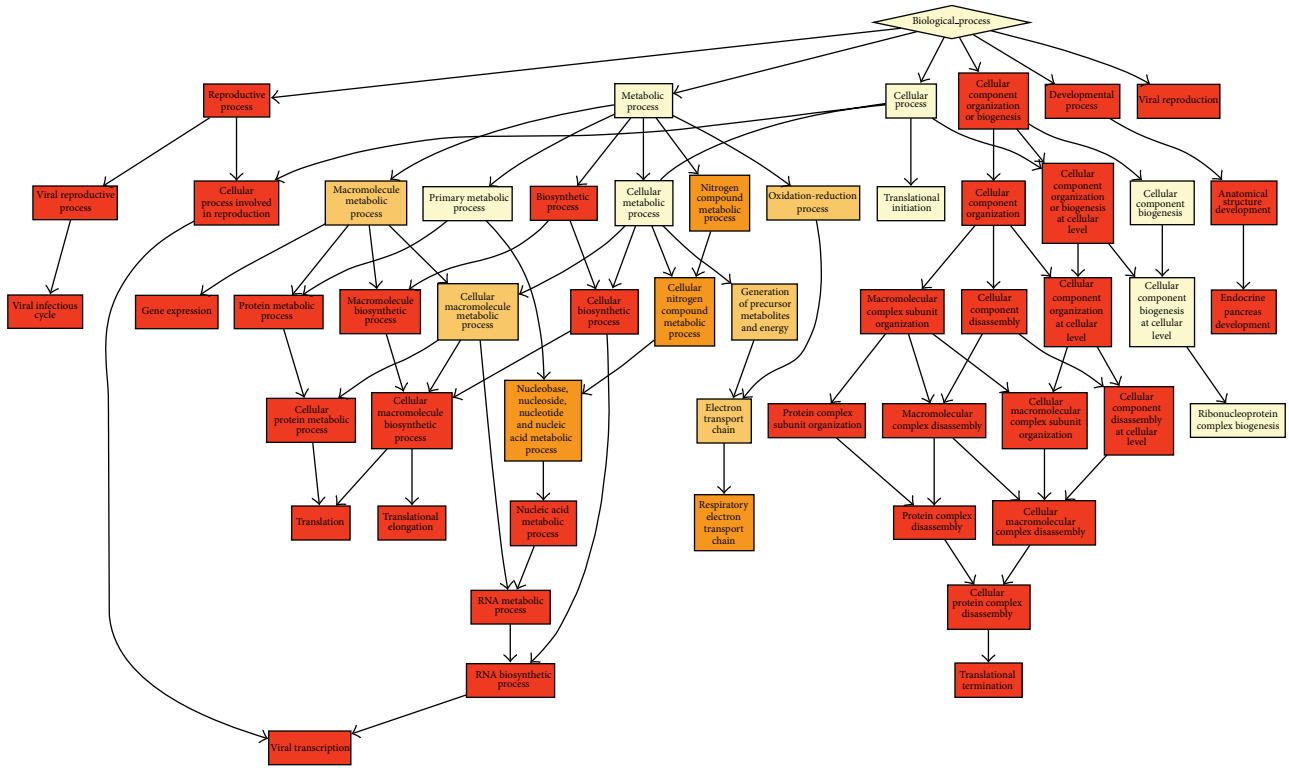



Figure 2: The biological process GO enrichment graph illustrates that the robust genes play basic roles of cell developments. The color in the rectangle is close to red; the genes are more enriched in that GO module, and the white color indicates the least enrichment in that module.

cancer association studies [22]. They found that nucleotide polymorphism in MLH1 displays a higher risk in sporadic colorectal carcinogenesis especially in men. A mechanism of genomic instability has been identified in colorectal cancer [23], the DNA mismatch repair genes MLH1 inactivated by hypermethylation of their promoter could cause microsatellite instability. We also found the expression of MLH1 in the early onset colorectal cancer investigation (GDS2609) is not significantly changed (Figure 4(b)). The expression profile analysis may overlook importance of robust gene MLH1. Therefore, we believe that diseases related robust genes are easily ignored in the expression analysis. In order to know the robust genes are enriched in what kinds of diseases, we established the relationships between diseases and gene expression sensitivities.

*3.3. The Relationships between Diseases and Gene Expression Sensitivity.* The 1469 diseases with their associated genes from GAD were divided into 16 groups. The disease genes are classified as robust disease gene groups and sensitive diseases genes groups based on whether they were included in the robust genes groups or sensitive genes groups. We performed the enrichment analysis of disease genes based on the hyper geometric distribution. The results (Table 2) show that the cancer, aging, and pharmacogenomics are enriched with robust genes (DGR), with $P$ values of 0.001309, 0.025063, and 0.06734 individually, and psych, chemical dependency, and reproduction are enriched with sensitive genes (DGS), with corresponding $P$ values of 0.001954, 0.028318, and 0.055457.

We annotated the gene from DGR and DGS with Gene Ontology [24]. Firstly the DGR and GDS were classified into seven groups (Figure 5(a)) according to whether the genes associated one or more diseases. In the DGR groups, cancer genes are mainly engaged in the mismatch repair (GO: 0006298), DNA catabolic process (GO: 0006308), and base excision repair (GO: 0006284). The GO analysis of DGS (Figure 5(b)) shows the genes only associated with psych diseases are enriched in the learning (GO:0007612) process and the genes associated both psych and chemical dependency diseases are engaged in the dopamine secretion (GO:0014046) and gamma-aminobutyric acid signaling pathway (GO:0007214) and so forth. We conclude that the DGR mainly engaged in more essential biological process compared with DGS which mainly involve in the regulation and response process.

## 4. Discussion and Conclusion

We parsed the GAD databases and selected diseases associated genes. Because the human genes show different expression sensitivity in response to the environmental perturbation [9], we evaluated the expression features of diseases genes with the method of gene expression sensitivity analysis. Because the finding of expression of robust genes is not easily changed in various biological conditions, we assumed that the disease related robust genes might be expressed stably in their corresponding disease conditions. The colorectal cancer associated robust genes HIF1A and MLH did not show significant

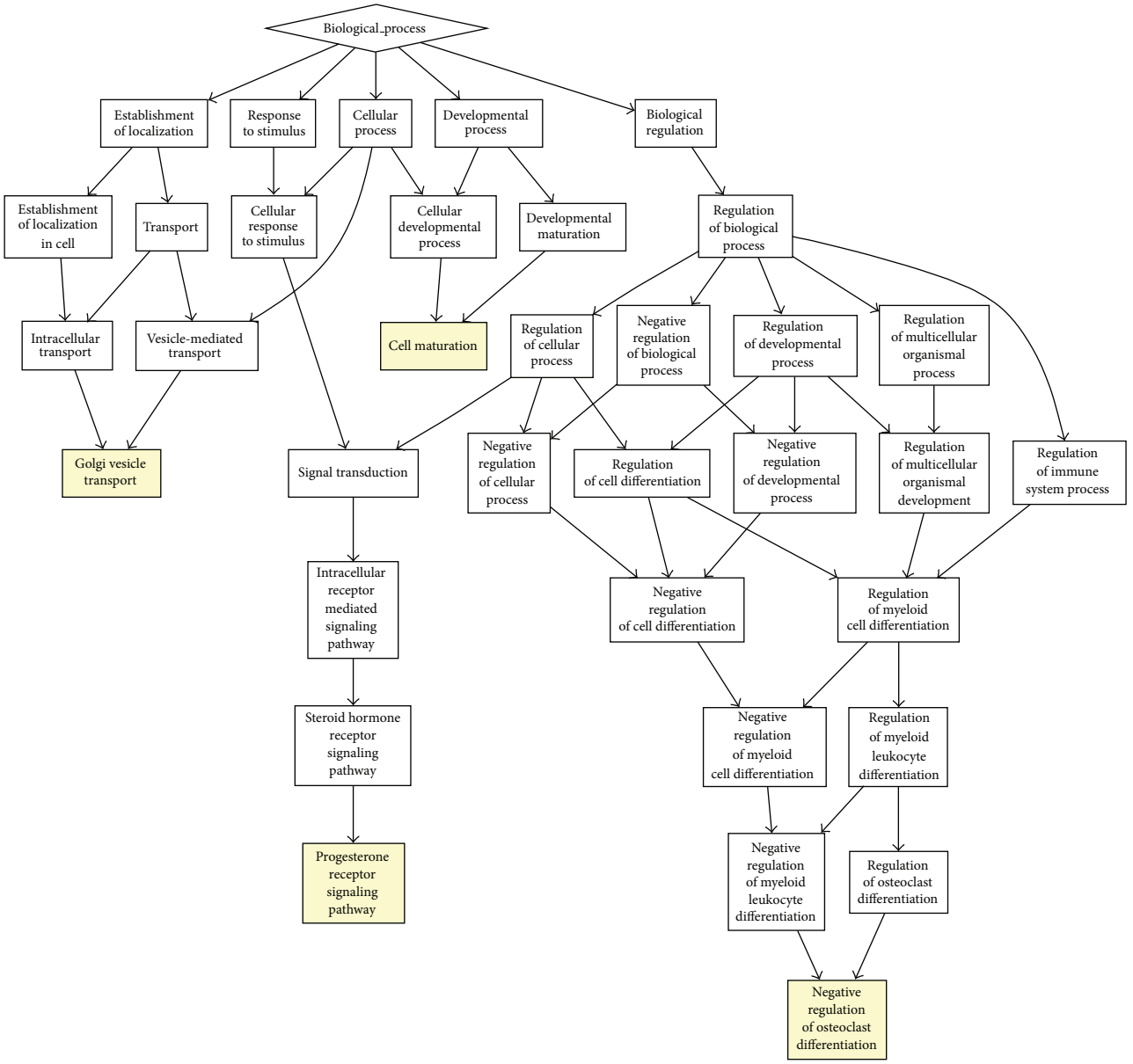



Figure 3: The biological process GO enrichment graph illustrates that the sensitive genes mainly regulate the cell metabolism. The sensitive genes are enriched in the GO module with a yellow color.

Table 2: Expression sensitivity analysis of diseases genes. The cancer, aging, and pharmacogenomic related genes reveal their robustness, and psych chemdependency and reproduction associated genes show their sensitiveness.

| Disease | Robustness | Sensitiveness | Disease | Sensitiveness | Robustness |
|---|---|---|---|---|---|
| Cancer | $1.31E-03$ | $7.08E-01$ | Psych | $1.95E-03$ | $8.09E-01$ |
| Aging | $2.51E-02$ | $4.87E-01$ | Chemdependency | $2.83E-02$ | $4.32E-01$ |
| Pharmacogenomic | $6.73E-02$ | $9.44E-01$ | Reproduction | $5.55E-02$ | $1.85E-01$ |

expression changes in the studies of colorectal cancer [17, 18, 20, 22]. Importantly our results suggested the genes associated with different diseases also reveal different sensitivities. We found the cancer, aging, and pharmacogenomics related genes display expression robustness, and psych, chemical dependency, and reproduction-associated genes are relatively sensitive. In our study, the robust disease related genes were investigated not only by combining the gene ontology but also by grouped disease information. The defect of robust genes could cause more lethal diseases, such as cancer and aging

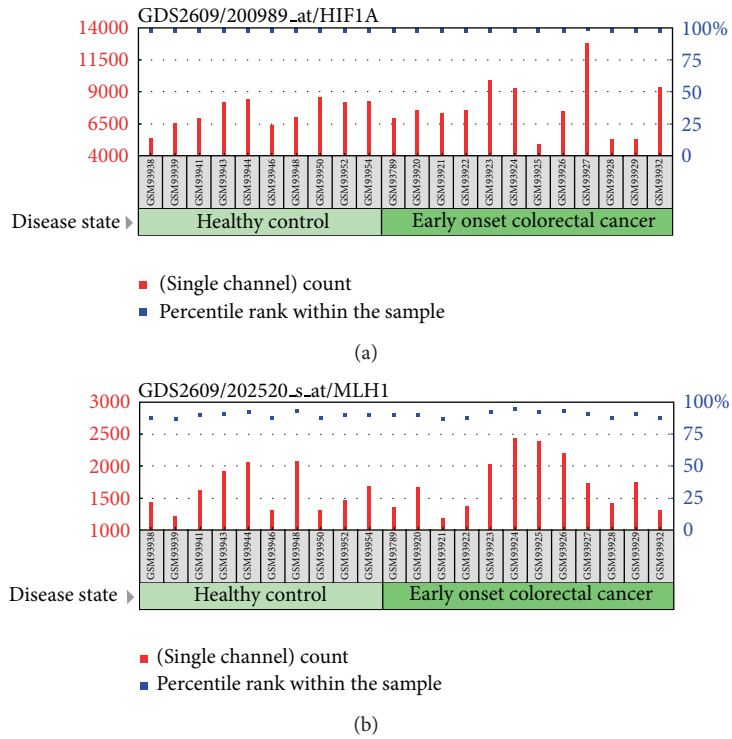



Figure 4: (a) The expression profile of gene HIF1A in GDS2609. (b) The expression profile of MLH1 in GDS2609.

diseases. Thus diseases related robust genes might play more essential roles to keep health for human.

Additionally, the protein interaction network and gene ontology provide extensive information to detail the relationships between different diseases genes. It was found that the structure of a cellar network and its functional properties were connected with protein or "Hubs" which are more likely encoded by essential genes [25, 26]. Human robust genes are higher degree centrality than the random groups of genes in the protein interaction network [9]. Gene ontology analysis also indicates the robust genes play an essential role in the cellar biology. The robust genes have shown its importance in different levels. Recent studies reveal that some kinds of diseases genes potentially encode hubs [27, 28]. Goh et al. suggested that cancer genes are more likely to encode hubs in the human disease networks and show higher coexpression with the rest of the genes in the cell [29], which means that cancer genes play critical roles in cellar development and growth. Age-related diseases tend to attack the center of the human protein network [30]. PPI network investigation above indicated that the cancer and aging related genes are potentially robust.

Based on the studies above, we believe that the different diseases genes reveal distinct expression sensitivity. Diseases that are associated with robust genes seem to be lethal, and the diseases associated with sensitive genes are nonlethal apparently. The gene ontology analysis indicates the robust genes are more essential when compared with sensitive genes. The robust genes those stably express in various environmental conditions are easily ignored in the expression analysis. Therefore the consideration of sensitivity of disease genes might be greatly helpful in elucidating of etiology and pathogenesis of the diseases. In practice, calculation of the diseases genes' sensitive values could be used to predict the potential harm to heath. In addition, if a robust gene is a potential drug target, it would have little therapeutic effects to these diseases by disturbing the expression level of the robust genes.

## Conflict of Interests

The authors declare that they have no conflict of interests.

## Authors' Contribution

Liang-Xiao Ma and Ya-Jun Wang statistical analyses for the study and drafted the paper. Liang-Xiao Ma, Ya-Jun Wang, and Jing-Fang Wang participated in the design of the study and provided guidance. Pei Hao and Xuan Li conceived the study and finalized the organization and contents of the paper. All authors approved the final paper.

## Acknowledgments

This work was supported by the National Basic Research Program of China (973 Program, nos. 2012CB517905 and 2012CB316501), Shanghai Natural Science Foundation (no. 11ZR1425700), and the National Natural Science Foundation

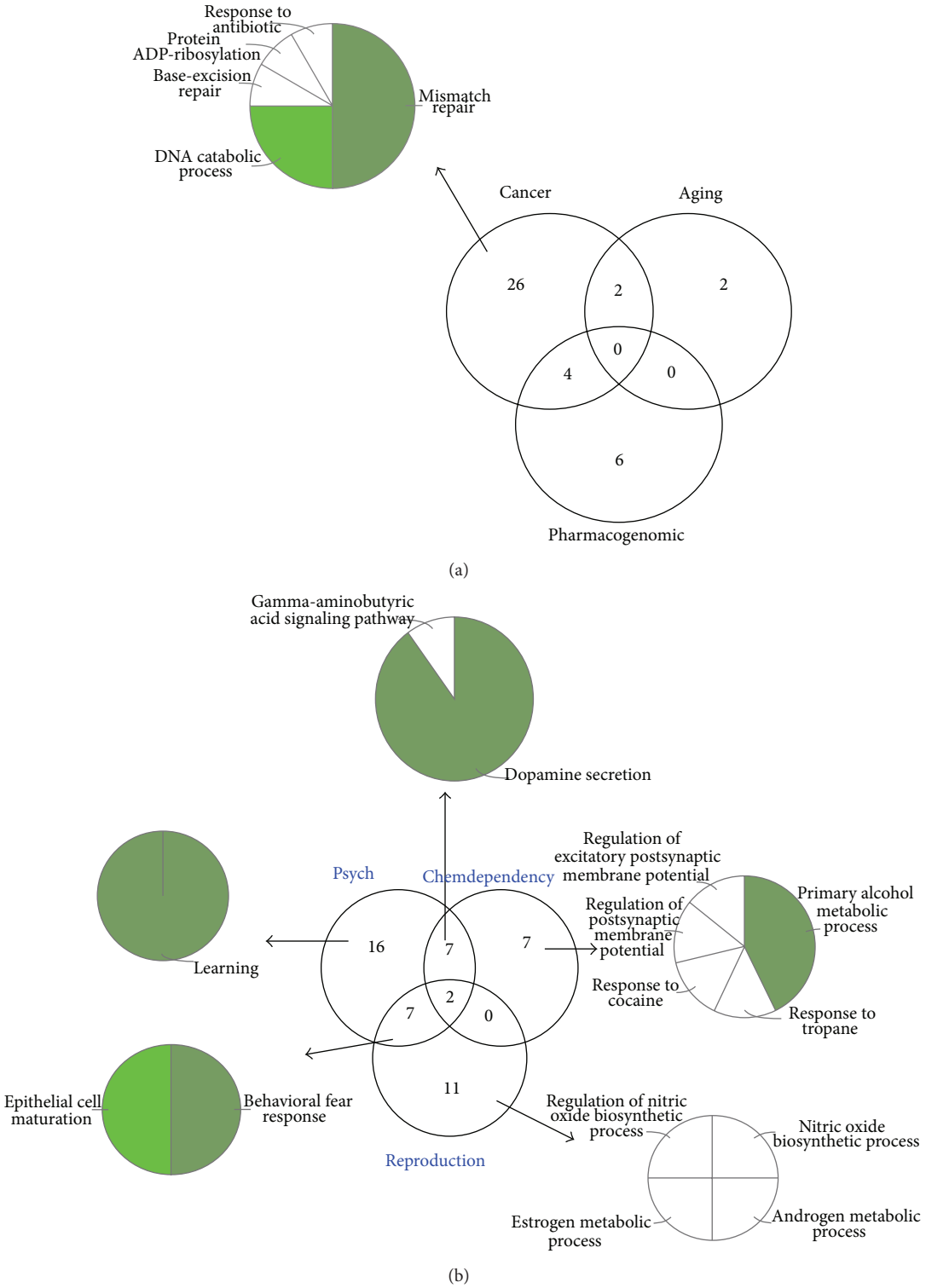



Figure 5: (a) Three kinds of disease genes show significant robustness. GO ontology analysis of cancer associated genes are on the left of graph. Cancer related robust genes apparently are more essential in the biological process. (b) Three kinds of diseases associated genes are relatively sensitive. The diseases associated with sensitive genes seem to be nonlethal.

of China (nos. 31200547 and 90913009). The authors gratefully acknowledge the support of the SA-SIBS Scholarship Program.

## References

[1] T. A. Manolio, "Genomewide association studies and assessment of the risk of disease," *The New England Journal of Medicine*, vol. 363, no. 2, pp. 166–176, 2010.

[2] U. Broeckel and N. J. Schork, "Identifying genes and genetic variation underlying human diseases and complex phenotypes via recombination mapping," *Journal of Physiology*, vol. 554, no. 1, pp. 40–45, 2004.

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

[4] N. Risch and K. Merikangas, "The future of genetic studies of complex human diseases," *Science*, vol. 273, no. 5281, pp. 1516–1517, 1996.

[5] J. Hardy and A. Singleton, "Genomewide association studies and human disease," *The New England Journal of Medicine*, vol. 360, no. 17, pp. 1759–1768, 2009.

[6] G. Jimenez-Sanchez, B. Childs, and D. Valle, "Human disease genes," *Nature*, vol. 409, no. 6822, pp. 853–855, 2001.

[7] L. A. Hindorff, P. Sethupathy, H. A. Junkins et al., "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 23, pp. 9362–9367, 2009.

[8] K. G. Becker, K. C. Barnes, T. J. Bright, and S. A. Wang, "The genetic association database," *Nature Genetics*, vol. 36, no. 5, pp. 431–432, 2004.

[9] P. Hao, S. Zheng, J. Ping et al., "Human gene expression sensitivity according to large scale meta-analysis," *BMC Bioinformatics*, vol. 10, no. 1, article S56, 2009.

[10] J. H. Ohn, J. Kim, and J. H. Kim, "Genomic characterization of perturbation sensitivity," *Bioinformatics*, vol. 23, no. 13, pp. i354–i358, 2007.

[11] Z. Tu, L. Wang, M. Xu, X. Zhou, T. Chen, and F. Sun, "Further understanding human disease genes by comparing with housekeeping genes and other genes," *BMC Genomics*, vol. 7, no. 1, article 31, 2006.

[12] E. Eisenberg and E. Y. Levanon, "Human housekeeping genes are compact," *Trends in Genetics*, vol. 19, no. 7, pp. 362–365, 2003.

[13] 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, no. 1, pp. D760–D765, 2007.

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

[15] E. Eden, R. Navon, I. Steinfeld, D. Lipson, and Z. Yakhini, "GOrilla: A tool for discovery and visualization of enriched GO terms in ranked gene lists," *BMC Bioinformatics*, vol. 10, no. 1, article 48, 2009.

[16] G. L. Wang, B.-H. Jiang, E. A. Rue, and G. L. Semenza, "Hypoxia-inducible factor 1 is a basic-helix-loop-helix-PAS heterodimer regulated by cellular O2 tension," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 92, no. 12, pp. 5510–5514, 1995.

[17] K. Fransén, M. Fenech, M. Fredrikson, C. Dabrosin, and P. Söderkvist, "Association between ulcerative growth and hypoxia inducible factor-1$\alpha$ polymorphisms in colorectal cancer patients," *Molecular Carcinogenesis*, vol. 45, no. 11, pp. 833–840, 2006.

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

[19] N. Papadopoulos, N. C. Nicolaides, Y.-F. Wei et al., "Mutation of a mutL homolog in hereditary colon cancer," *Science*, vol. 263, no. 5153, pp. 1625–1629, 1994.

[20] T. Liu, P. Tannergård, P. Hackman et al., "Missense mutations in hMLH1 associated with colorectal cancer," *Human Genetics*, vol. 105, no. 5, pp. 437–441, 1999.

[21] T. L. Chan, S. T. Yuen, J. W. C. Ho et al., "A novel germline 1.8-kb deletion of hMLH1 mimicking alternative splicing: a founder mutation in the Chinese population," *Oncogene*, vol. 20, no. 23, pp. 2976–2981, 2001.

[22] N. Nejda, D. Iglesias, M. Moreno Azcoita, V. Medina Arana, J. J. González-Aguilera, and A. M. Fernández-Peralta, "A MLH1 polymorphism that increases cancer risk is associated with better outcome in sporadic colorectal cancer," *Cancer Genetics and Cytogenetics*, vol. 193, no. 2, pp. 71–77, 2009.

[23] K. Söreide, E. Janssen, H. Söiland, H. Körner, and J. Baak, "Microsatellite instability in colorectal cancer," *British Journal of Surgery*, vol. 93, no. 4, pp. 395–406, 2006.

[24] G. Bindea, B. Mlecnik, H. Hackl et al., "ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks," *Bioinformatics*, vol. 25, no. 8, pp. 1091–1093, 2009.

[25] H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai, "Lethality and centrality in protein networks," *Nature*, vol. 411, no. 6833, pp. 41–42, 2001.

[26] J. D. J. Han, N. Bertin, T. Hao et al., "Evidence for dynamically organized modularity in the yeast protein-protein interaction network," *Nature*, vol. 430, no. 6995, pp. 88–93, 2004.

[27] P. F. Jonsson and P. A. Bates, "Global topological features of cancer proteins in the human interactome," *Bioinformatics*, vol. 22, no. 18, pp. 2291–2297, 2006.

[28] J. Xu and Y. Li, "Discovering disease-genes by topological features in human protein-protein interaction network," *Bioinformatics*, vol. 22, no. 22, pp. 2800–2805, 2006.

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

[30] J. Wang, S. Zhang, Y. Wang, L. Chen, and X.-S. Zhang, "Disease-aging network reveals significant roles of aging genes in connecting genetic diseases," *PLoS Computational Biology*, vol. 5, no. 9, Article ID e1000521, 2009.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 856325, 10 pages
http://dx.doi.org/10.1155/2013/856325

*Research Article*

# DeGNServer: Deciphering Genome-Scale Gene Networks through High Performance Reverse Engineering Analysis

**Jun Li,[1] Hairong Wei,[2,3] and Patrick Xuechun Zhao[1]**

[1] *Bioinformatics Lab, Plant Biology Division, Samuel Roberts Noble Foundation, 2510 Sam Noble Parkway, Ardmore, OK 73401, USA*
[2] *School of Forest Resources and Environmental Science, Michigan Technological University, 1400 Townsend Drive, Houghton, MI 49931, USA*
[3] *Department of Computer Science, Michigan Technological University, 1400 Townsend Drive, Houghton, MI 49931, USA*

Correspondence should be addressed to Patrick Xuechun Zhao; pzhao@noble.org

Received 22 August 2013; Accepted 1 October 2013

Academic Editor: Zhongming Zhao



Analysis of genome-scale gene networks (GNs) using large-scale gene expression data provides unprecedented opportunities to uncover gene interactions and regulatory networks involved in various biological processes and developmental programs, leading to accelerated discovery of novel knowledge of various biological processes, pathways and systems. The widely used context likelihood of relatedness (CLR) method based on the mutual information (MI) for scoring the similarity of gene pairs is one of the accurate methods currently available for inferring GNs. However, the MI-based reverse engineering method can achieve satisfactory performance only when sample size exceeds one hundred. This in turn limits their applications for GN construction from expression data set with small sample size. We developed a high performance web server, DeGNServer, to reverse engineering and decipher genome-scale networks. It extended the CLR method by integration of different correlation methods that are suitable for analyzing data sets ranging from moderate to large scale such as expression profiles with tens to hundreds of microarray hybridizations, and implemented all analysis algorithms using parallel computing techniques to infer gene-gene association at extraordinary speed. In addition, we integrated the SNBuilder and GeNa algorithms for subnetwork extraction and functional module discovery. DeGNServer is publicly and freely available online.

## 1. Introduction

The advent of high-throughput technologies including microarray experiments and RNA-Seq technologies has generated terabytes of gene expression data for systematically identifying transcriptional regulation and interactions through the reconstruction of gene networks on genome-wide scale. Analysis of whole genome-scale networks can provide a holistic view of all transcription regulations among and within different subnetworks and allows us to gain a more comprehensive understanding of regulation of cellular processes and events. In the past few years, large amount of gene expression data sets from numerous labs has been published and deposited in public databases such as ArrayExpress [1] and Gene expression Omnibus [2], and the volume of this kind of data is still exploding at an accelerated rate. Previous effort in analyzing these public available data has led to the discovery of large amount of novel biological knowledge, making it become increasingly clear that reverse engineering of such "big data" for genome-scale network reconstruction and analysis is one of the most efficient approaches for understanding how life functions through learning holistic transcription regulation and gene interaction.

To date, reverse engineering of aggregated high volume gene expression data for building accurate gene network is still very challenging. The challenge lies in the high dimensionality of gene space and large sample numbers that demand fast and high efficient algorithms, and enhanced computational power as well. A set of the algorithms operates under such a hypothesis that coexpressed [3–5], roughly coordinated genes [6, 7] and genes with dependency [8–10] across a set of samples indicate a functional relationship [11, 12]. As one of the best gene network construction methods, the context likelihood of relatedness (CLR) method [9] utilizing the mutual information (MI) for scoring the

similarity of gene pairs has been widely used to decipher gene networks for multiple species, such as yeast, bacteria, mammalian, and plants [9, 13–16]. However, it is computationally infeasible to decipher genome-scale networks for species with large genomes on a single computer due to physical limits on CPU speeds and memory capacities. For example, there are more than thirty-five thousand genes (transcripts) in human genome. To decipher a genome-scale network through such reverse engineering method, it will need to calculate more than 1.2 billion MI values if we evaluate genes in pairwise fashion, and it is more likely that we will need to evaluate genes in triples or quadrants. Even for those species with small genomes, it is still a big computational challenge to use this method. When CLR was used to construct global networks for *Escherichia coli* in [9], the authors had to trim the number of genes down to a few thousands in order to reduce the computational complexity to a manageable scale. Obviously, this kind of gene reduction prior network construction could miss many potential gene regulations and interactions in the constructed networks. This is because many important transcription factors or genes involved in signaling transduction are expressed at low level and do not necessarily have high variability in expression [17–19]. These genes can be easily eliminated during data trimming process.

Meanwhile, the estimation of mutual information adopted in CLR method heavily relies on the number of microarray data sets. The mutual information value could be estimated accurately only when the number of microarray profiles is larger than one hundred [20]. However, as more microarray and RNA-seq data become available in public database, this, in turn, demands fast, accurate, and less computational complexity. Therefore it is urgently called to develop a high performance reverse engineering system for large-scale gene network analysis through both innovations in efficient algorithm development and parallel computing implementation.

In this study, we integrated parallel computing technologies into DeGNServer to accelerate network reconstruction and subnetwork extraction, which enables DeGNServer to analyze the "big data" in at least one hundred times faster than the original mutual information based CLR, making it much feasible for reverse-engineering global gene networks using the data from a large genome and discovering novel biological knowledge. Meanwhile, we integrated multiple gene association methods into our DeGNServer for network construction. The benchmark data set demonstrated that most of these different association-based CLR methods could reach very similar accuracy as the original mutual information-based CLR method. In addition, we also integrated the SNBuilder [21] and GeNa [22] communities-finding algorithms for identifying subnetworks by providing some seed genes. The major purpose of our system is to provide a practical system to construct the gene association networks from large scale gene expression data.

## 2. Implementation

*2.1. Overview of Gene Network Analysis Methods and Data Analysis Workflow.* We extended the CLR method through integrating several gene-gene association estimation methods, of which includes Pearson, Spearman correlation [6], Kendall, Theil-sen [23], and Weighted Rank methods [24] as well as the mutual information-based method proposed in the original CLR method [9], in the DeGNServer. In addition, the recently published method, maximal information coefficient (MIC) method [25], which has demonstrated capability in discovering novel associations in large data sets, was also integrated into our DeGNServer. To help the biologists to interpret the inferred network, we integrated SNBuilder [21] and GeNa [22] approaches for subnetwork analysis/functional module discovery. All algorithms have been implemented and deployed on our in-house parallel computing platform, namely, BioGrid, which has dedicated over 700 CPU Cores. Figure 1 illustrates the data analysis workflow in DeGNServer. Utilizing our high performance DeGNServer, typical genome-scale gene networks involving 40,000~50,000 gene models could be constructed from expression data that consists of ~200 microarray hybridizations in less than 30 minutes.

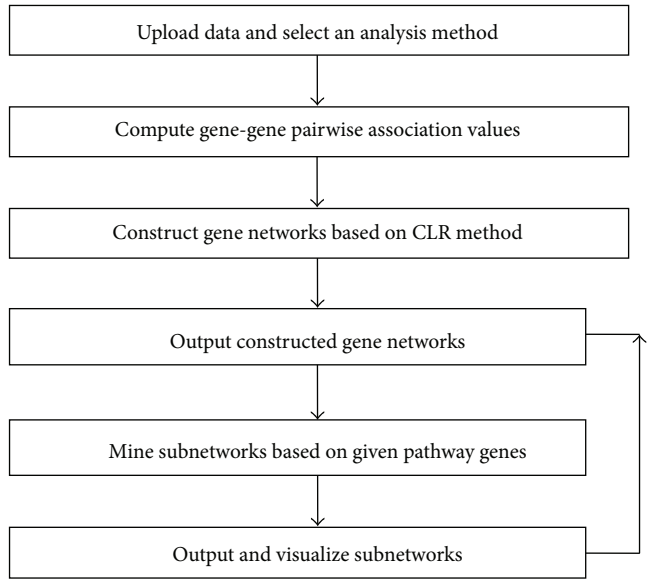



Figure 1: The DeGNServer data analysis workflow.

*2.2. Parallel Computing for the Accelerating of GN Construction.* To accelerate the GN construction through the parallel computing, we split the whole data sets of these gene pairs into multiple subsets. Let $M$ denote the $n \times m$ gene expression matrix, where $n$ denotes the number of genes and $m$ represents the number of gene expression profiles. The computational complexity of association value for all gene-gene pairs of is $O(n^2 \times m)$. The reconstruction of network will be very time-consuming when there exists massive number of expression profiles (e.g., $n > 20,000$ and $m > 1,000$). To tackle this issue, we implemented the GN analysis algorithms using parallel computing techniques. When this task is distributed to all the computing nodes in our Biogrid system, the total computational time complexity is then reduced to $O(n^2 \times m/p)$, where $p$ is the number of allocated processors.

When a gene regulatory network is inferred from $n$ genes, the algorithm will need to compute $n \times (n-1)/2$ pairwise associated values. A two-dimensional $n \times n$ matrix D is used to denote these gene pairs. For gene pair $(i, j)$, the association value of this gene pair will be calculated when the following requirements are satisfied.

(1) When $n$ is even

$$\text{if } i \le \left\lceil \frac{n}{2} \right\rceil, \text{ then } j \in \left[ i+1, \min\left( n-1, i-1+\left\lceil \frac{n}{2} \right\rceil \right) \right],$$
$$\text{if } i > \left\lceil \frac{n}{2} \right\rceil, \text{ then } j \in [i+1, n-1] \cup \left[ 0, i+1-\left\lceil \frac{n}{2} \right\rceil \right]. \quad (1)$$

(2) When $n$ is odd

$$\text{if } i \le \left\lceil \frac{n}{2} \right\rceil, \text{ then } j \in \left[ i+1, \min\left( n-1, i+\left\lceil \frac{n}{2} \right\rceil \right) \right],$$
$$\text{if } i > \left\lceil \frac{n}{2} \right\rceil, \text{ then } j \in [i+1, n-1] \cup \left[ 0, i-1-\left\lceil \frac{n}{2} \right\rceil \right]. \quad (2)$$

For every processor in our Biogrid system, we assign $n/p$ rows of matrix to this processor for the calculation of their corresponding association values.

*2.3. Input Detail.* DeGNServer accepts normalized expression data either in a tab-delimited text file or tab-delimited text. The server DeGNServer provides two options to construct different networks, that is, the coexpression networks and the CLR method-based association networks. Users may adjust the parameter settings, including gene-gene association estimation method and cut-off threshold, to control the size of constructed networks. After the networks are reconstructed, user may submit a list of genes-of-interest and select different subnetwork identification methods to further mine and visualize the same subnetwork generated from different extraction methods.

*2.4. Output Detail.* DeGNServer lists links to the constructed networks/subnetworks in Cytoscape [26] compatible text files, which can be easily imported into the popular Cytoscape software for downstream analysis. In addition, the DeGNServer output page provides interfaces for query and network visualization through Cytoscape web plug-in [24] for each identified subnetwork.

*2.5. Technical Detail.* The DeGNServer is currently deployed on Linux using resin Java server 4.0. It has been tested using the popular web browsers, such as Internet Explorer, Firefox, and Google Chrome. The web interfaces are implemented in JAVA and JSP scripts. All backend integrated analysis algorithms are implemented with parallel programming techniques in efficient C++ computing language and are deployed on an in-house developed Linux cluster, namely, BioGrid, which currently consists of about 700 CPU Cores, to achieve high performance computing capacity. Upon job submission through DeGNServer web server, the master node of BioGrid system firstly divides the gene expression matrix into multiple submatrixes and transfers these submatrixes to slave computing nodes in the Linux Cluster. Next, the master node remotely calls to execute the analysis pipelines and monitors analysis progresses in these computing nodes. Finally the master node collects the association values of all gene-gene pairs for gene network construction and subnetwork analysis. For those species with large genomes, the distributions of gene-gene pairs are close to the normal distribution, so we applied the normal distribution to calculate the $z$-score of gene-gene pairs. Based on the preset $z$-score threshold, those gene-gene pairs whose $z$-scores are less than the threshold would be discarded. Figure 2 illustrates the parallel implementation of the CLR Method.

## 3. Results

*3.1. Performance Evaluation with Synthetic Data.* To comprehensively evaluate performance of integrated network construction methods, we generated two groups of synthetic compendium gene expression data sets, each group with a series of data sets of various sizes, using the SynTReN software [27] and the regulatory network models based upon *Escherichia coli* experimental data as original seeds. The sampled sizes of Group A data sets are 30, 40, 50, 60, 70, 80, and 90, while the sizes for Group B are 100, 200, 300, 400, 500, 600, 700, 800, 900, and 1000 samples. We analyzed each of these compendium data sets with various sample sizes and then generated respective subnetworks containing 50 genes. The prediction accuracy against the corresponding reference network in SynTReN software uses the area under the receiver operating characteristic curve (ROC) curve, namely, the AUC scores [28], to represent the accuracy of each method. The AUC scores resulting from all compendium data sets within each group were averaged, and results of averaged AUC scores for all method in each group are shown in Figure 3.

The ROC curve indicates the change of sensitivity (true positive rate) versus specificity (true negative rate) under different thresholds, and AUC score can represent the accuracy of each method better because it is independent of different thresholds.

The following formula is used to calculate the sensitivity and the specificity:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$
$$\text{Specificity} = \frac{\text{FP}}{\text{FP} + \text{TN}}. \quad (3)$$

All methods were applied to construct GNs with each sampled data set in either Group A or B with the positive regulatory relationships being counted. We then calculated their respective AUC scores. For each group (smaller and large number of expression data sets), we compared their average AUC scores for different methods. Figure 3 shows that the prediction accuracies of Spearman-based CLR method have higher average AUC scores than other methods, suggesting that Spearman-based CLR method may produce better results in term of network construction.

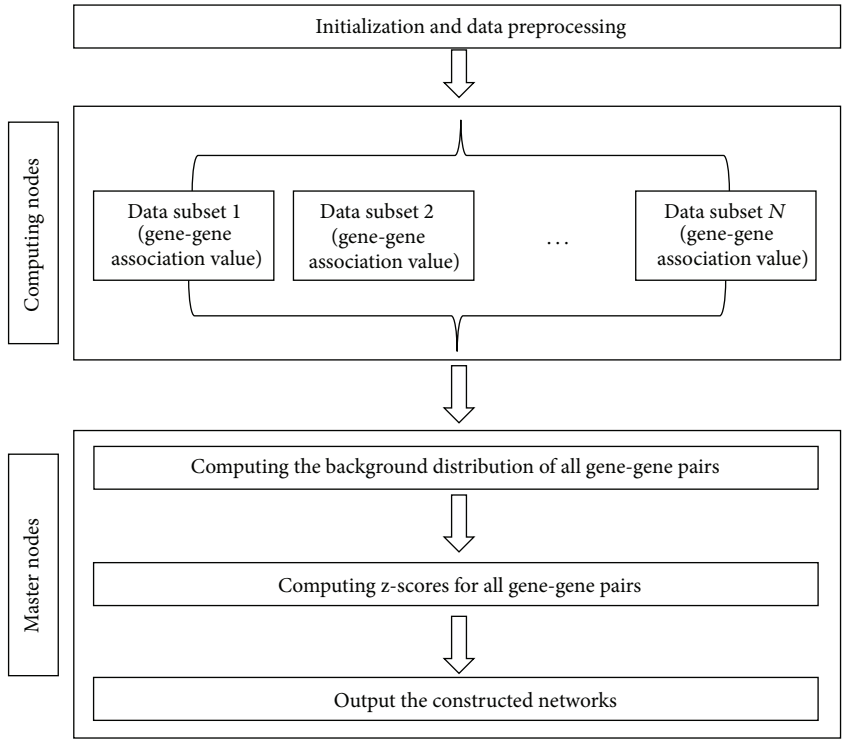



Figure 2: Parallel implementation of the CLR method.

### 3.2. Case Study 1: Deciphering Genome-Scale Pluripotency Networks in Human Embryonic Stem Cells

*3.2.1. Human Stem Cell Microarray Data Set.* To validate the performance of DeGNServer, we analyzed genome-scale networks from 189 human stem cell microarray profiles. These data sets were generated in 17 individual experiments in which human embryonic stem cells were treated with various reagents for inducing differentiation. Therefore, this compendium data set is enriched with regulatory events and interaction of pluripotency maintenance and transition from pluripotent stem cells to differentiated cell linages, and thus it can serve as an ideal testing data for the performance of DeGNServer in discovering functionally associated gene subnetworks governing these processes. Of these 189 microarray data sets, there are 104 high-density human gene expression arrays from HG17 assembly. This platform of microarray contains 388,634 probes from 36,494 human locus identifiers. These 104 chips were compiled from 15 experiments in which stem cells were treated with different reagents that disrupted pluripotency. The reagents and the conditions included 12-O-tetradecanoylphorbol-13-acetate (TPA) treatment in conditioned medium, TPA treatment in TeSR medium, BMP4 treatment with FGF, BMP4 treatment without FGF, and coculture with mouse OP9 cells. The remained 85 high-density human gene expression arrays have 381,002 probes from 47,633 human locus identifiers from the HG18. These 85 microarray data sets were compiled from a set of experiments where a variety of different growth factors were applied to human embryonic stem cells at varying conditions for 3 days. Both HG17 and HG18 microarray platforms were manufactured by NimbleGen Systems (http://www.nimblegen.com/). All probes are 60 mers and all chips were hybridized to Cy5 labeled mRNAs extracted from human embryonic stem cells (hESCs) from undifferentiated to differentiated stages. Raw data were extracted using NimbleScan software v2.1. The two data sets were joined by gene mapping via selection of shared common probes between the same genes on the two platforms. More than 99.5% of mapped genes share at least 6 common probes, and the signal intensities from these common probes were normalized with the Robust Multiple-chip Analysis (RMA) algorithm [29]. Thus, the whole data set obtained contains 36,398 genes.

*3.2.2. Results on Pluripotency Network Analysis in Human Embryonic Stem Cells.* The gene networks including 21,167 genes and 200,000 links were reconstructed in less than 20 minutes with a $z$-score threshold of 4.3 and spearman-based association method. The built network could be retrieved at http://plantgrn.noble.org/DeGNServer/Result.jsp?time4=&sessionid human&method=1_1&cutoff=4.3. We also tested with original mutual information-based CLR, and it took 53.3 hours to complete whole genome-scale network construction.

Generally, global networks with huge numbers of regulations and interactions are a "hairball", from which we can hardly identify any patterns. To facilitate the identification of subnetworks or modules that regulate a specific biological process or developmental program, we integrated both SNBuilder [21] and GeNa [22] methods to extract smaller subnetworks/functional modules by providing a few seed

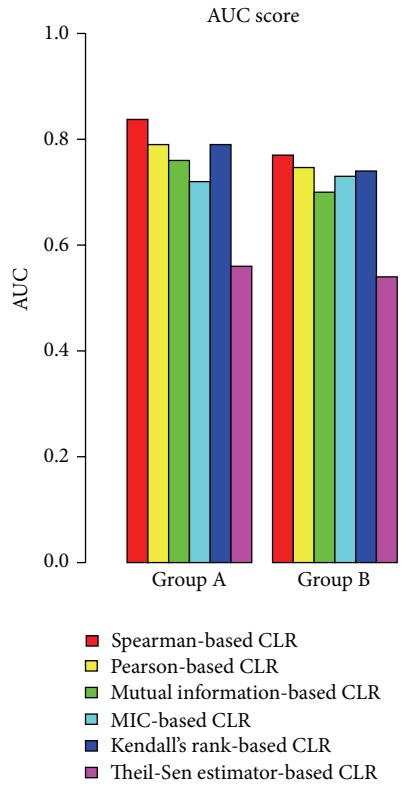



Figure 3: Average AUC scores from different association-based CLR methods for networks with larger and smaller numbers of expression profiles; Group A: networks constructed with smaller number of gene expression samples (30~90 samples), Group B: networks constructed with larger number of expression samples (100~1000). AUC scores were obtained through varying different threhold settings. A perfect model will have AUC score of 1, while random guessing will score an AUC around 0.5.

genes. We used NANOG, POU5F1, SOX2, and PHC1 as seed genes to bait the subnetwork shown in Figure 4.

Figure 4 shows the subnetwork that is implicated to control the pluripotency renewal of human embryonic stem cells. The literature evidence supporting the involvement of those transcription factors on inner ring in regulating pluripotency in human stem cells is already shown in our earlier publication and these TFs could be identified by our TF-Cluster that is capable of constructing gene association network with all TFs as an input [6]. However, it cannot be used to build the genome-wide GN mainly due to computational complexity. In this study, our DeGNServer identified 14 of 16 TF genes that were identified previously by TF-Cluster tool from the same data for governing pluripotency renewal. These 14 TFs include three master transcription factors, NANOG, POU5F1 (or OCT4), and SOX2, which are necessary for pluripotency maintenance, and they alone can convert skin cells to induced pluripotent cells [30]. Although two TFs were missed by our method, we identified six more other genes that are to be involved in pluripotency maintenance in human stem cells. In this study, we only examined the existing literature of six genes that are located on the outer rings (Figure 4). The developmental pluripotency-associated 2 (DPPA2) gene plays important roles in the maintenance of pluripotency and proliferation of human embryonic stem cells by regulating chromatin structures [31]. Although there is no direct evidence from human stem cells, study on mouse stem cells shows that DPPA2 knockdown induces the differentiation, while it represses proliferation of mouse embryonic stem cells [31]. PRDM14 is an important determinant of the human embryonic stem cell (ESC) identity, and it works in concert with the core ESC regulators to activate pluripotency-associated genes [32]. PRDM14 binds to silenced genes and serves as a direct repressor of differentiation genes in human stem cells though the exact mechanism of this repressive activity remains unknown. ZMYND8 encodes a zinc finger protein with a complex role in maintaining pluripotency. Although only expressed at low levels, either up- or down-regulation of ZMYND8 can induce differentiation in ES cells [33]. JARID2 is a component of chromatin modification complex PRC2 in embryonic stem cells and is required for multilineage differentiation. It plays a role in recruiting PRC1 and RNA Polymerase II to developmental regulators. We found that JARID2 and CD99 in our subnetwork and previous study have shown JARID2 functions together with CD99 in controlling autism spectrum disorder [34]. The exact function of DEPDC2 is currently unknown, but it is known that the promoter of DEPDC2 is bound by the three master transcription factors, NANOG, SOX 2, and POU5F1 as mentioned above [35]. DEPDC2 is a molecular marker for human stem cell [36] though its exact function remains unknown. Similarly, the exact function of CHST4 is currently unknown, but it is know that CHST4 is one of the 16 methylation markers of embryonic stem cells, and these 16 methylation markers also include PRDM14 as mentioned above [37].

To further examine the sensitivity, specificity, and prediction accuracy of the case study described above, we made some assumptions. (1) We assumed that the genes that are evidenced to be involved in pluripotency maintenance in the existing literature are all positive genes; we then counted the true positive (TP) and false positive (FP) genes within each subnetwork. The true negative (TN) and false negative (FN) genes were calculated from the rest of network that was adjusted to the same size of each subnetwork. For comparison, we rescaled all numbers to one hundred before we calculated sensitivity, specificity, and prediction accuracy. The results were shown in Table 1. The results demonstrate the high accuracy of the DeGNServer.

### 3.3. Case Study 2: Deciphering Genome-Scale Pluripotency Networks in Murine Heart Tissues

*3.3.1. Mouse Heart Microarray Data Set.* We also analyzed a compendium microarray data set from heart tissues of *Mus musculus* to evaluate the efficiency of the DeGNServer. This compendium data set includes 172 Affymetrix microarray chips of platform GPL1261, which contains 45,101 probes. The data was downloaded from NCBI Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/). These

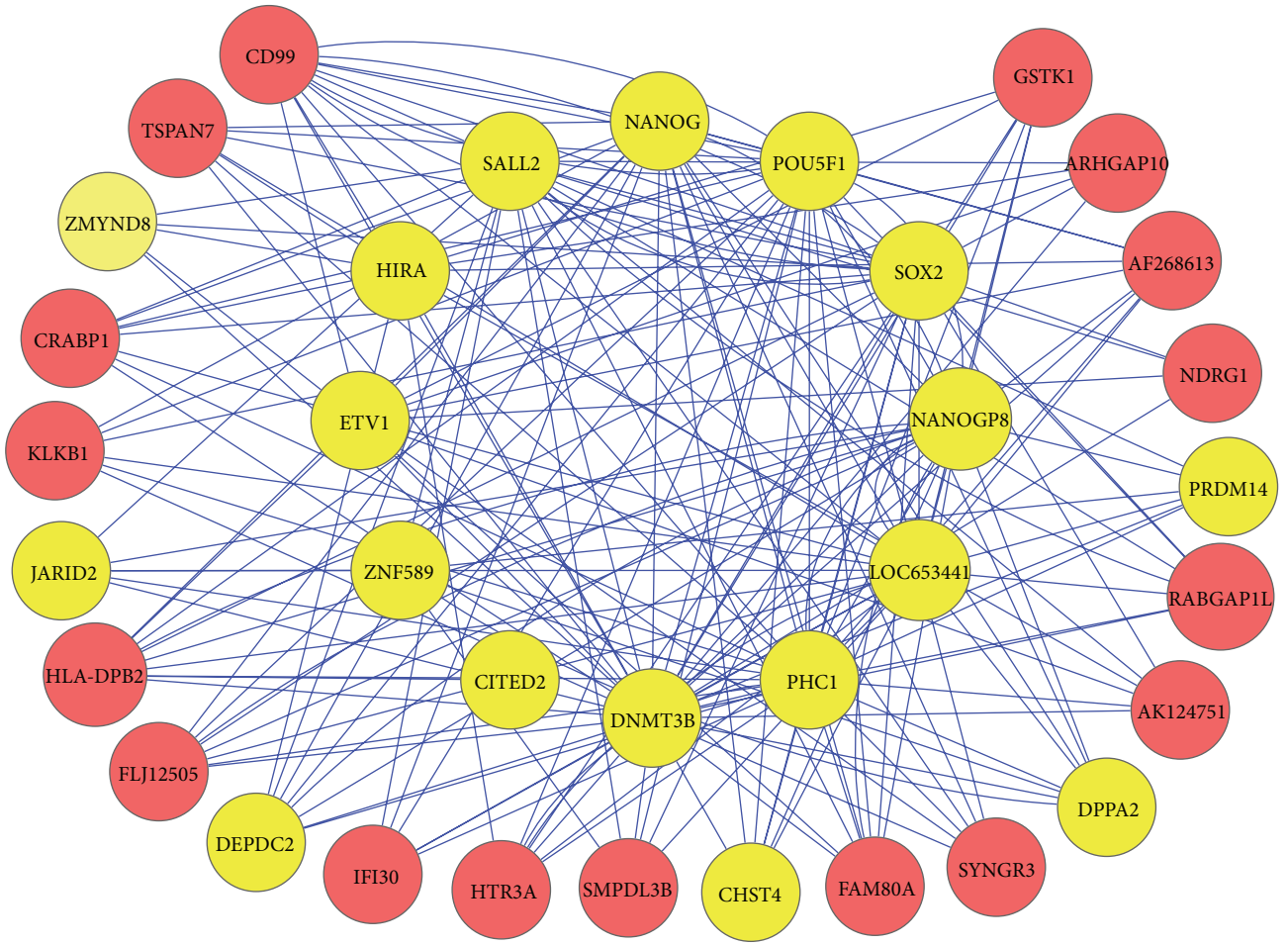



Figure 4: The identified subnetwork contains the essential transcription factors and other genes required for pluripotency maintenance. The twelve genes on the inner ring are transcription factors known to play essential or important role in pluripotency renewal of human embryonic stem cells. These include three master transcription factors, NANOG, POU5F1, and SOX2, which are absolutely required for pluripotency maintenance. The genes located on the outer ring were identified by DeGNServer for being closely coordinated with those transcription factors in the inner ring. The genes on outer ring, but highlighted in yellow, are those that are implicated by the existing literature to participate in the pluripotency renewal. This subnetwork was generated by using SNBuilder method [21] with NANOG, POU5F1, SOX2, and PHC1 as query seeds.

Table 1: Sensitivity, specificity, and prediction accuracy of two case studies.

| Case studies | TP | FP | TN | FN | Sensitivity | Specificity | Prediction accuracy |
|---|---|---|---|---|---|---|---|
| Human stem cell | 2.42 | 97.58 | 99.95 | 0.05 | 98% | 50.6% | 51.2% |
| Mouse heart | 39.6 | 60.4 | 97.5 | 2.50 | 94.1% | 61.7 % | 68.6% |

Prediction accuracy = ((TP + TN)/(TP + FP + TN + FN)) × 100%.

172 microarray data were from nine independent experiments that have the following GEO accession IDs: GSE11291, 15078, 19875, 29145, 30495, 3440, 38754, 5500, and 7781. The compendium data were generated through pooling the raw data of 172 microarray data and then normalized with RMA algorithm [29]. For quality control, we used two methods that were previously described [38].

*3.3.2. Overall Performance.* The gene networks including 41,742 genes and 3,869,157 links were deciphered in less than 30 minutes with a $z$-score threshold of 3.8 and spearman-based association method. The built network could be retrieved at http://plantgrn.noble.org/DeGNServer/Result.jsp?sessionid=1367625665687&method=1_1&cutoff=3.8#. We also tested with original mutual information-based CLR, which took 81.6 hours to complete whole genome-scale network construction.

*3.3.3. Subnetworks Controlling Murine Heart Development.* The pathway that controls murine heart development can be obtained from NCBI's BioSystems database with an accession number of 672437 [39]. From the pathway diagram, we can find the three central genes, Nkx2-5, Tbx1, and Mef2c, which play very important roles in heart development, as showed

up in the subnetwork, we obtained (Figure 5). Nkx2-5 is known to be involved in cardiac muscle cell differentiation [40], proliferation [41], contraction [42], and muscle tissue development [43]. Lack of Nkx2-5 can lead to the myogenic and morphogenetic defects in the heart tubes [43]. Mef2c and Nkx2.5 are known to control common downstream targets and exhibit striking phenotypic similarities when disrupted [43]. Tbx1 affects asymmetric cardiac morphogenesis by regulating Pitx2 in the secondary heart field [44]; it also controls regional coronary artery morphogenesis [45], aorta morphogenesis [46], and blood vessel development [47]. Prox1 is known to function as a direct upstream modifier of Nkx2.5 and is responsible for maintaining muscle structure and growth [48, 49]. CAMTAs promote cardiomyocyte hypertrophy and activate the ANF gene, at least in part, by associating with the cardiac homeodomain protein Nkx2-5 [50]. The transcriptional activity of CAMTAs is governed by association with class II histone deacetylases (HDACs), which negatively regulate cardiac growth [50]. Smarca4, as a nuclear notch signaling component required for the establishment of left-right asymmetry [51], is also essential for heart development by involving chromatin remodeling complexes [51]. Kdm6 interacts with Smarca4 to control T-box family member-dependent gene expression [52]. Wnt2 is required for atrial and inflow tract morphogenesis, and it regulates expansion of secondary heart field progenitors [53]. Myocd controls cardiac muscle cell proliferation, growth, and differentiation [54]. Eno3 is highly expressed in skeletal muscle and heart [55]. The specific function of murine Chst2 is currently unknown, but human umbilical vein endothelial cells predominantly express CHST2 [56, 57]. The heart requires glycerol as an energy substrate through aquaporin 7, a glycerol facilitator [58]. Glycerol is taken into cardiomyocytes and is finally converted to pyruvate by Gpd2 enzymes [59]. *EphA4* mutant mice exhibit defects in the coronal suture and neural crest-mesoderm boundary [60].

*3.4. Sensitivity, Specificity, and Prediction Accuracy of the above Two Case Studies.* To further examine the sensitivity, specificity, and prediction accuracy of the three case studies as shown above, we made some assumptions: (1) for human pluripotency renewal, we assumed that the genes that are evidenced to be involved in pluripotency maintenance in the existing literature are all positive genes; (2) for heart development, due to the large number of genes involved in these biological processes, we cannot search the literature evidence for all genes. We classified all genes involved in heart development to be positive based on gene ontologies. We then counted the true positive (TP) and false positive (FP) genes within each subnetwork. The true negative (TN) and false negative (FN) genes were calculated from the rest of network that was adjusted to the same size of each subnetwork. For comparison, we rescaled all numbers to one hundred before we calculated sensitivity, specificity, and prediction accuracy. The results were shown in Table 1.

## 4. Discussions

We developed the DeGNServer to enable the reconstruction of genome-scale GN using the increasingly accumulated large-scale gene expression data in public domain. Users may use it to generate whole genome scale GNs from large amount of gene expression data in any species. After whole genome GN construction, users can obtain the subnetworks by providing a few genes of interest. All subnetworks generated with different genes of interest and thresholds will be automatically listed online for downloading and studying. When genome-wide network construction was performed with 189 human microarray profiles as an input for DeGNServer, we could identify a subnetwork containing majority of genes involved in pluripotency maintenance in human embryonic stem cells [6, 30, 35]. It is worth mentioning that TF-Cluster pipeline that we developed earlier [6] is capable of building a coordinated network using the same human compendium data set and identifies only those transcription factors located on the inner ring in Figure 4, but it misses all genes that are located on the outer ring in Figure 4 mainly because it can build a local transcription factor coordination network rather than the whole genome-scale network. When genome-wide GNs were constructed using the DeGNServer, we could identify more genes (shown in outer ring in Figure 4) that regulate human pluripotency renewal together with those major transcription factors as shown in inner ring in Figure 4. To test if DeGNServer can identify true subnetworks in different circumstances, we also applied it to a murine compendium data set we downloaded and pooled from GEO database. The data is from heart tissues of *Mus musculus*. We obtained a subnetwork that contains functionally cohesive genes known to control the heart developmental program in mouse. This evidence clearly indicated that the use of DeGNServer can lead to the deciphering of the more comprehensive networks from which we can discover new genes involved in a specific biological process. We thus think that DeGNServer is useful in identifying genes governing a specific biological process, pathway, or a developmental program.

Although we have tested with synthetic data and found that Spearman-based CLR appears to have better performance than any of other methods including original mutual information based CLR, we still make all methods available in DeGNServer. This is because the efficiency of different methods may be dependent on the properties of biological data, as we showed in a previous study [7]. For subnetwork extraction, we integrated both SNBuilder and GeNa algorithms; both are found to be proficient in identifying the true subnetworks. However, GeNa usually produces small subnetworks with cohesive function.

## 5. Conclusions

We have developed a high performance web-based platform, namely, DeGNServer, for genome-scale GN construction and subnetwork extraction. DeGNServer is capable of analyzing gene expression data with very high dimensionality of gene space and very large number of gene expression profiles. As tested, it can analyze hundreds of microarray profiles of human (36,000 genes) for reconstruction of gene association networks within 30 minutes, mainly through the improvement of gene association estimation algorithms and parallel computing in combination. The DeGNServer

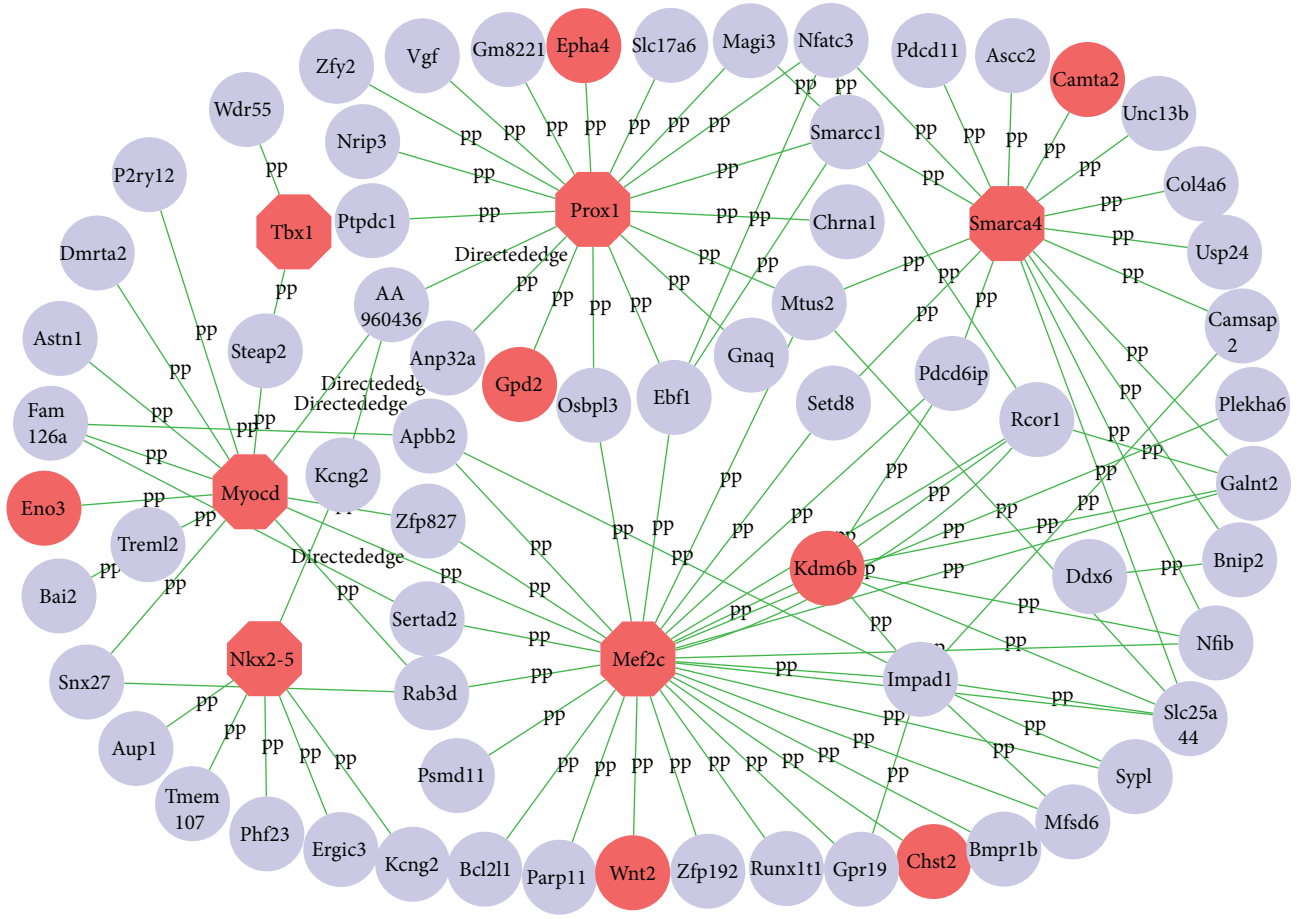



FIGURE 5: The subnetwork that is responsible for heart growth and development in mouse. The whole genome-scale network was constructed from 175 chips of GPL1261 platform using DeGNServer and then extracted using community-finding algorithm called GeNa [22] with Nkx2-5, Prox1, and Mef2c as query seeds. Genes highlighted in red are implicated by the existing literature to participate in heart growth and development.

is as accurate and sensitive as the original CLR method and runs hundreds to thousands times faster. Furthermore, through the integration of network decomposition methods, the DeGNServer is capable of identifying novel functional cohesive subnetworks or modules.

## Acknowledgments

The authors thank Dr. Xinbin Dai for his assistance in DeGNServer deployment. This work was supported by the National Science Foundation (Grant DBI : 0960897 to Patrick Xuechun Zhao) and the Samuel Roberts Noble Foundation.

## References

[1] H. Parkinson, U. Sarkans, M. Shojatalab et al., "ArrayExpress—a public repository for microarray gene expression data at the EBI," *Nucleic Acids Research*, vol. 33, pp. D553–D555, 2005.

[2] 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, no. 1, pp. D760–D765, 2007.

[3] H. Wei, S. Persson, T. Mehta et al., "Transcriptional coordination of the metabolic network in arabidopsis," *Plant Physiology*, vol. 142, no. 2, pp. 762–774, 2006.

[4] X. L. Zhu, Z. H. Ai, J. Wang, Y. L. Xu, and Y. C. Teng, "Weighted gene co-expression network analysis in identification of endometrial cancer prognosis markers," *Asian Pacific Journal Cancer Prevention*, vol. 13, no. 9, pp. 4607–4611, 2012.

[5] A. P. Presson, E. M. Sobel, J. C. Papp et al., "Integrated weighted gene co-expression network analysis with an application to chronic fatigue syndrome," *BMC Systems Biology*, vol. 2, article 95, 2008.

[6] J. Nie, R. Stewart, F. Ruan et al., "TF-cluster: a pipeline for identifying functionally coordinated transcription factors via network decomposition of the Shared Coexpression Connectivity Matrix (SCCM)," *BMC Systems Biology*, vol. 5, article 53, 2011.

[7] S. Kumari, R. Stewart, J. Nie et al., "Evaluation of gene association methods for coexpression network construction and biological knowledge discovery," *PLoS ONE*, vol. 7, no. 11, Article ID e50411, 2012.

[8] A. A. Margolin, I. Nemenman, K. Basso et al., "ARACNE: an algorithm for the reconstruction of gene regulatory networks

in a mammalian cellular context," *BMC Bioinformatics*, vol. 7, supplement 1, article S7, 2006.

[9] J. J. Faith, B. Hayete, J. T. Thaden et al., "Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles," *PLoS Biology*, vol. 5, no. 1, article e8, 2007.

[10] J. Li, H. Wei, T. Liu, and P. X. Zhao, "GPLEXUS: enabling genome-scale gene association network reconstruction and analysis for very large-scale expression data," *Nucleic Acids Research*, 2013.

[11] K.-C. Li, "Genome-wide coexpression dynamics: theory and application," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 26, pp. 16875–16880, 2002.

[12] S. Horvath and J. Dong, "Geometric interpretation of gene coexpression network analysis," *PLoS Computational Biology*, vol. 4, no. 8, Article ID e1000117, 2008.

[13] A. Madar, A. Greenfield, E. Vanden-Eijnden, and R. Bonneau, "DREAM3: network inference using dynamic context likelihood of relatedness and the inferelator," *PloS ONE*, vol. 5, no. 3, Article ID e9803, 2010.

[14] A. Nazri and P. Lio, "Investigating meta-approaches for reconstructing gene networks in a mammalian cellular context," *PLoS ONE*, vol. 7, no. 1, Article ID e28713, 2012.

[15] J. Xiong, D. Yuan, J. S. Fillingham et al., "Gene network landscape of the ciliate tetrahymena thermophila," *PLoS ONE*, vol. 6, no. 5, Article ID e20124, 2011.

[16] T. Michoel, R. de Smet, A. Joshi, Y. van de Peer, and K. Marchal, "Comparative analysis of module-based versus direct methods for reverse-engineering transcriptional regulatory networks," *BMC Systems Biology*, vol. 3, article 49, 2009.

[17] D. Wang, C. Zhang, D. J. Hearn et al., "Identification of transcription-factor genes expressed in the arabidopsis female gametophyte," *BMC Plant Biology*, vol. 10, article 110, 2010.

[18] R. Osorno and I. Chambers, "Transcription factor heterogeneity and epiblast pluripotency," *Philosophical Transactions of the Royal Society B*, vol. 366, article 1575, pp. 2230–2237, 2011.

[19] D. Linnekin, L. S. Park, and W. L. Farrar, "Dissociation of human cytokine receptor expression and signal transduction," *Blood*, vol. 80, no. 8, pp. 1896–1904, 1992.

[20] S. Edwards, "Elements of information theory, 2nd edition," *Information Processing & Management*, vol. 44, no. 1, pp. 400–401, 2008.

[21] X. Hu and F.-X. Wu, "Mining and state-space modeling and verification of sub-networks from large-scale biomolecular networks," *BMC Bioinformatics*, vol. 8, article 324, 2007.

[22] M. Aluru, J. Zola, D. Nettleton, S. Aluru et al., "Reverse engineering and analysis of large genome-scale gene networks," *Nucleic Acids Research*, vol. 41, no. 1, article e24, 2013.

[23] R. R. Wilcox, "A note on the Theil-Sen regression estimator when the regresser is random and the error term is heteroscedastic," *Biometrical Journal*, vol. 40, no. 3, pp. 261–268, 1998.

[24] I. G. Lyakhov, A. Krishnamachari, and T. D. Schneider, "Discovery of novel tumor suppressor p53 response elements using information theory," *Nucleic Acids Research*, vol. 36, no. 11, pp. 3828–3833, 2008.

[25] D. N. Reshef, Y. A. Reshef, H. K. Finucane et al., "Detecting novel associations in large data sets," *Science*, vol. 334, no. 6062, pp. 1518–1524, 2011.

[26] M. Kohl, S. Wiese, and B. Warscheid, "Cytoscape: software for visualization and analysis of biological networks," *Methods in Molecular Biology*, vol. 696, pp. 291–303, 2011.

[27] T. van den Bulcke, K. van Leemput, B. Naudts et al., "SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms," *BMC Bioinformatics*, vol. 7, article 43, 2006.

[28] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: an overview," *Bioinformatics*, vol. 16, no. 5, pp. 412–424, 2000.

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

[30] J. Yu, M. A. Vodyanik, K. Smuga-Otto et al., "Induced pluripotent stem cell lines derived from human somatic cells," *Science*, vol. 318, no. 5858, pp. 1917–1920, 2007.

[31] T. Watabe, "Roles of Dppa2 in the regulation of the present status and future of pluripotent stem cells," *Journal of Biochemistry*, vol. 152, no. 1, pp. 1–3, 2012.

[32] Y. S. Chan, J. Göke, X. Lu et al., "A PRC2-dependent repressive role of PRDM14 in human embryonic stem cells and induced pluripotent stem cell reprogramming," *Stem Cells*, vol. 31, no. 4, pp. 682–692, 2012.

[33] I. Barbaric and N. J. Harrison, "Rediscovering pluripotency: from teratocarcinomas to embryonic stem cells. Cardiff, 10–12 October 2011," *The International Journal of Developmental Biology*, vol. 56, no. 4, pp. 197–206, 2012.

[34] P. S. Ramos, S. Sajuthi, C. D. Langefeld et al., "Immune function genes CD99L2, JARID2 and TPO show association with autism spectrum disorder," *Molecular Autism*, vol. 3, no. 1, article 4, 2012.

[35] L. A. Boyer, I. L. Tong, M. F. Cole et al., "Core transcriptional regulatory circuitry in human embryonic stem cells," *Cell*, vol. 122, no. 6, pp. 947–956, 2005.

[36] R. Calloni, E. A. Cordero, J. A. Henriques, D. Bonatto et al., "Reviewing and updating the major molecular markers for stem cells," *Stem Cells and Development*, vol. 22, no. 9, pp. 1455–1476, 2013.

[37] M. Kim, T.-W. Kang, H.-C. Lee et al., "Identification of DNA methylation markers for lineage commitment of in vitro hepatogenesis," *Human Molecular Genetics*, vol. 20, no. 14, pp. 2722–2733, 2011.

[38] S. Persson, H. Wei, J. Milne, G. P. Page, and C. R. Somerville, "Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 24, pp. 8633–8638, 2005.

[39] D. Srivastava, "Making or breaking the heart: from lineage determination to morphogenesis," *Cell*, vol. 126, no. 6, pp. 1037–1048, 2006.

[40] J. W. Vincentz, R. M. Barnes, B. A. Firulli, and S. J. Conway, "Cooperative interaction of Nkx2.5 and Mef2c transcription factors during heart development," *Developmental Dynamics*, vol. 237, no. 12, pp. 3809–3819, 2008.

[41] O. W. J. Prall, M. K. Menon, M. J. Solloway et al., "An Nkx2-5/Bmp2/Smad1 negative feedback loop controls heart progenitor specification and proliferation," *Cell*, vol. 128, no. 5, pp. 947–959, 2007.

[42] M. Takeda, L. E. Briggs, H. Wakimoto et al., "Slow progressive conduction and contraction defects in loss of Nkx2-5 mice after

cardiomyocyte terminal differentiation," *Laboratory Investigation*, vol. 89, no. 9, pp. 983–993, 2009.

[43] I. Lyons, L. M. Parsons, L. Hartley et al., "Myogenic and morphogenetic defects in the heart tubes of murine embryos lacking the homeo box gene Nkx2-5," *Genes and Development*, vol. 9, no. 13, pp. 1654–1666, 1995.

[44] S. Nowotschin, J. Liao, P. J. Gage, J. A. Epstein, M. Campione, and B. E. Morrow, "Tbx1 affects asymmetric cardiac morphogenesis by regulating Pitx2 in the secondary heart field," *Development*, vol. 133, no. 8, pp. 1565–1573, 2006.

[45] M. Théveniau-Ruissy, M. Dandonneau, K. Mesbah et al., "The del22q11.2 candidate gene Tbx1 controls regional outflow tract identity and coronary artery patterning," *Circulation Research*, vol. 103, no. 2, pp. 142–148, 2008.

[46] Z. Zhang and A. Baldini, "In vivo response to high-resolution variation of Tbx1 mRNA dosage," *Human Molecular Genetics*, vol. 17, no. 1, pp. 150–157, 2008.

[47] S. Merscher, B. Funke, J. A. Epstein et al., "TBX1 is responsible for cardiovascular defects in velo-cardio-facial/DiGeorge syndrome," *Cell*, vol. 104, no. 4, pp. 619–629, 2001.

[48] C. A. Risebro, R. G. Searles, A. A. D. Melville et al., "Prox1 maintains muscle structure and growth in the developing heart," *Development*, vol. 136, no. 3, pp. 495–505, 2009.

[49] C. A. Risebro, L. K. Petchey, N. Smart et al., "Epistatic rescue of Nkx2. 5 adult cardiac conduction disease phenotypes by prospero-related homeobox protein 1 and HDAC3," *Circulation Research*, vol. 111, no. 2, pp. e19–e31, 2012.

[50] K. Song, J. Backs, J. McAnally et al., "The transcriptional coactivator CAMTA2 stimulates cardiac growth by opposing class II histone deacetylases," *Cell*, vol. 125, no. 3, pp. 453–466, 2006.

[51] H. Lickert, J. K. Takeuchi, I. Von Both et al., "Baf60c is essential for function of BAF chromatin remodelling complexes in heart development," *Nature*, vol. 432, no. 7013, pp. 107–112, 2004.

[52] S. A. Miller, S. E. Mohn, and A. S. Weinmann, "Jmjd3 and UTX play a demethylase-independent role in chromatin remodeling to regulate t-box family member-dependent gene expression," *Molecular Cell*, vol. 40, no. 4, pp. 594–605, 2010.

[53] Y. Tian, L. Yuan, A. M. Goss et al., "Characterization and in vivo pharmacological rescue of a Wnt2-Gata6 pathway required for cardiac inflow tract development," *Developmental Cell*, vol. 18, no. 2, pp. 275–287, 2010.

[54] D.-Z. Wang, P. S. Chang, Z. Wang et al., "Activation of cardiac gene expression by myocardin, a transcriptional cofactor for serum response factor," *Cell*, vol. 105, no. 7, pp. 851–862, 2001.

[55] J. Wu, D. Zhou, C. Deng, X. Wu, L. Long, and Y. Xiong, "Characterization of porcine ENO3: genomic and cDNA structure, polymorphism and expression," *Genetics Selection Evolution*, vol. 40, no. 5, pp. 563–579, 2008.

[56] X. Li, L. Tu, P. G. Murphy, T. Kadono, D. A. Steeber, and T. F. Tedder, "CHST1 and CHST2 sulfotransferase expression by vascular endothelial cells regulates shear-resistant leukocyte rolling via L-selectin," *Journal of Leukocyte Biology*, vol. 69, no. 4, pp. 565–574, 2001.

[57] X. Li and T. F. Tedder, "CHST1 and CHST2 sulfotransferases expressed by human vascular endothelial cells: cDNA cloning, expression, and chromosomal localization," *Genomics*, vol. 55, no. 3, pp. 345–347, 1999.

[58] T. Hibuse, N. Maeda, H. Nakatsuji et al., "The heart requires glycerol as an energy substrate through aquaporin 7, a glycerol facilitator," *Cardiovascular Research*, vol. 83, no. 1, pp. 34–41, 2009.

[59] S. Gambert, C. Héliès-Toussaint, and A. Grynberg, "Regulation of intermediary metabolism in rat cardiac myocyte by extracellular glycerol," *Biochimica et Biophysica Acta*, vol. 1736, no. 2, pp. 152–162, 2005.

[60] M.-C. Ting, N. L. Wu, P. G. Roybal et al., "EphA4 as an effector of Twist1 in the guidance of osteogenic precursor cells during calvarial bone growth and in craniosynostosis," *Development*, vol. 136, no. 5, pp. 855–864, 2009.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 620793, 8 pages
http://dx.doi.org/10.1155/2013/620793

*Research Article*

# Novel Natural Structure Corrector of ApoE4 for Checking Alzheimer's Disease: Benefits from High Throughput Screening and Molecular Dynamics Simulations

**Manisha Goyal,[1] Sonam Grover,[2] Jaspreet Kaur Dhanjal,[3] Sukriti Goyal,[1] Chetna Tyagi,[1] Sajeev Chacko,[4] and Abhinav Grover[2]**

[1] *Apaji Institute of Mathematics & Applied Computer Technology, Banasthali University, Tonk, Rajasthan 304022, India*
[2] *School of Biotechnology, Jawaharlal Nehru University, New Delhi 110067, India*
[3] *Department of Biotechnology, Delhi Technological University, New Delhi 110042, India*
[4] *Thematic Unit of Excellence on Computational Materials Science, S. N. Bose National Centre for Basic Sciences, Sector III, Block JD, Salt Lake, Kolkata 700098, India*

Correspondence should be addressed to Abhinav Grover; abhinavgr@gmail.com

Received 27 August 2013; Accepted 1 October 2013

Academic Editor: Zhongming Zhao



A major genetic suspect for Alzheimer's disease is the pathological conformation assumed by apolipoprotein E4 (ApoE4) through intramolecular interaction. In the present study, a large library of natural compounds was screened against ApoE4 to identify novel therapeutic molecules that can prevent ApoE4 from being converted to its pathological conformation. We report two such natural compounds PHC and IAH that bound to the active site of ApoE4 during the docking process. The binding analysis suggested that they have a strong mechanistic ability to correct the pathological structural orientation of ApoE4 by preventing repulsion between Arg 61 and Arg 112, thus inhibiting the formation of a salt bridge between Arg 61 and Glu 255. However, when the molecular dynamics simulations were carried out, structural changes in the PHC-bound complex forced PHC to move out of the cavity thus destabilizing the complex. However, IAH was structurally stable inside the binding pocket throughout the simulations trajectory. Our simulations results indicate that the initial receptor-ligand interaction observed after docking could be limited due to the receptor rigid docking algorithm and that the conformations and interactions observed after simulation runs are more energetically favored and should be better representations of derivative poses in the receptor.

## 1. Introduction

Alzheimer's disease (AD) is the most common form of dementia. AD is a harmful neurological disorder that affects about 5.4 million Americans of all ages [1]. One in every eight old Americans has AD, making it the sixth major cause of death in the United States [1]. In India, the annual incidence rate per 1,000 persons for AD is 11.67 for those above 55 years of age and even higher for those above 65 years [2]. AD, which affects memory, thinking ability, and behavior, is characterized by complex neuropathological features that include heaping of amyloid $\beta$ (A$\beta$) followed by synaptic dysfunction, formation of neurofibrillary tangles, and elements of degenerating neurons [3]. Degeneration causes a decrease in the acetylcholine levels and in the activities of choline acetyltransferase [4].

Although the U.S. Food and Drug Administration (FDA) has approved 5 drugs that temporarily improve the condition of patients suffering from AD, none is fully effective because of associated toxic effects [1]. Tacrine, donepezil, rivastigmine, and memantine, for example, have significant side effects such as elevation of serum aminotransferase concentration, nausea, vomiting, diarrhea, anorexia, anxiety, and agitation [5–7]. The toxic effects of these drugs necessitate the development of new therapeutic compounds.

To develop a new drug, a computational approach is worthwhile and saves time. This approach involves screening new ligands for a specific target within a relatively short span

of time. High throughput virtual screening (HTVS) is one of the most effective and rapid approaches for identifying probable inhibitors of the target protein [8]. Various potential drug targets have been reported to improve AD-associated pathological features such as acetylcholine esterases [9], NMDA receptor [10], and apolipoprotein E4 (ApoE4). ApoE plays a significant role in maintaining and repairing neurons. ApoE has three isoforms, namely, ApoE2, ApoE3, and ApoE4. The isoforms differ at residue positions 112 and 158 [11]. ApoE4 is the major genetic risk attributed to AD [12–17]. It acquires a pathological conformation through an intramolecular interaction, in which positively charged Arg 112 repels the side chain of Arg 61 in the aminoterminal domain, allowing the formation of a salt bridge between Arg 61 and Glu 255 at the carboxyl terminal domain [18, 19]. Forty to eighty percent of patients with AD are estimated to possess at least one ApoE4 allele [20]. ApoE4 is less effective in maintaining and repairing neuronal cells compared to ApoE2 and ApoE3 [21–23]. ApoE4 also disrupts the normal process by which cells release excess A$\beta$, resulting in elevated levels of A$\beta$ leading to its deposition in the brain [24–26]. ApoE4 uniquely performs neuron-specific proteolysis due to which harmful bioactive fragments are formed that can enter the cytosol, disrupt the mitochondrial energy balance, alter the cytoskeleton, and cause cell death [27–29]. ApoE is the only example of a susceptibility gene for AD [30] associated with lower glucose use and is believed to affect the hippocampus and cortex, areas found to be affected in patients with AD [31, 32]. It has been confirmed that the ApoE locus on chromosome 19 is strongly associated with the development of AD [12, 33, 34]. Small molecule structure correctors of ApoE4 have been suggested that effectively modulate the biophysical properties and the function of abnormal proteins. Some examples of ApoE4 structure correctors are GIND25 [35] and phthalazinone derivatives [36]. The evidential association of ApoE4 with increased risk of AD makes it a potential drug target for designing natural drug candidates for AD.

The present study focuses on identifying potential natural drug candidates as structure correctors for ApoE4. Keeping this goal in mind, a large database of natural compounds was screened against the 3D structure of ApoE4 using high throughput technology. *In silico* screening led to the identification of a new class of ApoE4 structure correctors that abolish the ApoE4 domain interaction. The molecular dynamics (MD) were then simulated to examine the dynamic behavior of molecular interactions between the screened compounds and the functional residues of ApoE4. This study paves the way for the development of novel leads for AD treatment that have improved binding properties and pose low toxicity to humans.

## 2. Materials and Methods

*2.1. Protein Preparation.* The crystal structure of human ApoE4 [PDB ID: 1GS9], determined at a resolution of 1.70 Å, was retrieved from the Protein Data Bank [37]. ApoE4 contains a single domain of 22 kD. To preprocess the retrieved structure of ApoE4, Protein Preparation Wizard in Schrodinger's Maestro interface [38] was used, followed by optimization [39].

*2.2. Grid Generation and Ligand Library Preparation.* The prepared protein structure was used to generate a grid using the receptor grid generation utility of the Glide docking module of the Schrodinger suite [40, 41]. Residues Arg-61, Glu-109, and Arg-112 form the catalytic triad in the active cleft of ApoE4 [36, 42]. The ligand library was prepared by extracting approximately 0.2 million natural compounds from the ZINC database [43] and processing them with Schrodinger's LigPrep Wizard [44] and using the Lipinski filter.

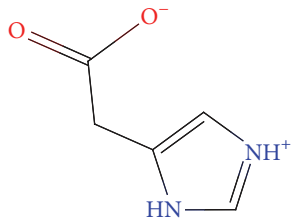
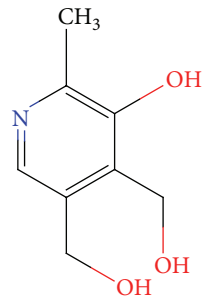
*2.3. High Throughput Virtual Screening and Docking Studies.* The prepared ligand library was screened with the Glide Program [41, 45]. Glide uses a systematic method for virtual screening based on incremental construction searching and provides the output as the GScore scoring function combined with various other parameters. Glide's HTVS and extraprecision (XP) algorithms combine to perform docking [46]. The screening against ApoE4 at the desired grid coordinates was performed through the HTVS docking algorithm [40]. Compounds with a significant docking score were subjected to Glide XP, a more precise docking algorithm for further refined screening.

*2.4. Molecular Dynamics Simulations of Docked Complexes.* The MD were simulated to study the dynamical behavior of the top-scoring docked complexes using the GROMACS package [47]. Initially, amber force fields were applied using the Amber tool package [48]. GROMACS topology files were created by converting amber topology files using the AnteChamber Python Parser interface script. To get electrically neutral complexes, the complexes were solvated in a cubic box of water molecules, and appropriate counterions were added. The solvated system was minimized for about 10,000 steps using the steepest descent and conjugate gradient methods until the force on each atom was less than 100 kJ/mol/nm. The geometrically minimized systems were then subjected to isothermal molecular dynamics simulations.

## 3. Results and Discussion

*3.1. Outcomes of High Throughput Virtual Screening and Docking Studies.* Human ApoE4, one of the most promising drug targets for treating AD, was virtually screened against approximately 0.2 million compounds of the ZINC database. The screened compounds were ranked according to their binding affinity, calculated as the scoring function called the Glide GScore. Of all compounds, a total of 10,000 compounds were identified from HTVS out of which those with a Glide score of less than −6.0 (64 compounds) were subjected to the Glide XP docking protocol. The top two scoring compounds and their properties are listed in Table 1. The values of the other docking parameters used for evaluating the selection criteria of the top-scoring ligands are shown in Table 2.

Table 1: Physical properties of potential structure correctors identified using virtual screening.

| Compound ZINC ID | Structure | log *P* value | Mol.wt. (g/mol) | HBD | HBA | tPSA (Å$^2$) | Heavy atoms |
|---|---|---|---|---|---|---|---|
| ZINC19735138 |  | −0.41 | 126.11 | 2 | 4 | 70 | 9 |
| ZINC00049154 |  | −0.55 | 169.18 | 3 | 4 | 74 | 12 |

Mol.wt.: Molecular weight, HBD: hydrogen bond donor, HBA: hydrogen bond Acceptor, tPSA: topological polar surface area.

Table 2: Binding affinity scores and energies of ApoE4 in complex with IAH and PHC.

| Compound | ZINC ID | Docking score | XP Gscore | Glide ligand efficiency | Glide evdw | Glide emodel | Glide energy |
|---|---|---|---|---|---|---|---|
| IAH | ZINC19735138 | −6.79 | −6.79 | 0.75 | −3.28 | −23.90 | −28.96 |
| PHC | ZINC00049154 | −6.76 | −6.76 | −0.56 | −6.18 | −32.97 | −26.97 |

The top-scoring compound (4-imidazoleacetic acid hydrochloride; ZINC19735138; IAH) had a Glide score of −6.79 kcal/mol, while the second compound (2-methyl, 3-hydroxy-4,5-dihydroxymethylpyridin or pyridoxine hydrochloride; ZINC00049154; PHC) had a score of −6.76 kcal/mol. The results revealed that IAH had a stronger binding affinity for human ApoE4 protein than PHC. Both ligands interacted with the two catalytic triad residues of ApoE4 in addition to other neighboring residues of the active site.

### 3.2. Binding Mode Analysis of Ligand-Docked ApoE4 Complexes

*3.2.1. ApoE4-IAH Complex.* In the case of the ApoE4-IAH complex, IAH interacted with the active site residues of ApoE4 (Figure 1(a)) with the formation of 3 hydrogen bonds and numerous hydrophobic contacts. Arg 61, Asp 65, and Glu 109 were the residues participating in hydrogen bond formation (Figure 1(b)). The NE and $NH_2$ atoms of basic catalytic amino acid Arg-61 formed 2 hydrogen bonds (3.28 Å, 2.73 Å) with the O atom of IAH. Other hydrogen bonds (2.49 Å, 2.71 Å) were formed by atom $N_1$ of IAH with the $OE_2$ atom of acidic active site residue Glu 109 and OD1 of neighboring acidic residue Asp-65 and atom N2 of IAH. In addition, Met-64 was involved in hydrophobic interaction in the ApoE4-IAH complex (Figure 1(c)). Among all these interacting residues, Arg 61 and Glu 109 (part of the catalytic triad) are crucial amino acids and play a prominent role in abolishing the structural orientation of ApoE4. IAH bound to these residues, thus preventing interaction among them and improving the functionality of ApoE4. Various chemical properties of IAH were considered that supported its drug-likeness for AD treatment (Table 1). The topological polar surface area was reasonably high, which indicated that it can readily be absorbed in the human intestine and can penetrate the blood-brain barrier (BBB). In IAH, the presence of 12 heavy atoms and a high potential energy of 50.33 kcal/mol suggested that this ligand molecule has a good binding affinity for human ApoE4.

*3.2.2. ApoE4-PHC Complex.* PHC is a single ringed structure with a molecular weight of 169.18 g/mol and lipophilicity value (log*P*) of −0.55 at pH 7. The topological polar surface area of PHC was also considered as it is very useful for identifying drug transport properties, human intestinal absorption, and BBB infiltration. The presence of a reasonable number of heavy atoms (9) and a good potential energy of 74 kcal/mol suggest that PHC is capable of binding strongly with ApoE4 (Figure 2(a)). In this study, PHC formed 4 hydrogen bonds and 1 hydrophobic contact with human ApoE4. As can be seen in Figure 2(b), 2 hydrogen bonds were formed between the $NH_2$ and NE atoms of active site residue Arg 61 and the $O_3$ atom of PHC with bond length 2.67 Å and 2.74 Å, respectively, while 2 others were formed with the OD1 and OD2 atoms of the neighboring residue Asp 65 and O1 and

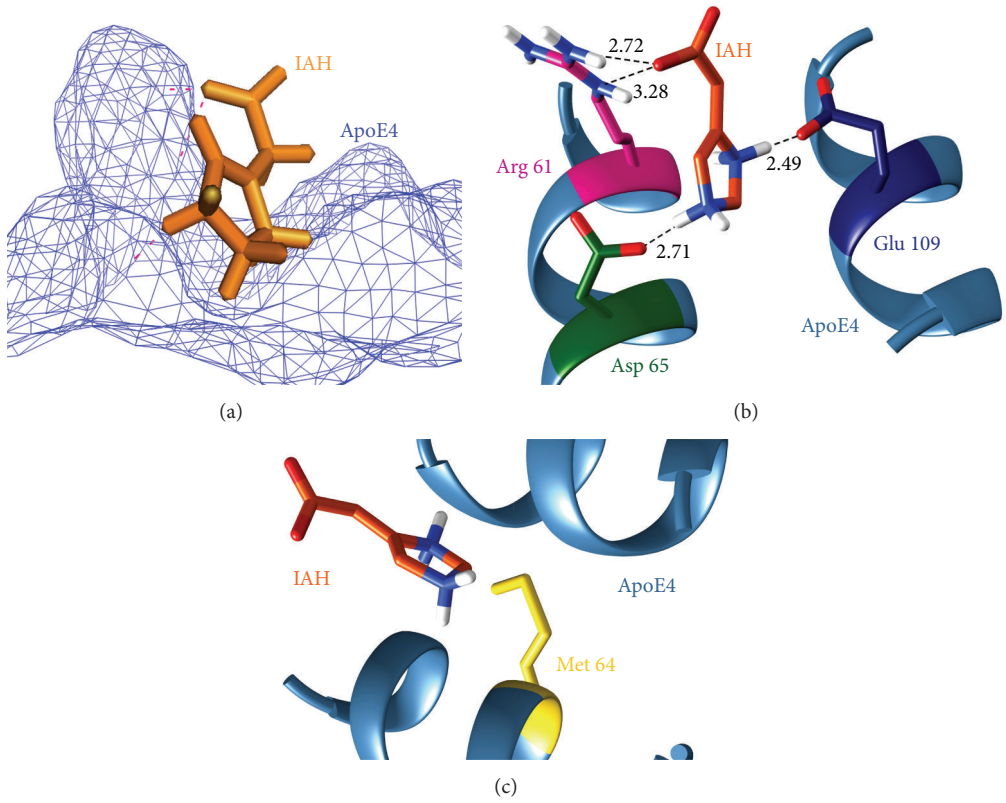



Figure 1: Molecular interactions between IAH (orange) and ApoE4 before MD simulations. (a) Position of IAH in the ligand-bound ApoE4 complex. (b) Hydrogen bond interactions. (c) Hydrophobic interactions.

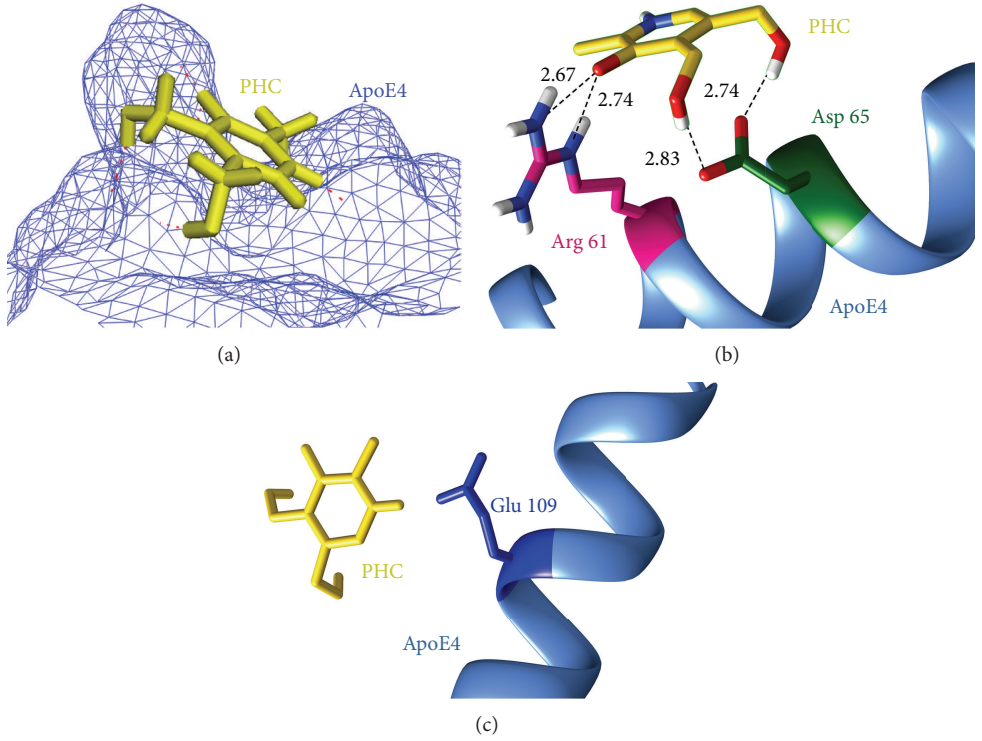



Figure 2: Molecular interactions between PHC (yellow) and ApoE4 before MD simulations. (a) Position of PHC in the ligand-bound docked complex. (b) Hydrogen bond interactions. (c) Hydrophobic interactions.

Table 3: Molecular interactions present in pre- and post-MD simulated IAH-bound ApoE4 complexes.

| ApoE4-IAH complex | Residues participating in hydrogen bonding | Residues governing hydrophobic contacts | Hydrogen bond length (Å) |
|---|---|---|---|
| Pre-MD | Arg-61 | Met-64 | 3.28, 2.73 |
| | Asp-65 | | 2.71 |
| | Glu-109 | | 2.49 |
| Post-MD | Met-64 | Arg-61, Asp-65, Met-68, Arg-112 | 2.98 |
| | Gly-105 | | 3.14 |
| | Glu-109 | | 3.08 |

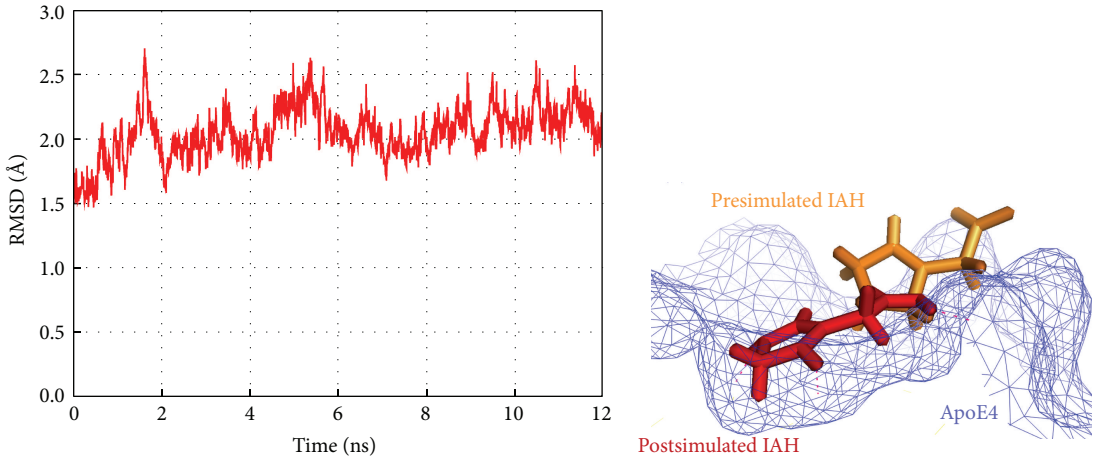



Figure 3: MD simulations trajectories: (a) RMSD trajectory of IAH in complex with ApoE4 obtained after MD simulations, (b) superimposition of pre-MD (orange) and post-MD (red) complexes of IAH with ApoE4.

O2 atoms of PHC with bond length of 2.83 Å and 2.74 Å, respectively. However, acidic amino acid Glu 109 of the catalytic triad was involved in making hydrophobic contact with PHC as illustrated in Figure 2(c). Of all these residues, Arg 61 and Glu 109 as part of the catalytic triad are responsible for the structural aberration in the human ApoE4 protein. These interactions of PHC with the crucial residues of ApoE4 suggest that this is a promising ligand that could correct the functionality of abnormal ApoE4.

### 3.3. Molecular Dynamics Simulations of Ligand-Bound ApoE4 Complexes

*3.3.1. Interaction Analysis of the ApoE4-PHC Complex.* For further refinement and stabilization of both docked complexes, the MD were simulated using the GROMACS package. The simulation lengths used in the study were long enough to allow rearrangement of the side chains of the native and the ligand-complexed protein thus facilitating the most stable binding mode. As is evident in Figure 3(a), the backbone of the protein acquired stability after 8 ns with a root mean square deviation (RMSD) of only about 2.5 Å from its initial position. However, the MD simulations for ApoE4-PHC complex conducted for up to 24 ns revealed interesting results. PHC moved away from the binding site of ApoE4 during the simulations and lost all interactions formed in the initial docked pose. Figure 4 illustrates the binding instability snapshots of PHC with ApoE4 during the simulation trajectory. During the MD simulations, the position of PHC in the ligand-bound complex was constantly altered. As can be seen from the snapshots at 6 ns and 8 ns, PHC moved far away from the binding site while staying at the surface of the protein. However, at 20 ns PHC was highly destabilized and split. Thus, it can be inferred that during the docking procedure the interactions of PHC with residues Arg 61, Arg 65, and Glu 109 of ApoE4 were only the result of static contacts. These pseudointeractions readily vanished when dynamics was considered in the study.

*3.3.2. Interaction Analysis of MD-Stabilized ApoE4-IAH Complex.* In the energetically stable ApoE4-IAH complex, the IAH molecule interacted with the residues Arg 61, Glu 109, and Arg 112 of the catalytic triad of ApoE4. The IAH molecule also formed contact with the residues Met 64, Asp 65, Met 68, and Gly 105. Though some deviation of IAH was observed from its initial position leading to a change in its binding mode, the binding was stable inside the ApoE4 cavity. A comparative analysis of the interaction profiles of ApoE4-IAH complex before and after the MD simulations is described in Table 3. The superimposition of the ligand IAH in the pre- and post-MD simulated complex structures inside the active site of ApoE4 is depicted in Figure 3(b).

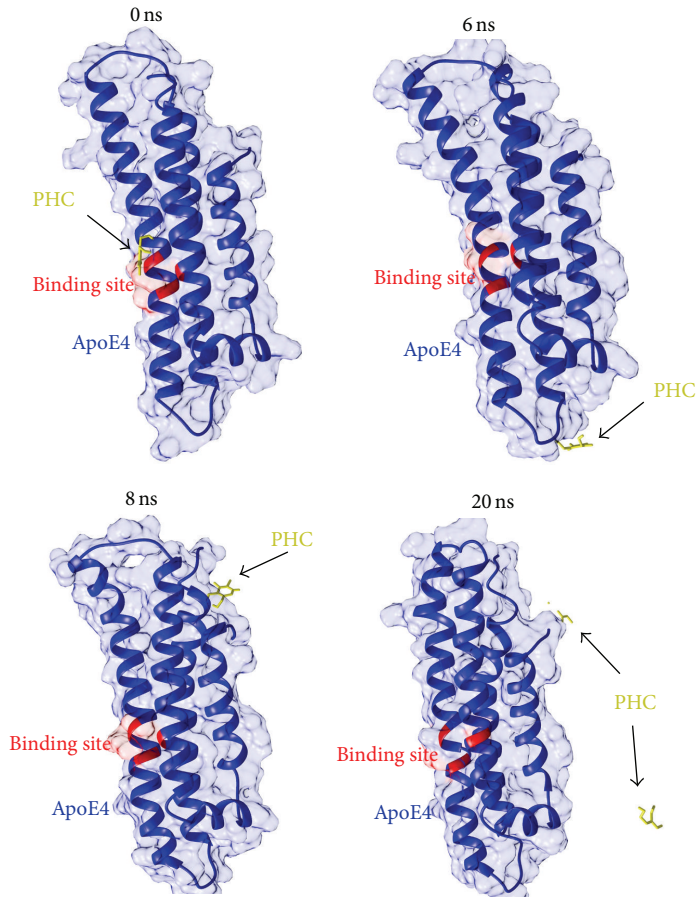



Figure 4: Snapshots depicting the binding instability of PHC with APoE4 during the MD simulations trajectory.

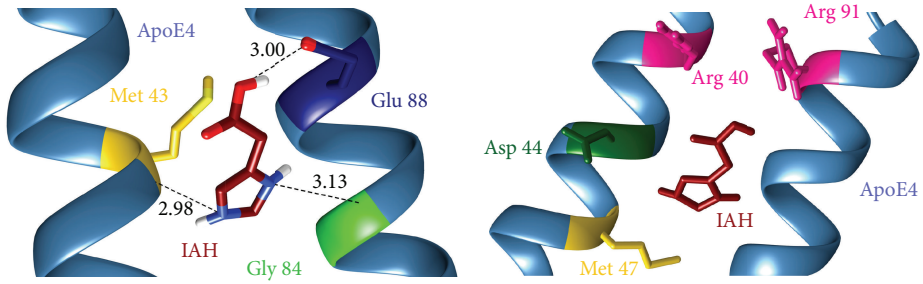



Figure 5: Molecular interactions between IAH (orange) and ApoE4 after MD simulations: (a) Hydrogen bond interactions and (b) hydrophobic interactions.

Initially, IAH formed 4 hydrogen bonds with the residues Arg 61, Asp 65, and Glu 109 of ApoE. After the simulations, 3 hydrogen bonds with the residues Arg 61 and Asp 65 had been replaced with 2 new hydrogen bonds involved with amino acids Gly 105 and Met 64. The hydrogen bond with the residue Glu 109 remained consistent with a slight change in the bond length (Figure 5(a)). The only hydrophobic contact with Met 64 was present in IAH-bound ApoE4 before MD disappeared during the MD simulations. However, after the MD simulations IAH formed strong hydrophobic contacts with 4 residues of ApoE4 (Figure 5(b)). The stability of IAH in the binding pocket of APoE4 is prominently governed by these hydrophobic contacts. After the MD simulations, IAH acquired a more stable conformation within the active site of ApoE4 by placing itself deep inside the cavity.

## 4. Conclusion

In the present work, we screened two top-scoring compounds, IAH and PHC, which possess high Glide XP scores of −6.79 kcal/mol and −6.76 kcal/mol, respectively, against human ApoE4. These compounds interacted with

the catalytic triad residues of ApoE4 that are crucial for maintaining its aberrant structure. The binding of these ligands suggests that they have a strong mechanistic ability to correct the pathological structural orientation of ApoE4 by preventing repulsion between Arg 61 and Arg 112, thus inhibiting the formation of a salt bridge between Arg 61 and Glu 255. The chemical properties of these potent structure-correctors are in line with the stipulated requirements of drug-like compounds for further experimental analysis. After the MD simulations, the interactions formed by IAH were consistent. However, a comparison between the conformations obtained from docking and that from molecular dynamics simulations for the second ligand PHC revealed substantial changes in binding conformations. Our simulation results indicate that the initial receptor-ligand interaction observed after docking can be limited due to the receptor rigid docking algorithm and that the conformations and interactions observed after the simulation runs are more energetically favored and should be better representations of the derivative poses in the receptor. Our detailed binding analysis of IAH substantiated by its dynamic structural stability provides considerable evidence for use as a potent natural lead against Alzheimer's. Results from this study would also be helpful in designing novel neuroregenerative drugs with improved binding properties and low toxicity.

## Acknowledgments

Abhinav Grover is thankful to the Science and Engineering Research Board, the Department of Science and Technology, Government of India, for the Fast Track Young Scientist Grant. The author also acknowledges support from Jawaharlal Nehru University for usage of all computational facilities.

## References

[1] "Alzheimer's Association: Alzheimer's disease facts and figures," *Alzheimer's & Dementia*, vol. 8, pp. 1–67, 2012.

[2] P. S. Mathuranath, A. George, N. Ranjith et al., "Incidence of Alzheimer's disease in India: a 10 years follow-up study," *Neurology India*, vol. 60, no. 6, pp. 625–630, 2012.

[3] M. J. Sadowski, J. Pankiewicz, H. Scholtzova et al., "Blocking the apolipoprotein E/amyloid-$\beta$ interaction as a potential therapeutic approach for Alzheimer's disease," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 49, pp. 18787–18792, 2006.

[4] D. L. Price, D. R. Borchlet, L. J. Martin, B. J. Crain, S. S. Sisodiya, and J. C. Troncoso, "Neuropathology of Alzheimer's disease and animal models," in *Neuropathology of Dementing Disorders*, W. R. Markesbery, Ed., pp. 121–141, 1998.

[5] E. L. Conway, "A review of the randomized controlled trials of tacrine in the treatment of Alzheimer's disease: methodologic considerations," *Clinical Neuropharmacology*, vol. 21, no. 1, pp. 8–17, 1998.

[6] S. I. Gracon, M. J. Knapp, W. G. Berghoff et al., "Safety of tacrine: clinical trials, treatment IND, and postmarketing experience," *Alzheimer Disease and Associated Disorders*, vol. 12, no. 2, pp. 93–101, 1998.

[7] R. Mayeux and M. Sano, "Treatment of Alzheimer's disease," *The New England Journal of Medicine*, vol. 341, no. 22, pp. 1670–1679, 1999.

[8] D. A. Ostrov, J. A. Hernández Prada, P. E. Corsino, K. A. Finton, N. Le, and T. C. Rowe, "Discovery of novel DNA gyrase inhibitors by high-throughput virtual screening," *Antimicrobial Agents and Chemotherapy*, vol. 51, no. 10, pp. 3688–3698, 2007.

[9] A. Lleó, S. M. Greenberg, and J. H. Growdon, "Current pharmacotherapy for Alzheimer's disease," *Annual Review of Medicine*, vol. 57, pp. 513–533, 2006.

[10] S. K. Sonkusare, C. L. Kaul, and P. Ramarao, "Dementia of Alzheimer's disease and other neurodegenerative disorders—memantine, a new hope," *Pharmacological Research*, vol. 51, no. 1, pp. 1–17, 2005.

[11] K. H. Weisgraber, "Apolipoprotein E: structure-function relationships," *Advances in Protein Chemistry*, vol. 45, pp. 249–302, 1994.

[12] W. J. Strittmatter, A. M. Saunders, D. Schmechel et al., "Apolipoprotein E: high-avidity binding to $\beta$-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 90, no. 5, pp. 1977–1981, 1993.

[13] E. H. Corder, A. M. Saunders, W. J. Strittmatter et al., "Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families," *Science*, vol. 261, no. 5123, pp. 921–923, 1993.

[14] R. W. Mahley, K. H. Weisgraber, and Y. Huang, "Apolipoprotein E4: a causative factor and therapeutic target in neuropathology, including Alzheimer's disease," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 15, pp. 5644–5651, 2006.

[15] R. W. Mahley, K. H. Weisgraber, and Y. Huang, "Apolipoprotein E: structure determines function, from atherosclerosis to Alzheimer's disease to AIDS," *Journal of Lipid Research*, supplement 50, pp. S183–S188, 2009.

[16] R. W. Mahley and Y. Huang, "Alzheimer disease: multiple causes, multiple effects of apolipoprotein E4, and multiple therapeutic approaches," *Annals of Neurology*, vol. 65, no. 6, pp. 623–625, 2009.

[17] J. Kim, J. M. Basak, and D. M. Holtzman, "The role of apolipoprotein E in Alzheimer's disease," *Neuron*, vol. 63, no. 3, pp. 287–303, 2009.

[18] L.-M. Dong, C. Wilson, M. R. Wardell et al., "Human apolipoprotein E. Role of arginine 61 in mediating the lipoprotein preferences of the E3 and E4 isoforms," *Journal of Biological Chemistry*, vol. 269, no. 35, pp. 22358–22365, 1994.

[19] L.-M. Dong and K. H. Weisgraber, "Human apolipoprotein E4 domain interaction. Arginine 61 and glutamic acid 255 interact to direct the preference for very low density lipoproteins," *Journal of Biological Chemistry*, vol. 271, no. 32, pp. 19053–19057, 1996.

[20] L. A. Farrer, L. A. Cupples, J. L. Haines et al., "Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease: a meta-analysis," *Journal of the American Medical Association*, vol. 278, no. 16, pp. 1349–1356, 1997.

[21] R. W. Mahley, "Apolipoprotein E: cholesterol transport protein with expanding role in cell biology," *Science*, vol. 240, no. 4852, pp. 622–630, 1988.

[22] R. W. Mahley and S. C. Rall Jr., "Apolipoprotein E: far more than a lipid transport protein," *Annual Review of Genomics and Human Genetics*, vol. 1, no. 2000, pp. 507–537, 2000.

[23] K. H. Weisgraber and R. W. Mahley, "Human apolipoprotein E: the Alzheimer's disease connection," *FASEB Journal*, vol. 10, no. 13, pp. 1485–1494, 1996.

[24] R. E. Tanzi and L. Bertram, "Twenty years of the Alzheimer's disease amyloid hypothesis: a genetic perspective," *Cell*, vol. 120, no. 4, pp. 545–555, 2005.

[25] J. Hardy and D. J. Selkoe, "The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics," *Science*, vol. 297, no. 5580, pp. 353–356, 2002.

[26] K. Blennow, M. J. de Leon, and H. Zetterberg, "Alzheimer's disease," *The Lancet*, vol. 368, no. 9533, pp. 387–403, 2006.

[27] M. E. Risner, A. M. Saunders, J. F. B. Altman et al., "Efficacy of rosiglitazone in a genetically defined population with mild-to-moderate Alzheimer's disease," *Pharmacogenomics Journal*, vol. 6, no. 4, pp. 246–254, 2006.

[28] Y. Huang and R. W. Mahley, "Commentary on "perspective on a pathogenesis and treatment of Alzheimer's disease." Apolipoprotein E and the mitochondrial metabolic hypothesis," *Alzheimer's & Dementia*, vol. 2, no. 2, pp. 71–73, 2006.

[29] A. D. Roses and A. M. Saunders, "Perspective on a pathogenesis and treatment of Alzheimer's disease," *Alzheimer's & Dementia*, vol. 2, no. 2, pp. 59–70, 2006.

[30] A. M. Saunders, M. K. Trowers, R. A. Shimkets et al., "The role of apolipoprotein E in Alzheimer's disease: pharmacogenomic target selection," *Biochimica et Biophysica Acta*, vol. 1502, no. 1, pp. 85–94, 2000.

[31] E. M. Reiman, R. J. Caselli, K. Chen, G. E. Alexander, D. Bandy, and J. Frost, "Declining brain activity in cognitively normal apolipoprotein E $\varepsilon 4$ heterozygotes: a foundation for using positron emission tomography to efficiently test treatments to prevent alzheimer's disease," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 6, pp. 3334–3339, 2001.

[32] E. M. Reiman, K. Chen, G. E. Alexander et al., "Functional brain abnormalities in young adults at genetic risk for late-onset Alzheimer's dementia," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 1, pp. 284–289, 2004.

[33] A. M. Saunders, W. J. Strittmatter, D. Schmechel et al., "Association of apolipoprotein E allele $\varepsilon 4$ with late-onset familial and sporadic Alzheimer's disease," *Neurology*, vol. 43, no. 8, pp. 1467–1472, 1993.

[34] M. A. Pericak-Vance, L. H. Yamaoka, C. S. Haynes et al. et al., "Genetic linkage studies in Alzheimer's disease families," *Experimental Neurology*, vol. 102, no. 3, pp. 271–279, 1988.

[35] S. Ye, Y. Huang, K. Müllendorff et al., "Apolipoprotein (apo) E4 enhances amyloid $\beta$ peptide production in cultured neuronal cells: ApoE structure as a potential therapeutic target," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 51, pp. 18700–18705, 2005.

[36] H.-K. Chen, Z. Liu, A. Meyer-Franke et al., "Small molecule structure correctors abolish detrimental effects of apolipoprotein E4 in cultured neurons," *Journal of Biological Chemistry*, vol. 287, no. 8, pp. 5253–5266, 2012.

[37] "Protein Data Bank," http://www.rcsb.org/pdb/home/home.do.

[38] Schrodinger, *Maestro*, version 9, LLC, New York, NY, USA, 2009.

[39] S. Sreeramulu, H. R. A. Jonker, T. Langer, C. Richter, C. R. D. Lancaster, and H. Schwalbe, "The human Cdc37·Hsp90 complex studied by heteronuclear NMR spectroscopy," *Journal of Biological Chemistry*, vol. 284, no. 6, pp. 3885–3896, 2009.

[40] T. A. Halgren, R. B. Murphy, R. A. Friesner et al., "Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening," *Journal of Medicinal Chemistry*, vol. 47, no. 7, pp. 1750–1759, 2004.

[41] R. A. Friesner, J. L. Banks, R. B. Murphy et al., "Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy," *Journal of Medicinal Chemistry*, vol. 47, no. 7, pp. 1739–1749, 2004.

[42] C. Frieden and K. Garai, "Structural differences between apoE3 and apoE4 may be useful in developing therapeutic agents for Alzheimer's disease," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 23, pp. 8913–8918, 2012.

[43] J. J. Irwin and B. K. Shoichet, "ZINC—a free database of commercially available compounds for virtual screening," *Journal of Chemical Information and Modeling*, vol. 45, no. 1, pp. 177–182, 2005.

[44] Schrodinger, *Ligprep*, version 2.3, LLC, New York, NY, USA, 2009.

[45] Schrodinger, *Glide*, version 5.5, LLC, New York, NY, USA, 2009.

[46] M. Sándor, R. Kiss, and G. M. Keseru, "Virtual fragment docking by glide: a validation study on 190 protein-fragment complexes," *Journal of Chemical Information and Modeling*, vol. 50, no. 6, pp. 1165–1172, 2010.

[47] D. van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen, "GROMACS: fast, flexible, and free," *Journal of Computational Chemistry*, vol. 26, no. 16, pp. 1701–1718, 2005.

[48] D. A. Case, T. A. Darden, T. E. Cheatham et al., *AMBER 12*, University of California, San Francisco, Calif, USA, 2012.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 865181, 4 pages
http://dx.doi.org/10.1155/2013/865181

*Research Article*

# QPLOT: A Quality Assessment Tool for Next Generation Sequencing Data

**Bingshan Li,[1] Xiaowei Zhan,[2] Mary-Kate Wing,[2] Paul Anderson,[2] Hyun Min Kang,[2] and Goncalo R. Abecasis[2]**

[1] *Department of Physiology and Biophysics, Center for Human Genetics Research, Vanderbilt University, Nashville, TN 37232, USA*
[2] *Department of Biostatistics, Center for Statistical Genetics, University of Michigan, Ann Arbor, MI 48109, USA*

Correspondence should be addressed to Bingshan Li; bingshan.li@vanderbilt.edu and Goncalo R. Abecasis; goncalo@umich.edu

Received 13 September 2013; Accepted 15 October 2013

Academic Editor: Zhongming Zhao

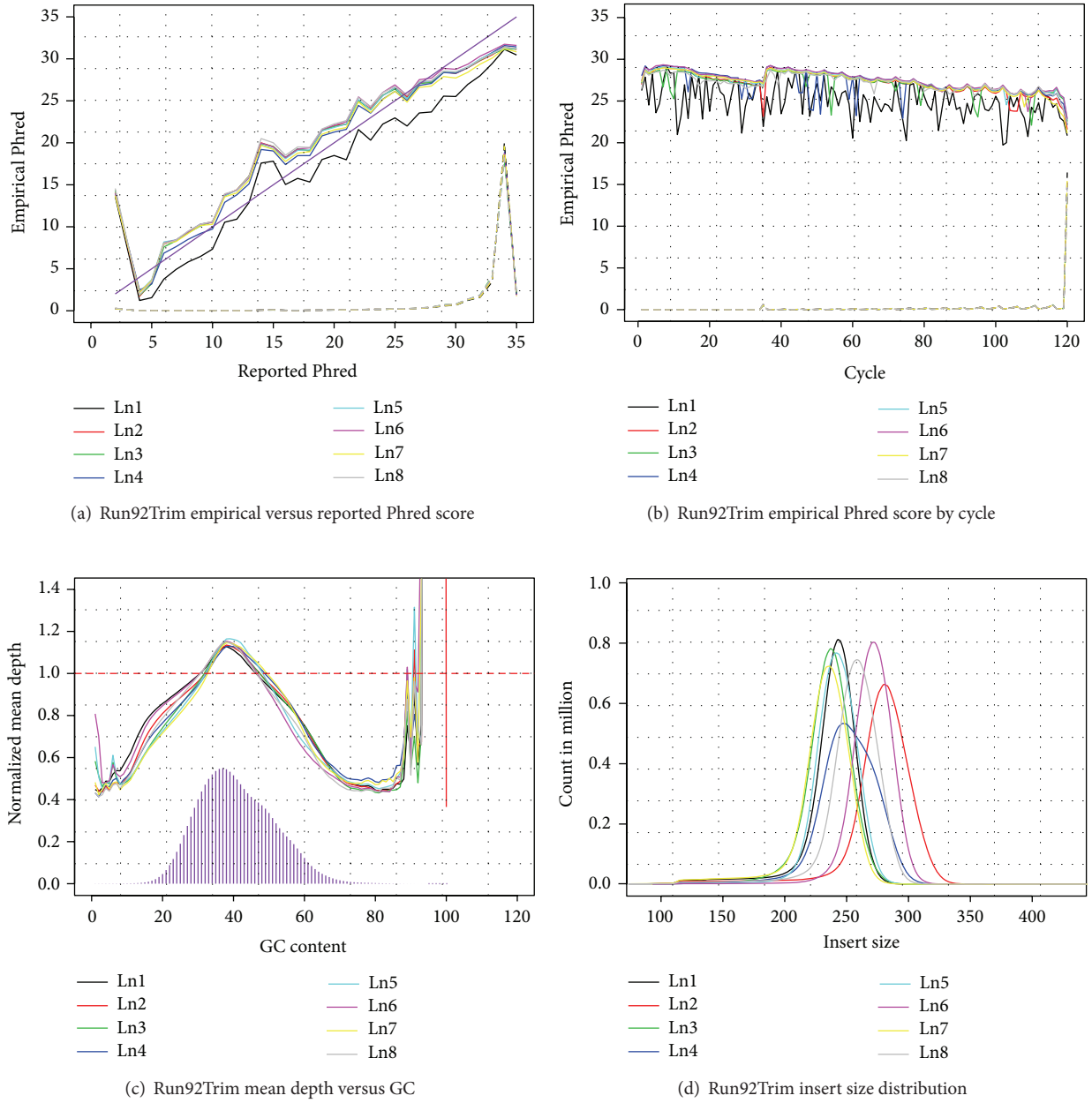


*Background.* Next generation sequencing (NGS) is being widely used to identify genetic variants associated with human disease. Although the approach is cost effective, the underlying data is susceptible to many types of error. Importantly, since NGS technologies and protocols are rapidly evolving, with constantly changing steps ranging from sample preparation to data processing software updates, it is important to enable researchers to routinely assess the quality of sequencing and alignment data prior to downstream analyses. *Results.* Here we describe QPLOT, an automated tool that can facilitate the quality assessment of sequencing run performance. Taking standard sequence alignments as input, QPLOT generates a series of diagnostic metrics summarizing run quality and produces convenient graphical summaries for these metrics. QPLOT is computationally efficient, generates webpages for interactive exploration of detailed results, and can handle the joint output of many sequencing runs. *Conclusion.* QPLOT is an automated tool that facilitates assessment of sequence run quality. We routinely apply QPLOT to ensure quick detection of diagnostic of sequencing run problems. We hope that QPLOT will be useful to the community as well.

## 1. Introduction

Next generation sequencing (NGS) is a revolutionary technology for biomedical research and is being deployed in a variety of applications, ranging from the identification of rare variants, *de novo* mutations, and somatic mutations in human disease studies to assessments of transcriptome and epigenome states in cultured cells. Since NGS provides more complete results than traditional array technologies and is rapidly decreasing in cost, it is becoming more widely used for genomics studies. Whole exome sequencing, which is the targeted sequencing of the entire collection of protein coding regions in the genome, has already led to great advances in Mendelian disorder genetics [1, 2], complex traits [3, 4],and cancer genomics [5, 6]. The 1000 Genomes Project [7, 8] is leading an effort to provide a comprehensive catalog of human variation across the world through whole genome sequencing. Several underway studies are now deploying whole genome and whole exome sequencing to study large collections of human disease samples.

The success of NGS studies depends on appropriately understanding the quality of underlying data. However, unlike traditional array platforms, analysis of sequencing data is much more complex, making real time monitoring of data quality more challenging. NGS technologies and associated set of protocols are constantly evolving, and updates to several different components of the process (including, for example, software, sample preparation, and/or reagents) can result in important and sometimes unexpected changes in data quality. We believe that the ability to generate automated visual summaries that help identify common problems is critical. To achieve this, we developed QPLOT, a tool for quick quality assessment in NGS data. QPLOT calculates and graphs summary statistics




(a) Run92Trim empirical versus reported Phred score

(b) Run92Trim empirical Phred score by cycle

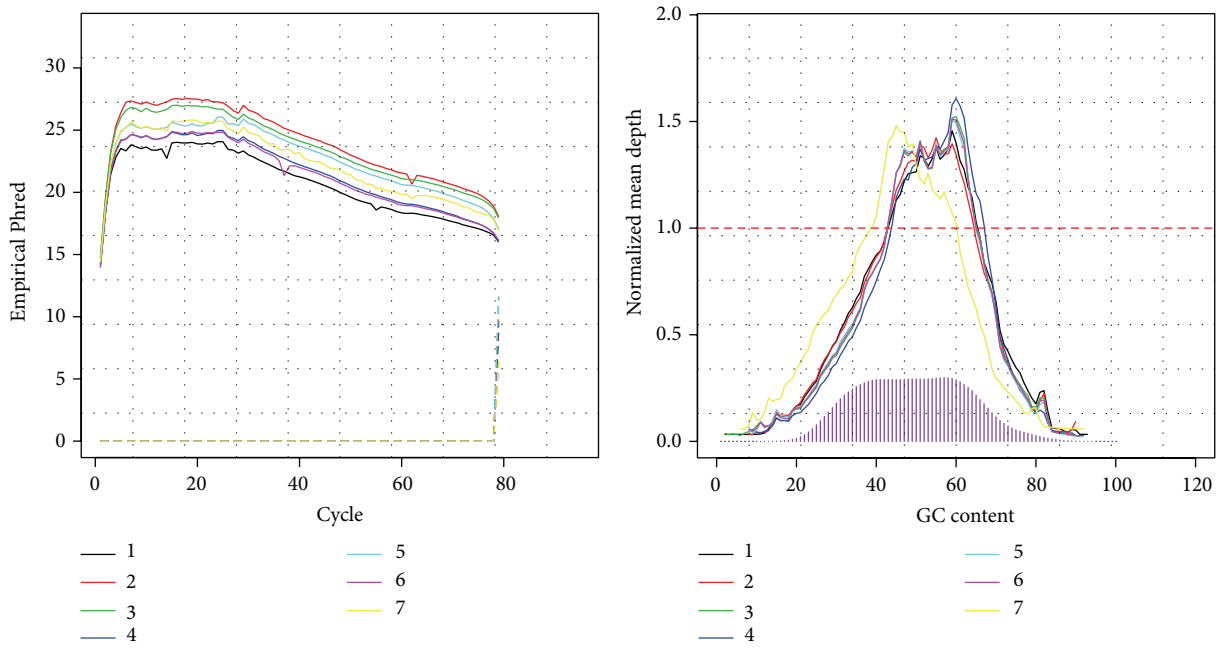
(c) Run92Trim mean depth versus GC

(d) Run92Trim insert size distribution

FIGURE 1: A subset of figures generated by QPLOT on an Illumina run. (a) Empirical base quality scores versus the scores stored in the BAM files. (b) Empirical base quality scores by cycles. (c) Bias of depth by GC content. (d) Insert size distribution.

describing sequence and alignment quality. Data quality is assessed both through reported base quality scores and empirically obtained metrics by comparing aligned bases to the reference genome. In this way, it is possible to track the number of high quality bases along the length of a read (to choose a read length that maximizes the yield of high quality bases and compare run quality over time) or identify the presence of adaptor sequence and other problems in alignment (these can result in high empirical mismatch rates near the ends of RNA-sequencing reads, due to difficulties in correctly placing splice junctions—a problem that can be ameliorated by excluding these bases from variant calling and RNA editing analyses *after* alignment). We constantly interact with our sequencing core and other collaborators generating sequence data to improve QPLOT and facilitate efforts to drive up the quality of next generation sequence data.

QPLOT differs from tools that only inspect unaligned sequence reads (such as FastQC [9] and SolexaQA [10]), because it can identify common problems in alignment and provide diagnostic descriptions of read mapping. For example, it generates empirically calibrated base quality scores and insert size distributions, two features that have substantial impact on variant calling and other downstream analyses. QPLOT also tries to improve packages designed specifically for handling aligned data (such as SAMStat [11] and Picard




(a) Empirical Phred score by cycle

(b) Mean depth versus GC

FIGURE 2: Exemplar diagnosis plots of RNA-sequencing data. (a) Empirical base quality scores by cycles. (b) Differential GC biases across multiple samples.

[12]) through its computational efficiency (QPLOT can sample regions of the genome randomly so as to rapidly evaluate very large alignments) and its ability to handle many samples (which helps to identify batch effects and other transient data processing problems). Importantly we note that genome-wide summary statistics can be extrapolated based on randomly sampled regions with little loss of accuracy. When the number of input files is very large, QPLOT can generate XML and text files with raw summary data and an interactive webpage that allows users to explore available quality metrics and graphs. XML and text output can be conveniently stored in a tracking database. In addition to graphical representation, key features are also summarized to generate a concise representation of the quality measurement (for example, a mean squared difference is used to summarize concordance of empirical and reported base quality scores, and the impact of GC content is summarized in a similar fashion based on the deviation of the depth for each GC bin from uniform coverage).

## 2. Materials and Methods

QPLOT is implemented in C++ and invokes R to generate figures. Available statistics include summaries of base quality, both overall and along each position in a read, comparisons of reported and empirical quality base scores, summaries of insert size for paired end libraries, global evaluations of coverage as well as more detailed evaluations of coverage as a function of GC content, and the regions targeted for enrichment. Empirical base scores are calculated as Phred scaled mismatch rates, that is, $-10 \times \log_{10}$ (number of matches/(number of matches + number of mismatches)), where number of matches and number of mismatches are the counts of aligned sequence bases that are concordant or discordant with the expected base in the reference genome, respectively, excluding known variant sites; these mismatches are dominated by genuine sequencing errors and provide a basis for base quality recalibration. To describe potential GC bias in sequencing runs, we calculate the mean depth of coverage for each GC content bin (0–100 representing 0–100% GC composition) for a series of windows along the genome (or, in the case of targeted sequencing experiments, within targeted regions). After normalization by the expected depth based on total mapped reads, the normalized depth for each GC content bin reflects biases of each experiment and can be compared with sequenced samples. Details of other summary statistics are available on the QPLOT website (http://genome.sph.umich.edu/wiki/QPLOT). QPLOT can be run as a stand-alone tool or incorporated into automated data processing pipelines.

## 3. Results and Discussion

We regularly use QPLOT in our sequencing projects including whole genome sequencing, RNA-seq, and targeted sequencing. Results for one Illumina run in a whole genome low pass sequencing study are shown in Figure 1. In this run the reported base quality scores deviate from empirically assessed quality, indicating that base quality recalibration is recommended (Figure 1(a)). As expected, empirical base quality scores decrease with increasing position along reads (Figure 1(b)), which is typical of Illumina sequencing.

However at position 36 empirical quality scores appear to increase, an artifact of the −q 15 option used in BWA [13] when mapping these data. The −q 15 option trims portions of reads with base quality <15, but always leaves at least 36 bases in each read (in our experience, this option increases the fraction of mapped reads and the number of mapped high quality bases). In this run, sequences with very high or very low GC content are underrepresented (below 1 in the relative depth curve, Figure 1(c)). Assessment of paired reads shows a distribution of insert sizes with peaks ranging from ~240 bp to ~300 bp (Figure 1(d)). In this case, since reads are 120 bases long, many paired reads overlap (particularly in lanes 1, 3, 5, and 7);these overlaps, if ignored, can result in PCR artifacts that look like sequence variants—suggesting that the protocol might be tweaked to increase library insert sizes. When we compared metrics generated by evaluating the complete data and those extrapolated from random 5 Mb segments of the genome, the two sets of summary statistics were remarkably similar (see QPLOT webpage for examples), but computing time was reduced from 38 minutes to 13 minutes.

In a second example, Figure 2 summarizes the results of an RNA-sequencing run. Here, empirical base quality scores are unexpectedly low near the beginning of each read (Figure 2(a)). When we remapped all reads after trimming the first several bases, the same pattern was repeated, suggesting that the observation is not due to high sequencing error rates or residual adapter sequences (trimming and remapping usually solve problems with residual adapter sequences, in our experience). Instead, the observation is the result of alignment artifacts when exon boundaries fall near the beginning or end of reads, a common problem in RNA-sequencing analyses. To avoid artifacts in downstream analyses, we suggest trimming the beginning and end bases of each read *after* mapping. Figure 2(b) shows that lane 7 has a GC content pattern that is dramatically different from the others, recommending great caution before comparing gene expression levels estimated for that sample and the others [14].

## 4. Conclusions

NGS has revolutionized the way genomics and biomedical studies are conducted. However the technologies are still rapidly evolving, and analysis of NGS data is challenging. Simple and convenient tools are important to help monitor data production and processing. Here we describe QPLOT, a computationally efficient tool that we hope will be helpful in quality assessment and diagnosis of NGS performance. We hope that information conveyed in these plots and statistics will facilitate the understanding of sequencing data to enable improved downstream processing and constant quality improvements.

## Authors' Contribution

Bingshan Li and Xiaowei Zhan should be regarded as joint first authors.

## Acknowledgments

The authors would like to thank the University of Michigan DNA Sequencing Core, especially Brendan Tarrier and Christine Brennan, for their helpful feedbacks on QPLOT.

## References

[1] S. B. Ng, A. W. Bigham, K. J. Buckingham et al., "Exome sequencing identifies *MLL2* mutations as a cause of kabuki syndrome," *Nature Genetics*, vol. 42, no. 9, pp. 790–793, 2010.

[2] 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.

[3] J. A. Tennessen, A. W. Bigham, T. D. O'Connor et al., "Evolution and functional impact of rare coding variation from deep sequencing of human exomes," *Science*, vol. 337, no. 6090, pp. 64–69, 2012.

[4] M. J. Emond, T. Louie, J. Emerson, W. Zhao, R. A. Mathias et al., "Exome sequencing of extreme phenotypes identifies *DCTN4* as a modifier of chronic *Pseudomonas aeruginosa* infection in cystic fibrosis," *Nature Genetics*, vol. 44, pp. 886–889, 2012.

[5] D. M. Muzny, M. N. Bainbridge, K. Chang et al., "Comprehensive molecular characterization of human colon and rectal cancer," *Nature*, vol. 487, pp. 330–337, 2012.

[6] P. S. Hammerman, M. S. Lawrence, D. Voet et al., "Comprehensive genomic characterization of squamous cell lung cancers," *Nature*, vol. 489, pp. 519–525, 2012.

[7] G. R. Abecasis, D. Altshuler, A. Auton et al., "A map of human genome variation from population-scale sequencing," *Nature*, vol. 467, pp. 1061–1073, 2010.

[8] G. R. Abecasis, A. Auton, L. D. Brooks et al., "An integrated map of genetic variation from 1,092 human genomes," *Nature*, vol. 491, pp. 56–65, 2012.

[9] A. Simon, http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

[10] M. P. Cox, D. A. Peterson, and P. J. Biggs, "SolexaQA: at-a-glance quality assessment of lllumina second-generation sequencing data," *BMC Bioinformatics*, vol. 11, article 485, 2010.

[11] T. Lassmann, Y. Hayashizaki, and C. O. Daub, "SAMStat: monitoring biases in next generation sequencing data," *Bioinformatics*, vol. 27, no. 1, pp. 130–131, 2011.

[12] A. Wysoker, http://picard.sourceforge.net.

[13] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, 2009.

[14] J. K. Pickrell, J. C. Marioni, A. A. Pai et al., "Understanding mechanisms underlying human gene expression variation with RNA sequencing," *Nature*, vol. 464, no. 7289, pp. 768–772, 2010.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 915636, 7 pages
http://dx.doi.org/10.1155/2013/915636

*Research Article*

# Comparative Study of Exome Copy Number Variation Estimation Tools Using Array Comparative Genomic Hybridization as Control

**Yan Guo,[1] Quanghu Sheng,[1] David C. Samuels,[2] Brian Lehmann,[3] Joshua A. Bauer,[3] Jennifer Pietenpol,[3] and Yu Shyr[1]**

[1] *Center for Quantitative Sciences, Vanderbilt University, Nashville, TN 37027, USA*
[2] *Center for Human Genetics Research, Vanderbilt University, Nashville, TN 37037, USA*
[3] *Department of Biochemistry, Vanderbilt University, Nashville, TN 37027, USA*

Correspondence should be addressed to Yan Guo; yan.guo@vanderbilt.edu and Yu Shyr; yu.shyr@vanderbilt.edu

Received 3 September 2013; Accepted 24 September 2013

Academic Editor: Zhongming Zhao



Exome sequencing using next-generation sequencing technologies is a cost-efficient approach to selectively sequencing coding regions of the human genome for detection of disease variants. One of the lesser known yet important applications of exome sequencing data is to identify copy number variation (CNV). There have been many exome CNV tools developed over the last few years, but the performance and accuracy of these programs have not been thoroughly evaluated. In this study, we systematically compared four popular exome CNV tools (CoNIFER, cn.MOPS, exomeCopy, and ExomeDepth) and evaluated their effectiveness against array comparative genome hybridization (array CGH) platforms. We found that exome CNV tools are capable of identifying CNVs, but they can have problems such as high false positives, low sensitivity, and duplication bias when compared to array CGH platforms. While exome CNV tools do serve their purpose for data mining, careful evaluation and additional validation is highly recommended. Based on all these results, we recommend CoNIFER and cn.MOPs for nonpaired exome CNV detection over the other two tools due to a low false-positive rate, although none of the four exome CNV tools performed at an outstanding level when compared to array CGH.

## 1. Introduction

Next-generation sequencing technology, piloted by the Illumina platform, has substantially decreased the cost of sequencing on large genomic regions. However it is still financially prohibitive to perform whole genome sequencing on a large number of subjects, especially for large scale genetic epidemiology association studies, at a sufficient depth for accurate genotype calls. The human exome represents about 1–3% of the human genome with approximately 30–50 million base pairs but accounts for over 85% of all mutations identified in Mendelian disorders [1]. As a result, exome sequencing is currently an attractive and practical approach for investigating coding variations.

Exome sequencing is typically used to identify single nucleotide polymorphisms (SNPs), somatic mutations (through paired sample comparison), and small and large structural variations. A lesser-known application of exome sequencing data is to identify copy number variations (CNV). CNVs are a structural variation in which cells have an abnormal number of copies of one or more sections of the DNA. Normal cells are diploid containing two copies of DNA and abnormal CNVs refer to large regions of the chromosome that have been deleted or duplicated. CNV characterization is important for both the basic understanding of many diseases and their diagnoses. CNVs have been linked to various diseases including autism [2], obesity [3], breast cancer [4], colorectal cancer [5], and lung cancer [6].

Traditionally, CNV detection has been performed with cytogenetic techniques such as fluorescent in situ hybridization, array comparative genomic hybridization (array CGH), and with virtual karyotyping using SNP arrays. Array CGH is commonly considered to be a reliable method for discovering

novel CNVs because of the relatively even distribution of probes [7]. Many high-impact copy number studies [8–10] were based on results derived from array CGH methods.

Whole genome sequencing data are relatively even in coverage, thus making it ideal for CNV discovery. Many CNV methods [11–17] have been developed for whole genome sequencing data. On the other hand, exome sequencing's depth is strongly affected by the enrichment regions, thus making it less ideal for CNV discovery. However, given the popularity of exome sequencing and the massive amount of exome sequencing data accumulated thus far, there is much interest in inferring CNVs from exome sequencing data. Thus, multiple CNV tools targeting exome sequencing data have been developed. We have cataloged sequencing data based CNV tools in Table S1 at the Supplementary Material available online at http://dx.doi.org/10.1155/2013/915636.

To determine if exome sequencing could provide reliable CNV detection, we performed array CGH and exome sequencing on 16 breast cancer cell lines. The data obtained from this study provides us an opportunity to evaluate the CNV discovery method based on exome sequencing while using array CGH as the reference. To date, there have been seven CNV tools targeting exome sequencing data: ExomeCNV [18], CoNIFER [19], cn.MOPS [20], exomeCopy [21], ExomeDepth [22], CNANorm [23], and CONTRA [24]. CNANorm, CONTRA, and ExomeCNV are specifically designed for paired tumor and normal samples. The other four do not require paired sample as input. These four tools cover a unique aspect of exome sequencing data. As exome sequencing become more commercially affordable, large epidemiology studies which only have blood samples available are more likely to choose exome sequencing over SNP array. The unpaired exome CNV tools will become the only suitable tools for CNV analysis. In this study, we systematically evaluated the performance of these four tools against each other using array CGH as the reference. We present our findings in detail and make a recommendation for the best unpaired exome CNV discovery tool based on our findings.

## 2. Materials and Methods

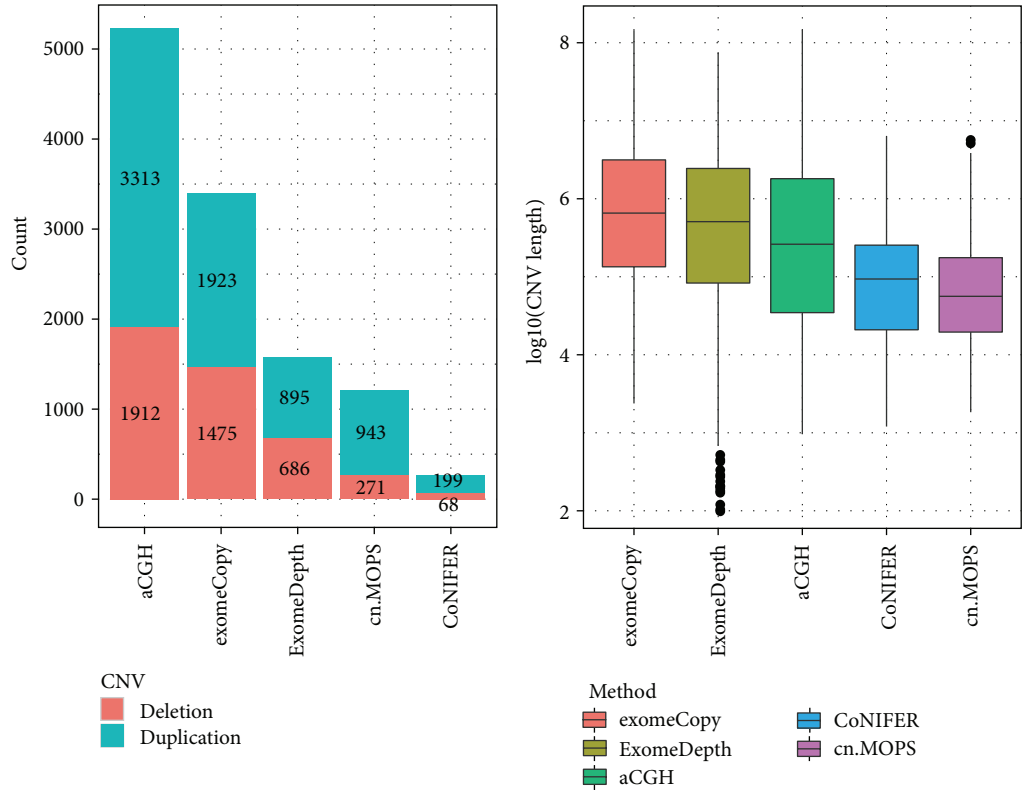
We performed array CGH on 16 breast cancer cell lines (Table S2) using the Agilent SurePrint G3 Human CGH Microarray Kit. This array CGH kit contains 963,029 distinct probes with 2.1 KB overall median probe spacing. The array CGH chips were scanned using the GenePix 4000B scanner, and probe intensities were normalized using Agilent's Feature Extraction software. CNVs were called using the Aberration Detection Method 2 (ADM2), a very broadly used CNV detection method for array CGH platform through GeneSpring Software. Exome sequencing data analysis was also performed on the same 16 breast cancer cell lines using Illumina's TrueSeq exome enrichment kit on Illumina's HiSeq 2000 platform. The sequencing reads are pair end 75 base pair long. The pooled, barcoded raw data produced by the Illumina HiSeq 2000 high-throughput sequencer was first split using barcode splitting software to obtain raw data for each individual. The raw data were aligned using BWA [25], which was designed based on the Borrows-Wheeler Transformation. The Human reference genome HG19 was used for alignment. The aligned BAM [26] files were locally realigned using the Genome Analysis Toolkit (GATK) [27] developed by the Broad Institute. The local realignment step aims to correct misalignment caused by the presence of insertions or deletions (indels). To further increase the local realignment accuracy, after local realignment, we performed base quality score recalibration on the realigned BAM files using GATK's recalibration tool. The recalibration tool attempts to correct for variations in quality with machine cycle and sequence context. The resulting BAM files contain not only more accurate base quality scores but also more widely dispersed ones. The recalibrated BAM files were filtered by removing all reads with mapping quality Phred score [28] less than 20 and all bases with base quality Phred score less than 20 (meaning that the probability of the base call being wrong is less than 0.01). CNVs on the processed BAM files were called using CoNIFER, cn.MOPS, exomeCopy, and ExomeDepth. Each of the four tools provides a wide range of parameters. We either consulted with the authors of the tools for the best parameters or used the author recommended parameters for the analysis. The exact command line used for each tool is listed in Table S3. Results of CNV detection from these four tools were compared to the array CGH results to determine the strength and weakness of each program.

## 3. Results and Discussion

*3.1. Results.* We generated high quality exome sequencing data using the Illumina TrueSeq enrichment kit on the HiSeq 2000 platform. All samples' raw data passed the initial quality control using FASTQC. On average, each sample had 117 million (range: 73 to 183 million) reads sequenced. The average capture efficiency was 48% (range: 39% to 62%). No notable quality issues were observed for the exome sequencing data (Table S2).

Across all 16 samples, array CGH identified 5,225 CNVs. Among the four exome CNV tools, exomeCopy identified the most CNVs (3,398), and CoNIFER identified the least (267). ExomeDepth (1,581) and cn.MOPS (1,214) identified a moderate number of CNVs (Figure 1(a)). The median CNV length identified by array CGH was 261,400 base pairs (range: 959 to 146,900,000 base pairs). ExomeDepth and exomeCopy identified the CNVs with longer average length than did array CGH, while CNVs identified by CoNIFER and cn.MOPS had shorter average length compared to array CGH (Figure 1(b)).

We also determined the deletion-duplication ratio for array CGH and the four exome CNV tools by sample. Each sample has distinct molecular characteristics that result in distinct deletion and duplication ratios. Consistently observing more duplication than deletions or vice versa across all samples may be an indication of an algorithm-specific bias. For array CGH, across all 16 samples, we observed 9 samples with more duplication and 7 samples with more deletions, a rather ideal scenario (Figure 2(a)). For exomeCopy and cn.MOPS, we observed 10 samples with more duplication




(a) (b)

Figure 1: Overview of the CNVs detected by array CGH and four algorithms. (a) Barplot of the duplication and deletion CNVs detected by five methods. (b) Boxplot of the CNV length detected by five methods.

Table 1: Kullback-Leibler test on similarity with array CGH.

| | aCGH | cn.MOPS | exomeCopy | ExomeDepth | CoNIFER |
|---|---|---|---|---|---|
| Deletion CNVs proportion similarity | | | | | |
| aCGH | 0 | 0.14 | 0.16 | 0.17 | 2.24 |
| cn.MOPS | 0.15 | 0 | 0.22 | 0.24 | 1.84 |
| exomeCopy | 0.16 | 0.23 | 0 | 0.2 | 1.88 |
| ExomeDepth | 0.22 | 0.27 | 0.23 | 0 | 2.56 |
| CoNIFER | 0.8 | 0.97 | 0.87 | 0.72 | 0 |
| Duplication CNVs proportion similarity | | | | | |
| aCGH | 0 | 0.4 | 0.4 | 0.47 | 0.66 |
| cn.MOPS | 0.36 | 0 | 0.42 | 0.66 | 0.61 |
| exomeCopy | 0.42 | 0.59 | 0 | 0.18 | 0.26 |
| ExomeDepth | 0.55 | 1.06 | 0.25 | 0 | 0.55 |
| CoNIFER | 0.77 | 0.58 | 0.26 | 0.42 | 0 |

and 6 samples with more deletions (Figures 2(b) and 2(c)). For ExomeDepth, we observed 11 samples with more duplication and 5 samples with more deletions (Figure 2(d)). For CoNIFER, we observed 14 samples with more duplication and 2 samples with more deletions (Figure 2(e)). We conducted paired Wilcoxon signed rank tests to see if there is any duplication or deletion bias. We found that array CGH showed unbiased duplication and deletions with $P$ value = 0.11, exomeCopy also showed unbiased duplication and deletions with $P$ value = 0.1. CoNIFER had strong bias toward duplication with $P$ values equal to 0.025. ExomeDepth and Cn.MOPS showed marginal bias toward duplication with $P$ value = 0.064. To identify the exome CNV tool with the most similar deletion duplication ratio, we conducted pairwise Kullback-Leibler divergence distance on both duplication and deletions proportions (Table 1). The values in Table 1 are measures of the difference between the tested method and the array CGH method, with smaller values indicating less difference.

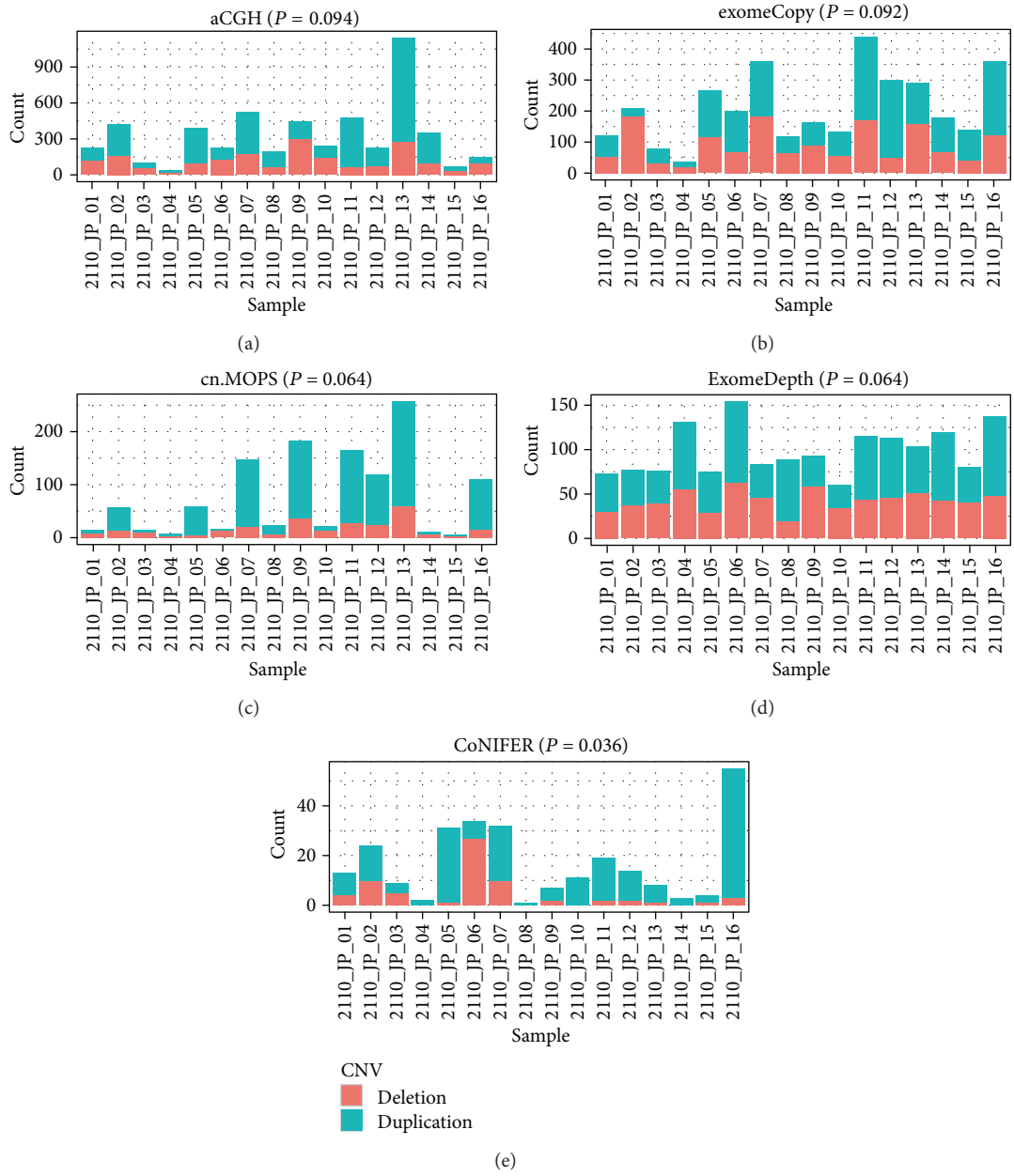



Figure 2: Barplot of duplication and deletion CNVs detected from each sample by five methods. The $P$ value beside each method name was calculated by paired Wilcoxon signed rank tests following FDR correction. It indicated the detection bias between duplication and deletion CNVs of that method. Array CGH and exomeCopy showed unbiased duplication and deletion while CoNIFER had strong bias toward duplication. Cn.MOPS and ExomeDepth showed marginal bias toward duplication.

For both duplication and deletions, cn.MOPS showed the shortest distance to array CGH, with 0.15 for deletions and 0.36 for duplication.

To measure the consistency with array CGH, we determined the overlap of CNVs identified between each of the exome CNV tools with array CGH. Overlapping CNVs were defined as regions that share at least 50% of their base pairs. We also used a less strict option where two CNVs are considered consistent if only 1% of the base pairs overlapped. However, regardless of which option we use, the results were very similar, since if two CNVs from two methods overlapped, most of them overlapped by at least 50% (Table S4). Compared to the array CGH platform, cn.MOPS had the best true positive rate for duplication with 76.9%, and CoNIFER had the best true positive rate for deletions with 83.8% (Figure 3). ExomeDepth and exomeCopy had comparable true positive rates for duplication with CoNIFER but lower true positive rates for deletions. Also interestingly, all four exome CNV tools identified some CNVs with opposite direction (deletion instead of duplication or vice versa) compared to array CGH. ExomeDepth and exomeCopy had relatively low proportion of CNVs with opposite direction

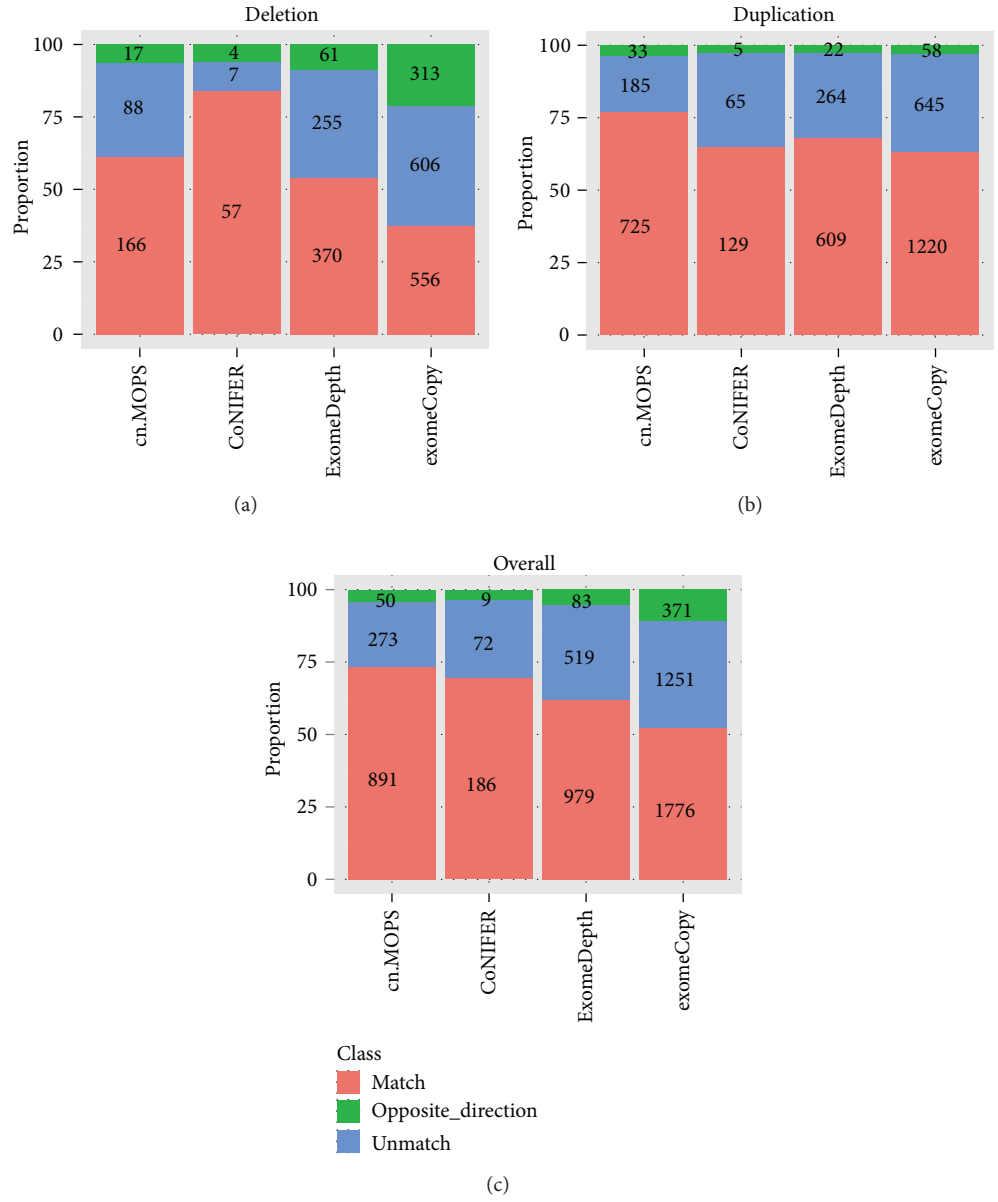



Figure 3: Specificity of four algorithms for CNV detection. CoNIFER identified many fewer CNVs but with a high true positive rate at deletion detection. ExomeDepth and ExomeCopy showed comparable specificity with CoNIFER on duplication detection but many more false positives on deletion detection. Cn.MOPS showed best specificity at duplication detection and second best specificity at deletion identified many more CNVs than CoNIFER. Overall, cn.MOPS achieved the highest specificity among all four algorithms.

on duplication (2.5% and 3.0%) but a moderate proportion of CNVs with opposite direction on deletion (8.9% and 10.9%), while cn.MOPS and ConNIFER had a relatively low proportion of CNVs with opposite direction (3.5% and 2.5% for duplication, 6.3% and 5.9% for deletion). CoNIFER and cn.MOPS detected CNVs with a much lower false-positive rate. In such a scenario, CoNIFER and cn.MOPS are much more desirable, because it is impossible to tell true-positives from false-positives without any prior knowledge. ExomeDepth and exomeCopy also demonstrated comparable performance for detecting duplication with CoNIFER.

*3.2. Discussion.* Exome sequencing is widely used to conduct genomic research. Identifying CNVs through exome sequencing data has been a popular topic over the last few years. Compared to array-based methods, identifying CNVs through exome sequencing data has some shortcomings. First, the exons within the genome are not evenly placed. They are located at fixed positions, unlike probes which can be designed to be placed evenly across the whole genome. Thus if only depth information from unevenly located exons can be used for CNV assessment then CNV detection over a long intergenic region would be unreliable. Also, a probe can

be designed to avoid hybridization at problematic genomic regions such as regions with high GC content. A high GC content region can affect the sequencing depth for exome sequencing, which makes identifying CNVs using exome sequencing data even more complicated. Additional normalization to correct noise caused by effects such as GC content is desirable.

With these known difficulties, many exome CNV tools have been developed over the last few years. In this study, we evaluated the effectiveness of four popular unpaired exome CNV tools: cn.MOPS, CoNIFER, ExomeDepth, and exomeCopy using 16 breast cancer cell lines. We identified CNVs using these four tools and verified the results against array CGH results from the same samples. CoNIFER and cn.MOPS identified much fewer CNVs but with a high true-positive rate. ExomeDepth and exomeCopy produced comparable performance for duplication detection with CoNIFER. In terms of duplication-deletion proportion, we found that with the exception of exomeCopy, the exome CNV tools showed a significant bias toward duplication. This could be the result of an artifact of the exome CNV algorithm that underestimates the normal copy number based on depth. Using the Kullback-Leibler divergence distance, we found that cn.MOPS is the closest to array CGH in terms of duplication or deletion proportion across samples. Based on all these results, we recommend CoNIFER and cn.MOPS for nonpaired exome CNV detection over the other two tools due to a low false-positive rate, although none of the four exome CNV tools performed at an outstanding level when compared to array CGH. In summary, there is value in identifying CNVs using exome sequencing data but extra caution needs to be taken into consideration due to the high false positive rate. Identifying CNVs is almost never the primary goal of the exome sequencing study, and it should stay that way due to the noise introduced by exome sequencing data. Identifying CNVs using exome sequencing data is potentially a good secondary data mining technique. Based on our comparison of the methods, results generated from exome CNV tools should be evaluated thoroughly, and additional validation is highly recommended to eliminate false-positives and to ensure quality data.

## 4. Conclusions

Using array CGH result as control, we systematically compared four popular exome CNV tools (CoNIFER, cn.MOPS, exomeCopy, and ExomeDepth) on exome sequencing data generated from 16 breast cancer cell lines. Among evaluated four tools, we recommend CoNIFER and cn.MOPS for nonpaired exome CNV detection due to a low false-positive rate. Our results suggest that exome CNV tools are subjected to high false positive rat, low sensitivity, and duplication bias when compared to array CGH platform. Thus careful evaluation and additional validation is highly recommended.

## Authors' Contribution

Yan Guo and Quanghu Sheng have equally contributed to this paper.

## Acknowledgments

All data were generated from Vanderbilt Technologies for Advanced Genomics. The authors would like to thank Margot Bjoring for her editorial support.

## References

[1] 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.

[2] J. T. Glessner, K. Wang, G. Cai et al., "Autism genome-wide copy number variation reveals ubiquitin and neuronal genes," *Nature*, vol. 459, no. 7246, pp. 569–572, 2009.

[3] T. L. Yang, Y. Guo, H. Shen et al., "Copy number variation on chromosome 10q26.3 for obesity identified by a genome-wide study," *The Journal of Clinical Endocrinology & Metabolism*, vol. 98, no. 1, pp. E191–E195, 2013.

[4] A. Bergamaschi, Y. H. Kim, P. Wang et al., "Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer," *Genes Chromosomes and Cancer*, vol. 45, no. 11, pp. 1033–1040, 2006.

[5] M. Moroni, S. Veronese, S. Benvenuti et al., "Gene copy number for epidermal growth factor receptor (EGFR) and clinical response to antiEGFR treatment in colorectal cancer: a cohort study," *The Lancet Oncology*, vol. 6, no. 5, pp. 279–286, 2005.

[6] F. Cappuzzo, C. Ligorio, L. Toschi et al., "EGFR and HER2 gene copy number and response to first-line chemotherapy in patients with advanced non-small cell lung cancer (NSCLC)," *Journal of Thoracic Oncology*, vol. 2, no. 5, pp. 423–429, 2007.

[7] A. Vasson, C. Leroux, L. Orhant et al., "Custom oligonucleotide array-based CGH: a reliable diagnostic tool for detection of exonic copy-number changes in multiple targeted genes," *European Journal of Human Genetics*, vol. 21, no. 9, pp. 977–987, 2013.

[8] H. Lybæk, L. A. Meza-Zepeda, S. H. Kresse, T. Høysæter, V. M. Steen, and G. Houge, "Array-CGH fine mapping of minor and cryptic HR-CGH detected genomic imbalances in 80 out of 590 patients with abnormal development," *European Journal of Human Genetics*, vol. 16, no. 11, pp. 1318–1328, 2008.

[9] J. Schoumans, K. Nielsen, I. Jeppesen et al., "A comparison of different metaphase CGH methods for the detection of cryptic chromosome aberrations of defined size," *European Journal of Human Genetics*, vol. 12, no. 6, pp. 447–454, 2004.

[10] K. Hashimoto, N. Mori, T. Tamesa et al., "Analysis of DNA copy number aberrations in hepatitis C virus-associated hepatocellular carcinomas by conventional CGH and array CGH," *Modern Pathology*, vol. 17, no. 6, pp. 617–622, 2004.

[11] A. Abyzov, A. E. Urban, M. Snyder, and M. Gerstein, "CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing," *Genome Research*, vol. 21, no. 6, pp. 974–984, 2011.

[12] V. Boeva, T. Popova, K. Bleakley et al., "Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data," *Bioinformatics*, vol. 28, no. 3, pp. 423–425, 2012.

[13] D. Y. Chiang, G. Getz, D. B. Jaffe et al., "High-resolution mapping of copy-number alterations with massively parallel sequencing," *Nature Methods*, vol. 6, no. 1, pp. 99–103, 2009.

[14] A. Magi, M. Benelli, S. Yoon, F. Roviello, and F. Torricelli, "Detecting common copy number variants in high-throughput

sequencing data by using JointSLM algorithm," *Nucleic Acids Research*, vol. 39, no. 10, article e65, 2011.

[15] P. Medvedev, M. Fiume, M. Dzamba, T. Smith, and M. Brudno, "Detecting copy number variation with mated short reads," *Genome Research*, vol. 20, no. 11, pp. 1613–1622, 2010.

[16] C. A. Miller, O. Hampton, C. Coarfa, and A. Milosavljevic, "ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads," *PLoS ONE*, vol. 6, no. 1, Article ID e16327, 2011.

[17] S. M. Waszak, Y. Hasin, T. Zichner et al., "Systematic inference of copy-number genotypes from personal genome sequencing data reveals extensive olfactory receptor gene content diversity," *PLoS Computational Biology*, vol. 6, no. 11, Article ID e1000988, 2010.

[18] J. F. Sathirapongsasuti, H. Lee, B. A. J. Horst et al., "Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV," *Bioinformatics*, vol. 27, no. 19, pp. 2648–2654, 2011.

[19] N. Krumm, P. H. Sudmant, A. Ko et al., "Copy number variation detection and genotyping from exome sequence data," *Genome Research*, vol. 22, no. 8, pp. 1525–1532, 2012.

[20] G. Klambauer, K. Schwarzbauer, A. Mayr et al., "cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate," *Nucleic Acids Research*, vol. 40, no. 9, article e69, 2012.

[21] M. I. Love, A. Myišičková, R. Sun, V. Kalscheuer, M. Vingron, and S. A. Haas, "Modeling read counts for CNV detection in exome sequencing data," *Statistical Applications in Genetics and Molecular Biology*, vol. 10, no. 1, article 52, 2011.

[22] V. Plagnol, J. Curtis, M. Epstein et al., "A robust model for read count data in exome sequencing experiments and implications for copy number variant calling," *Bioinformatics*, vol. 28, no. 21, pp. 2747–2754, 2012.

[23] A. Gusnanto, H. M. Wood, Y. Pawitan, P. Rabbitts, and S. Berri, "Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data," *Bioinformatics*, vol. 28, no. 1, pp. 40–47, 2012.

[24] 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.

[25] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, 2009.

[26] H. Li, B. Handsaker, A. Wysoker et al., "The sequence alignment/Map format and SAMtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009.

[27] A. McKenna, M. Hanna, E. Banks et al., "The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data," *Genome Research*, vol. 20, no. 9, pp. 1297–1303, 2010.

[28] 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.

Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 493019, 7 pages
http://dx.doi.org/10.1155/2013/493019

*Research Article*

# New aQTL SNPs for the CYP2D6 Identified by a Novel Mediation Analysis of Genome-Wide SNP Arrays, Gene Expression Arrays, and CYP2D6 Activity

**Guanglong Jiang,[1,2] Arindom Chakraborty,[1,2] Zhiping Wang,[1,2] Malaz Boustani,[3] Yunlong Liu,[1,2] Todd Skaar,[4] and Lang Li[1,2,4]**

[1] *Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN 46202, USA*
[2] *Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN 46202, USA*
[3] *Regenstrief Institute, Indianapolis, IN 46202, USA*
[4] *Division of Clinical Pharmacology, Department of Medicine, Indiana University School of Medicine, Indianapolis, IN 46202, USA*

Correspondence should be addressed to Lang Li; lali@iu.edu

Received 30 August 2013; Accepted 16 September 2013

Academic Editor: Zhongming Zhao



*Background*. The genome-wide association studies (GWAS) have been successful during the last few years. A key challenge is that the interpretation of the results is not straightforward, especially for transacting SNPs. Integration of transcriptome data into GWAS may provide clues elucidating the mechanisms by which a genetic variant leads to a disease. *Methods*. Here, we developed a novel mediation analysis approach to identify new expression quantitative trait loci (eQTL) driving CYP2D6 activity by combining genotype, gene expression, and enzyme activity data. *Results*. 389,573 and 1,214,416 SNP-transcript-CYP2D6 activity trios are found strongly associated ($P < 10^{-5}$, FDR = 16.6% and 11.7%) for two different genotype platforms, namely, Affymetrix and Illumina, respectively. The majority of eQTLs are trans-SNPs. A single polymorphism leads to widespread downstream changes in the expression of distant genes by affecting major regulators or transcription factors (TFs), which would be visible as an eQTL hotspot and can lead to large and consistent biological effects. Overlapped eQTL hotspots with the mediators lead to the discovery of 64 TFs. *Conclusions*. Our mediation analysis is a powerful approach in identifying the trans-QTL-phenotype associations. It improves our understanding of the functional genetic variations for the liver metabolism mechanisms.

## 1. Introduction

Genome-wide association studies (GWAS) have identified hundreds of genetic variants associated with complex human diseases, clinical conditions, and traits. These studies have also provided valuable insights into the genetic architecture. Unfortunately, GWAS studies have achieved limited success. The variants discovered usually explain only a small fraction of the overall heritability of the disease [1]. The identification of specific causal genes or mutations from associated regions is a challenge especially for the transacting SNPs which fall either far from genes or a region with many equally plausible causative genes. To make the situation more complicated, sometimes, a single locus can contain multiple independent risk variants (common or rare). Even when a locus is identified by SNP association, the causal mutation itself needs not to be a SNP [2]. For example, GWAS have associated the IRGM gene with Crohn's disease, but a subsequent study showed that the causal mutation is a deletion of the upstream of the promoter affecting tissue-specific expression [3].

There is a substantial gap in understanding the SNP traits associations from a genome-wide association study and the contribution of the locus to a disease. An eQTL approach investigates how the abundance of a gene transcript is directly modified by polymorphism in regulatory elements. The validity of eQTL has been shown in multiple tissue types,

in which high heritability has been observed in widespread gene transcripts [4–8]. This indicates that genetic influences on gene expression are common. The potential of genome-wide eQTL identification has been shown originally in the yeast *Saccharomyces cerevisiae* [9] and then in humans, animals, and plants [10, 11]. One of the most important consequences of eQTL mapping is the link that it provides between genetic markers of a disease identified in GWAS and the expression of a specific gene or genes. In particular, the power of these studies depends upon the identification of specific genetic markers that are simultaneously associated with a disease and eQTLs. For example, a study generated genome-wide transcriptional profiles of lymphocyte samples from participants in the San Antonio Family Heart Study and showed that high density lipoprotein cholesterol concentration was influenced by the cis-regulated VNN1 [5, 12]. Another study of postmortem brain tissue identified eQTLs affecting the MAPT and APOE genes, which play an important part in Alzheimer's disease. Utilizing human lymphoblastoid cell lines from the HapMap project, recent pharmacogenomics study reveals novel genetic variants that contribute to etoposide-induced toxicity through affecting gene expression, which included genes that may play a role in cancer (AGPAT2, IL1B, and WNT5B) [13].

The substantial gap between associated regions from GWAS and the identification of causal variations that contribute to a disease might be filled by eQTL analysis. The functional effects of DNA polymorphism on a multifactorial disease can be mediated through several mechanisms. Polymorphisms responsible for the alteration in protein function can have important effects. However, systematic studies of complex diseases with known nonsynonymous SNPs have not yielded many highly significant results, and many associations implicate nonprotein coding regions. It has been shown that 5% of the human genome is evolutionary conserved and thus functional, whereas less than one-third of this 5% consists of genes that encode proteins [2]. Variation in gene expression is probably a more important mechanism underlying susceptibility to complex disease [2, 14].

Three major different methodologies have been developed and applied to the integrated eQTL and GWAS analyses. The first method focused on the overlapped SNPtrait, SNP-gene expression, and gene expression-trait associations [13, 15]. The second method employed the causal inference framework to identify causal model, reactive model, and independent model among SNP, gene expression, and traits. This approach brought in a more molecular mechanism in analyzing the data [16]. The third approach constructed a Bayesian network for the gene expression and traits, while the network construction was weighted by SNP-gene expression correlation [17].

A multistep procedure for identifying key driver of a complex trait has been described by Schadt et al. [16]. Pairwise regressions among genotype variation, gene expression, and complex trait are investigated first. Then the likelihood based causal model selection (LCMS) test is used to identify expression profiles that sit between the complex-trait QTL and complex trait. In this approach, without applying the statistical test for causality, three different models (causal model, reactive model, and independent model) are used. The particular model with the lowest AIC (Akaike information criterion) value is considered to be the best fit for the data. One great advantage of this procedure is that when a correlation between an expression trait and a clinical phenotype does exist, it can distinguish causal, reactive, or independent relationship between them.

*1.1. Mediation Analysis.* Mediation analysis is the study of the causal chain or the indirect effect, to identify the possible underlying causal mechanisms. Mediation analysis is widely used across many disciplines such as social sciences, to identify the underlying causal mechanisms or to guide the experiments design [18]. A lot of research works focus on the relations between two variables, $X$ and $Y$. Much has been written about two-variable relations, including conditions under which $X$ can be considered a possible cause of $Y$. To this $X \rightarrow Y$ relation, one can add a third variable by using mediation, whereby $X$ causes the mediator, $M$, and $M$ causes $Y$, so $X \rightarrow M \rightarrow Y$ (see Figure 1). If $X$ leads to $Y$ through $M$, this is called the *indirect effect*. Ignoring $M$ leads to incorrect inference about the relation of $X$ and $Y$, since the effect of $M$ is confounded. If $M$ is related to $X$ and/or $Y$, so that information about $M$ improves the prediction of $Y$ by $X$ but does not substantially alter the relation of $X$ to $Y$ when $M$ is included in the analysis, then we consider $M$ as a covariate. In another situation, $M$ may also modify the relation of $X$ to $Y$ such that the relation of $X$ to $Y$ differs at different values of $M$. This is referred to as a moderator or interaction effect (see MacKinnon et al. [18] and references therein).

To establish this indirect relationship, Baron and Kenny [19] proposed a four-step approach in which several regression analyses are conducted, and the significance of the coefficients is examined at each step. In step 1, a simple regression analysis with $X$ predicting $Y$ is conducted (see Figure 1(a)) to test for path $\beta_1$ as

$$Y = \alpha_1 + \beta_1 X + \varepsilon_1. \quad (1)$$

In step 2, another simple regression analysis is performed with $X$ predicting $M$ to test for path $\beta_2$ as

$$M = \alpha_2 + \beta_2 X + \varepsilon_2. \quad (2)$$

And in step 3, the following regression equation is fitted with $M$ predicting $Y$ to test for path $\beta_3$:

$$Y = \alpha_3 + \beta_3 M + \varepsilon_3. \quad (3)$$

Step 2 and step 3 are combined in Figure 1(b). The final step is to conduct a multiple regression analysis with $X$ and $M$ predicting $Y$ as (see Figure 1(c))

$$Y = \alpha_4 + \beta_4 M + \beta_5 X + \varepsilon_4. \quad (4)$$

In all the above steps, it is assumed that independently, $\varepsilon_k \sim N(0, \sigma_k^2)$, $k = 1, 2, 3, 4$. The purpose of step 1–step 3 is to establish that zero-order relationships among the variables exist. One proceeds to step 4 assuming that there are significant relationships from steps 1 through 3. To

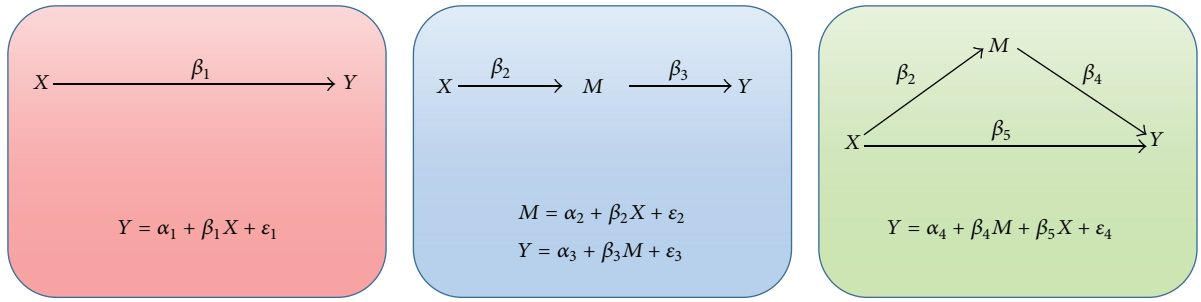



Figure 1: Mediation test.

identify potential causal links between genotype and clinical phenotypes, Huang et al. [20] designed a three-way model based on a stepwise regression analysis with genotype, gene expression, and cytotoxicity data as follows:

S1: SNP is associated with cytotoxicity,

S2: SNP is associated with gene expression,

S3: gene expression is associated with cytotoxicity.

Theoretical evidence in the form of "Causality Equivalence Theorem" has been proposed by Chen et al. [21] to establish causal relationship. According to the theorem, under the assumption that $X$ is randomized, the following conditions are needed to establish a causal relation:

C1: $X$ and $M$ are associated,

C2: $X$ and $Y$ are associated,

C3: $X$ is independent of $Y \mid M$.

If both $X$ and $M$ are significant predictors of $Y$, then *partial mediation* is achieved, whereas if $X$ is no longer significant when $M$ is controlled, this supports the condition of *full mediation*. However, there are some limitations of this test as mentioned by MacKinnon et al. [22]. This includes a low power to detect mediation and biased estimates. It does not test for the significance for the indirect pathway. An alternative and preferable approach to estimate the indirect effect is by multiplying two regression coefficients, $\beta_2 \times \beta_4$ [23].

In this paper, we introduce a new method, mediation analysis, which is somewhere between the overlap analysis (the first method) and causal inference (the second method). We use the human liver consortium data to demonstrate its application and performance. We use genome-wide genotype and gene expression data to explore functional mutation for an important pharmacogene, CYP2D6, which is a member of the cytochrome P450 mixed-function oxidase system and is responsible for the metabolism of 25% of all drugs on the market.

## 2. Material and Methods

*2.1. Human Liver Cohort Dataset.* Human liver cohort (HLC) data are collected from Sage Bionetworks Repository and Gene Expression Omnibus (GEO) database as described in the literature [17]. The dataset includes 2 genotype arrays (Illumina Sentrix human Hap650Y genotyping beadchip and Affymetrix 500 K genotyping array), gene expressions (30,128 probes × 466 samples) and enzyme activities (10 activity measurements of 9 enzymes × 488 samples), and demographic information. Genotype data for 219 Illumina and 214 Affymetrix that are publicly accessible are used. Patients with genotyping call rate less than 95% are removed from further analysis. This filtration reduces the sample sizes to 204 and 207 for Affymetrix and Illumina platforms, respectively. 167 Illumina genotyping has both gene expression data and enzyme activity data. In case of Affymetrix platform, 180 samples overlapped with gene expression and enzyme activity data.

SNPs whose genotyping call rate are less than 95% or Hardy-Weinberg equilibrium tests are significant ($P < 0.001$) or minor allele frequency <10% are discarded. For Affymetrix platform, 214,399 SNPs, and for Illumina, 471,394 SNPs are used for mediation and eQTL analysis. Enzyme activity and gene expression data are corrected with age and gender and then are normalized with normal quartile normalization.

*2.2. Mediation Analysis.* The mediation analysis method is developed to assess the indirect effects of genetic variant to CYP2D6 activity mediated by gene expressions. The tests are performed by parallel programming using C and MPICH. The computations are run on a Linux cluster computing environment with 200 compute nodes, and each node takes around 36 hours.

MacKinnon proposed a permutation test for mediation that makes use of the permutation-of-raw-data approach for testing a regression coefficient [22, 24]. It is referred to as the *permutation test of* $\beta_2 \times \beta_4$. To test for regression coefficients, permutation tests have been applied in several ways [24–26]. Applying this method requires, first, that the regression models in (2) and (4) are estimated for the original, nonpermuted data to find the values of $\beta_2$ and $\beta_4$. Values of the outcome variable, $Y$, are then permuted $10^9$ times and reassigned to nonpermuted scores on the predictor, $X$, and mediator, $M$, to create many permuted samples. The permuted $Y$ values, labeled $Y^+$, are then regressed on the nonpermuted $X$ and $M$ values in each permuted sample

Table 1: Mediation analysis.

| Genotype dataset | Enzyme | SNP effect | No. of sig. trios $P < 10^{-5}$ | No. of sig. SNPs $P < 10^{-5}$ | No. of sig. exp $P < 10^{-5}$ | FDR (trios) |
|---|---|---|---|---|---|---|
| Affymetrix | CYP2D6 | Gene dose | 389,573 | 103,369 | 3,545 | 16.63% |
| Illumina array | CYP2D6 | Gene dose | 1,214,416 | 251,738 | 4,770 | 11.73% |

Table 2: eQTL analysis.

| Genotype dataset | No. of pairs $P < 10^{-5}$ | No. of SNPs Correlated with >1 gene (total SNPs) | No. of SNPs Correlated with >20 genes |
|---|---|---|---|
| Affymetrix | 65,763 | 28,089 (214,399) | 295 |
| Illumina | 154,546 | 63,643 (471,394) | 724 |

(as in (4)), and the coefficient for $M$ in each permuted sample is labelled $\beta_4^*$. Similarly, values of the mediator, $M$, are permuted $10^9$ times and reassigned to values of the predictor $X$ to create many permuted samples. The permuted $M$ values, labeled $M^+$, are regressed on $X$ in each permuted sample (as in (2)), and the coefficient for $X$ in each permuted sample is labelled $\beta_2^+$. Finally, corresponding pairs of $\beta_2^+$ and $\beta_4^+$ values are multiplied to yield $\beta_2^+\beta_4^+$, and $\widehat{\beta}_2\widehat{\beta}_4$, the estimate of the mediated effect from the original data, is compared to the distribution of $\beta_2^+\beta_4^+$ to perform a test of the null hypothesis of no mediation.

The mediated effect is estimated by the product of coefficients ($\beta_2 \times \beta_4$) then divided by its standard error, which is derived by Sobel [23], under the assumption of multivariate normality for the standard error of the indirect effect, using the multivariate delta method as

$$\sigma_{\beta_2\beta_4} = \sqrt{\beta_2^2\sigma_4^2 + \beta_4^2\sigma_2^2}. \quad (5)$$

Hence, the test statistics are

$$\Delta_{\text{Indirrect effect}} = \frac{\widehat{\beta}_2\widehat{\beta}_4}{\text{se}\left(\widehat{\beta}_2\widehat{\beta}_4\right)}. \quad (6)$$

*2.3. Genome-Wide Association Based on Mediation Analysis.* The huge sizes of SNP and gene expression probes in mediation analysis introduce problems related to multiple hypotheses testing. False discovery rate (FDR) is used to control type I error for multiple testing. FDR is calculated as

$$\text{FDR} = \frac{\#\text{ significance by chance}}{\#\text{ significance results}}. \quad (7)$$

A stringent threshold is needed to avoid high FDR. Comparing to cis-acting variations, more transacting variations are detected by GWAS. In GWAS analysis, transeffects are usually weaker than cis-effects but are more numerous than the latter [14]. The trans-acting SNPs having smaller effects than cis-acting SNPs are more likely to be missed if more stringent threshold is applied.

*2.4. eQTL Analysis.* Transcript abundance is highly heritable in human populations and can be considered as a quantitative trait and be mapped to particular genomic loci, known as expression quantitative loci (eQTL). Not only gene expression is itself a complex trait, but also it acts as an intermediate phenotype between genetic loci and higher level cellular or clinical phenotypes, such as disease risk or individual drug response [27].

Linear model is fitted with genome-wide genotype and gene expression profiles. eQTL analysis is run in parallel on the same computing cluster with R language program. eQTL hotspots are defined as SNPs enriched in correlations with expression profiles across the genome (SNPs correlated with at least 20 gene expression profiles). The correlation $P$ values between SNP and expression probe less than $10^{-5}$ are considered to be significant and used for hotspot analysis. To test the enrichment of significant correlation between eQTL and all gene expression probes, exact binomial tests are conducted and corrected with Bonferroni method, and the corrected $P$ values are used as the enrichment scores.

## 3. Results

*3.1. Mediation Analysis.* The result of mediation analysis is summarized in Table 1. To find the significant trios, $P$ values less than $10^{-5}$ are considered. Using the same criteria for both platforms, the number of significant trios differs. For Affymetrix platform, we have 389,573 trios having $P$ values less than $10^{-5}$. For the other platform, this number is 1,214,416. The FDR for Illumina platform is found to be 11.73%, whereas for Affymetrix platform it is a bit higher (16.63%).

*3.2. eQTL Analysis.* In Table 2, the result corresponding to eQTL analysis of the HLC data is reported. The Affymetrix dataset has 214,399 SNPs after the implementation of the quality control out of which 28,089 are correlated with at least one gene at $P < 10^{-5}$ significance level, and there are total 65,763 SNP-gene pairs significantly correlated. 295 SNPs are correlated with at least 20 genes. Those 295 hotspots are used to check for overlapping with the results of mediation analysis. 289 eQTL hotspots are found correlated with 1542 gene expression profiles at $P < 10^{-5}$ significance level (Table 3). In contrast, Illumina dataset has higher quality

Table 3: QTL overlapping.

| Overlapping | Affymetrix | | Illumina | |
|---|---|---|---|---|
| | No. of eQTL hotspots | No. of mediation trios | No. of eQTL hotspots | No. of mediation trios |
| | 295 | 389,573 | 724 | 1,214,416 |
| No. of eQTL hotspot trios (No. of SNPs, No. of genes) | 9,296 (289, 1,542) | | 34,880 (719, 2,444) | |

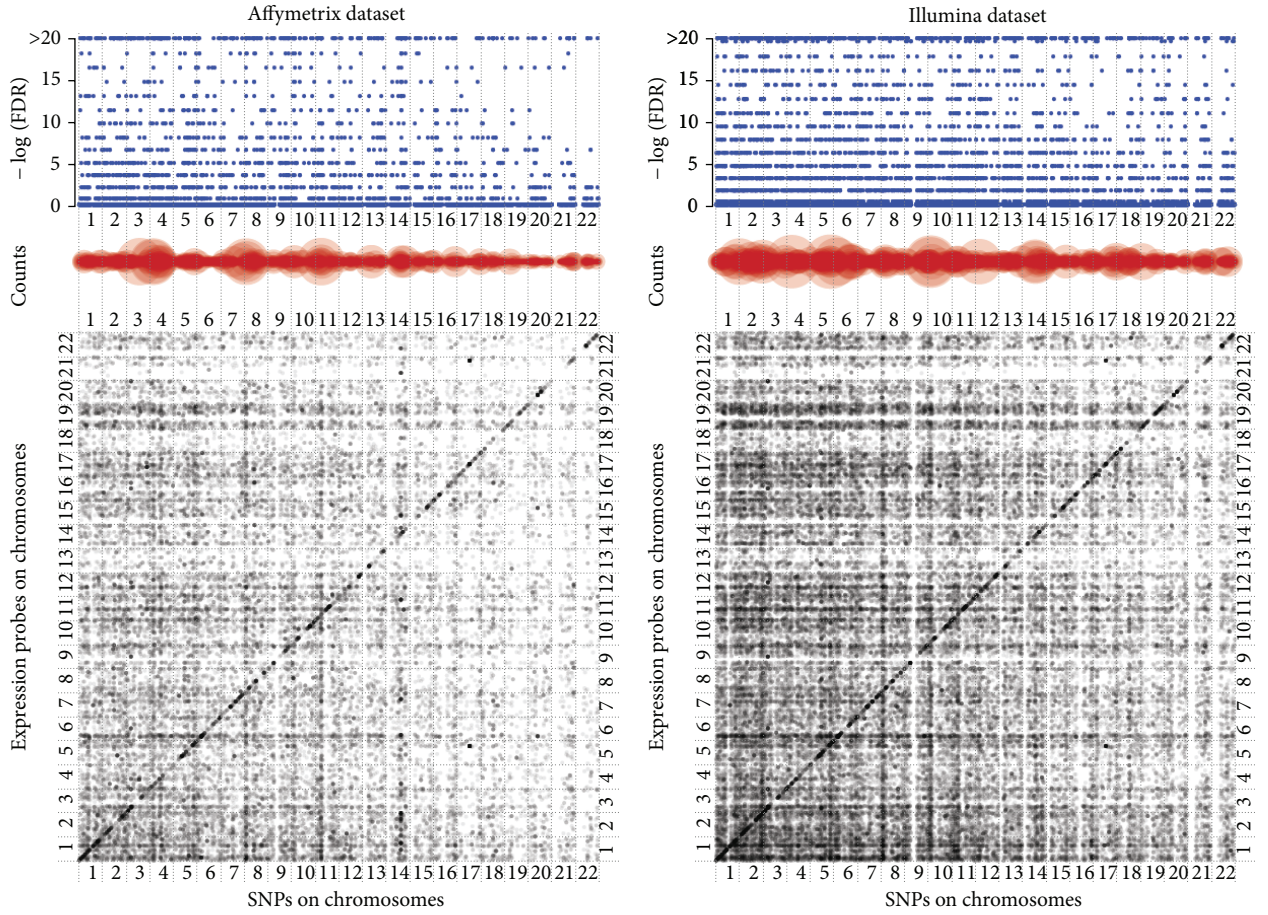



Figure 2: eQTL visualization. The main plot at the bottom is the scatter plot of the eQTL-transcript association. Each dot denotes a significant association between a SNP and a transcript ($P$ value $< 10^{-5}$). Gray color shows the level of significance where dark means more significant association. SNPs are arranged according to their chromosomal loci along the $X$-axis from chromosome 1 to 22, and genes are arranged along $Y$-axis in the same way. The dots along diagonal line indicate cis-eQTLs, otherwise, trans-eQTLs. The counts plot in the middle gives the number of genes that a SNP correlated with significantly ($P$ value $< 10^{-5}$). Large size means more genes associated with that SNP. The $-\log_{10}$ (FDR) plot at the top presents the enrichment score of a SNP associated with multiple transcripts comparing with that by chance. SNP has a large circle in counts plot and a high enrichment score in $-\log_{10}$ (FDR) plot which indicates eQTL hotspots.

with more SNPs passed quality control tests. Out of 471,394 SNPs, 63,643 SNPs are found to be correlated with at least one gene at $P < 10^{-5}$ significance level. Numbers of SNPs that are correlated with at least 20 genes are found to be 724, and 719 of the hotspots are significantly correlated with 2,444 genes in mediation analysis (Table 3). In Figure 2, a pictorial depiction of this eQTL analysis is given for both platforms. The significant SNP-expression pairs ($P < 10^{-5}$) are plotted as a dot according to the locations of the SNP and the gene on 22 chromosomes along $X$-axis and $Y$-axis. The grey colors show the level of significance, with darker dots representing smaller $P$ values. The counts of significant SNP-expression pairs and $-\log_{10}$ (FDR) for a given SNP are also plotted above the eQTL image. For each SNP, the count gives the number of genes that are correlated with this particular SNP, as the larger radius of the circle indicates that the SNP is correlated with more genes. In that case, it may be considered to be a potential eQTL hotspot. The dots along diagonal line indicate

Table 4: Functional annotations of the mediators.

| Types | Affymetrix no. Mediator genes | Illumina no. Mediator genes |
|---|---|---|
| Cytokine | 5 | 7 |
| Enzyme | 246 | 368 |
| G-protein coupled receptor | 17 | 20 |
| Growth factor | 5 | 11 |
| Ion channel | 13 | 18 |
| Kinase | 52 | 62 |
| Ligand-dependent nuclear receptor | 5 | 7 |
| Other | 373 | 596 |
| Peptidase | 31 | 39 |
| Phosphatase | 15 | 21 |
| Transcription regulator | 82 | 118 |
| Translation regulator | 6 | 10 |
| Transmembrane receptor | 12 | 16 |
| Transporter | 77 | 127 |
| Sum | 939 | 1420 |

cis-effects. It can be seen that cis-eQTLs have bigger effect on expression profile compared to trans-eQTLs.

*3.3. Functional Analysis of Hotspots Mediators.* 1,542 and 2,444 hotspot mediators from Affymetrix and Illumina platforms annotated to 1,388 and 2,187 unique genes separately. 939 and 1420 genes are successfully mapped in Ingenuity database for two platforms. The functional annotations of these genes are summarized in Table 4. Five (CCL16, CCL20, CMTM5, IL6, and SPP1) and 7 (CCL16, CCL20, CKLF, CKLFSF5, EPO, FAM3C, and SPP1) cytokines, 5 (AR, NR1I2, NR1I3, NR2F6, and PPARA) and 7 (AR, ESR1, NR1I2, NR1I3, PPARA, RORA, and RORC) ligand-dependent nuclear receptors, and 80 and 113 transcription regulators are found to mediate the relationship between genetic variant and CYP2D6 activity for Affymetrix and Illumina platforms. 64 transcription regulators overlapped between the two platforms (Gene List 1). Among the 64 transcription factors predicted mediateding genetic regulation of CYP2D6 activity, YY1 is reported putatively binding to gene CYP2D6 promoter region and regulating the expression of CYP2D6 and CYP2D4 [28, 29].

## 4. Conclusion

Cytochrome P450 constitutes a large subfamily of enzymes that plan an important role in the metabolism of endogenous compounds and the activation of chemical carcinogens. In this work, the regulations of P450 expression and activities have been intensely studied. Several other studies have found that P450 are subject to regulation by liver-enriched transcription factors, cytokines, and nuclear receptors. Our study provides some new clues on the regulation of CYP2D6 enzyme activity. Our mediation analysis is a powerful approach in identifying the trans-SNP-phenotype associations. We found a rich class of functional categories of mediators that potentially control the CYP2D6 activities, which include many new transcription factors. This method has some limitations too. In this work, the relationship between genetic variants, gene expression, and phenotype is assumed to be a simple one. However, in most of the situations, this relationship may become very complex. More sophisticated methods are required to analyze those complex models. In mediation analysis, we are only interested in testing the product of two regression coefficients. Mediation analysis cannot provide causal inference. The mediation analysis assumes that there is some causal relationship. It will be necessary to test for the assumption. We need to be extra cautious about drawing the conclusion of the causal relationship. Our studies provide insights into the comprehension of the complex regulatory network of CYP2D6 and improve our understanding of the functional genetic variations for the liver metabolism mechanisms.

## 5. Genes List

64 TFs overlapped between Affymetrix and Illumina datasets, including AATF, ALYREF, ARHGAP35, ASB8, ATF4, CBX4, CEBPG, CSDA, DDIT3, E2F5, ETV7, FOXN3, FOXN3, FUBP1, GPS2, HDAC10, HMGN1, ID1, INVS, IRF9, KANK1, KAT2B, KHDRBS1, KLF12, MAF, MAML2, MEIS2, MLXIPL, MXD4, MYBBP1A, MYCL1, NCOA7, NCOR1, NFIA, NFKB2, NFYA, NOLC1, NPM1, PEX14, PYCARD, SAP18, SATB1, SIM2, SLC2A4RG, SMARCC1, SNAI3, SNW1, SOX5, TCERG1, TCF7L2, TEAD3, TEAD4, TFDP2, TFEB, TOB1, TP53, YWHAB, YY1, ZGPAT, ZHX3, ZKSCAN1, ZNF132, ZNF256, and ZNF263.

## Abbreviations

GWAS: Genome-wide association study
eQTL: Expression quantitative trait loci
aQTL: Enzyme activity quantitative trait loci
FDR: False discovery rate
HLC: Human liver cohort
TFs: Transcription factors.

## Conflict of Interests

The authors have declared that no competing interests exist.

## Acknowledgments

This work is supported by the US National Institutes of Health Grant R01 GM74217 (Lang Li), R01 GM088076 (Todd Skaar), and AHRQ Grant R01HS019818-01 (Malaz Boustani). The content is solely the responsibility of the authors and does not necessarily represent the official views of the Agency for Healthcare Research and Quality.

## References


[1] J. H. Moore, F. W. Asselbergs, and S. M. Williams, "Bioinformatics challenges for genome-wide association studies," *Bioinformatics*, vol. 26, no. 4, Article ID btp713, pp. 445–455, 2010.

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

[3] S. A. McCarroll, A. Huett, P. Kuballa et al., "Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease," *Nature Genetics*, vol. 40, no. 9, pp. 1107–1112, 2008.

[4] A. L. Dixon, L. Liang, M. F. Moffatt et al., "A genome-wide association study of global gene expression," *Nature Genetics*, vol. 39, no. 10, pp. 1202–1207, 2007.

[5] H. H. H. Göring, J. E. Curran, M. P. Johnson et al., "Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes," *Nature Genetics*, vol. 39, no. 10, pp. 1208–1216, 2007.

[6] S. A. Monks, A. Leonardson, H. Zhu et al., "Genetic inheritance of gene expression in human cell lines," *American Journal of Human Genetics*, vol. 75, no. 6, pp. 1094–1105, 2004.

[7] E. Petretto, J. Mangion, N. J. Dickens et al., "Heritability and tissue specificity of expression quantitative trait loci," *PLoS Genetics*, vol. 2, no. 10, Article ID e172, 2006.

[8] P. M. Visscher, W. G. Hill, and N. R. Wray, "Heritability in the genomics era–concepts and misconceptions," *Nature Reviews Genetics*, vol. 9, no. 4, pp. 255–266, 2008.

[9] R. B. Brem, G. Yvert, R. Clinton, and L. Kruglyak, "Genetic dissection of transcriptional regulation in budding yeast," *Science*, vol. 296, no. 5568, pp. 752–755, 2002.

[10] M. V. Rockman and L. Kruglyak, "Genetics of global gene expression," *Nature Reviews Genetics*, vol. 7, no. 11, pp. 862–872, 2006.

[11] E. E. Schadt, S. A. Monks, T. A. Drake et al., "Genetics of gene expression surveyed in maize, mouse and man," *Nature*, vol. 422, no. 6929, pp. 297–302, 2003.

[12] A. J. Myers, J. R. Gibbs, J. A. Webster et al., "A survey of genetic human cortical gene expression," *Nature Genetics*, vol. 39, no. 12, pp. 1494–1499, 2007.

[13] R. S. Huang, S. Duan, W. K. Bleibel et al., "A genome-wide approach to identify genetic variants that contribute to etoposide-induced cytotoxicity," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 23, pp. 9758–9763, 2007.

[14] W. Cookson, L. Liang, G. Abecasis, M. Moffatt, and M. Lathrop, "Mapping complex disease traits with global gene expression," *Nature Reviews Genetics*, vol. 10, no. 3, pp. 184–194, 2009.

[15] N. Niu, Y. Qin, B. L. Fridley et al., "Radiation pharmacogenomics: a genome-wide association approach to identify radiation response biomarkers using human lymphoblastoid cell lines," *Genome Research*, vol. 20, no. 11, pp. 1482–1492, 2010.

[16] E. E. Schadt, J. Lamb, X. Yang et al., "An integrative genomics approach to infer causal associations between gene expression and disease," *Nature Genetics*, vol. 37, no. 7, pp. 710–717, 2005.

[17] X. Yang, B. Zhang, C. Molony et al., "Systematic genetic and genomic analysis of cytochrome P450 enzyme activities in human liver," *Genome Research*, vol. 20, no. 8, pp. 1020–1036, 2010.

[18] D. P. MacKinnon, C. M. Lockwood, C. H. Brown, W. Wang, and J. M. Hoffman, "The intermediate endpoint effect in logistic and probit regression," *Clinical Trials*, vol. 4, no. 5, pp. 499–513, 2007.

[19] R. M. Baron and D. A. Kenny, "The moderator-mediator variable distinction in social psychological research. Conceptual, strategic, and statistical considerations," *Journal of Personality and Social Psychology*, vol. 51, no. 6, pp. 1173–1182, 1986.

[20] B. Huang, S. Sivaganesan, P. Succop, and E. Goodman, "Statistical assessment of mediational effects for logistic mediational models," *Statistics in Medicine*, vol. 23, no. 17, pp. 2713–2728, 2004.

[21] L. S. Chen, F. Emmert-Streib, and J. D. Storey, "Harnessing naturally randomized transcription to infer regulatory relationships among genes," *Genome Biology*, vol. 8, no. 10, Article ID R219, 2007.

[22] D. P. MacKinnon, A. J. Fairchild, and M. S. Fritz, "Mediation analysis," *Annual Review of Psychology*, vol. 58, pp. 593–614, 2007.

[23] M. E. Sobel, "Asymptotic confidence intervals for indirect effects in structural equation models," *Sociological Methodology*, vol. 13, pp. 290–312, 1982.

[24] B. F. J. Manly, *Randomization, Bootstrap and Monte Carlo Methods in Biology*, vol. 70, Chapman & Hall, New York, NY, USA, 1997.

[25] M. J. Anderson and P. Legendre, "An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model," *Journal of Statistical Computation and Simulation*, vol. 62, no. 3, pp. 271–303, 1999.

[26] C. J. F. Terbraak, "Permutation versus bootstrap significance tests in multiple-regression and Anova," in *Bootstrapping and Related Techniques*, vol. 376 of *Lecture Notes in Economics and Mathematical Systems*, pp. 79–85, 1992.

[27] E. R. Gamazon, W. Zhang, A. Konkashbaev et al., "SCAN: SNP and copy number annotation," *Bioinformatics*, vol. 26, no. 2, pp. 259–262, 2010.

[28] X. L. Gong, Y. Liu, X. Zhang et al., "Systematic functional study of cytochrome P450 2D6 promoter polymorphisms in the Chinese Han population," *Plos One*, vol. 8, no. 2, Article ID e57764, 2013.

[29] D. Mizuno, Y. Takahashi, T. Hiroi, S. Imaoka, T. Kamataki, and Y. Funae, "A novel transcriptional element which regulates expression of the CYP2D4 gene by Oct-1 and YY-1 binding," *Biochimica et Biophysica Acta*, vol. 1627, no. 2-3, pp. 121–128, 2003.