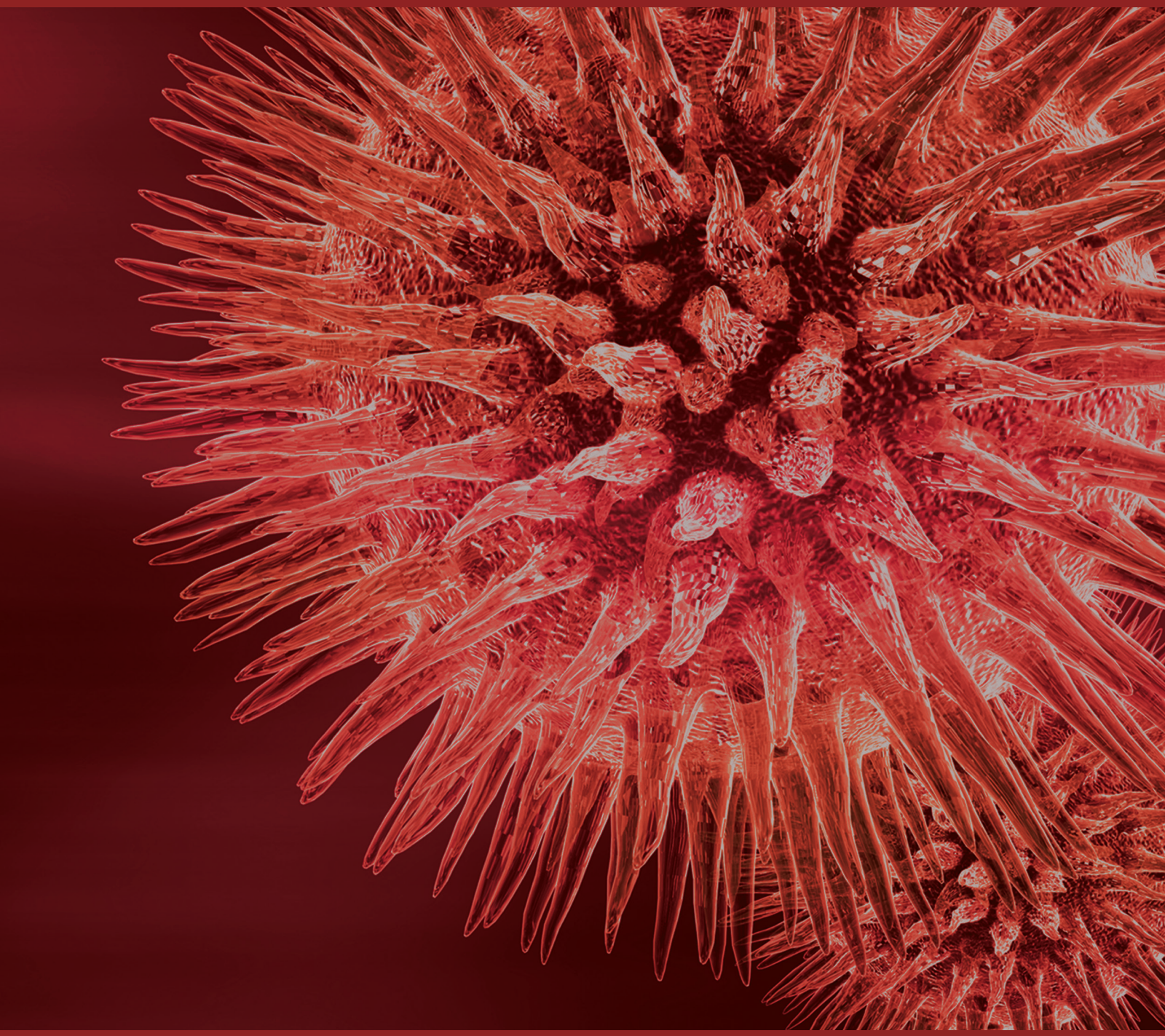
BioMed Research International

Advances in Computational Genomics

Guest Editors: Leng Han, Yan Guo, Zhixi Su, Siyuan Zheng, and Zhixiang Lu



Advances in Computational Genomics

BioMed Research International

Advances in Computational Genomics

Guest Editors: Leng Han, Yan Guo, Zhixi Su, Siyuan Zheng, and Zhixiang Lu

Contents

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 187803, 2 pages
http://dx.doi.org/10.1155/2015/187803

Editorial

Advances in Computational Genomics

Leng Han,[1] Yan Guo,[2] Zhixi Su,[3] Siyuan Zheng,[1] and Zhixiang Lu[4]

[1]*Department of Bioinformatics and Computational Biology, University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA*
[2]*Department of Cancer Biology, Vanderbilt University, Nashville, TN 37232, USA*
[3]*State Key Laboratory of Genetic Engineering and MOE Key Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University, Shanghai 200433, China*
[4]*Department of Microbiology, Immunology & Molecular Genetics, University of California, Los Angeles, CA 90095, USA*

Correspondence should be addressed to Leng Han; lhan1@mdanderson.org

Received 14 September 2014; Accepted 13 October 2014

High-throughput technologies, such as microarray and next generation sequencing (NGS), have emerged as a powerful tool less than a decade ago and have produced an avalanche of genome sequences. Computational genomics, which focus on computational analysis from genome sequences to other postgenomic data, including both DNA and RNA sequences, protein profiling, and epigenetic profiling, have become one of the most important avenues for biological discovery. Transforming genomic information into biomedical and biological knowledge requires creative and innovative new computational methods for all aspects of genomics. In this special issue, we selected fifteen high-quality papers in computational genomics field after in-depth peer review, and we briefly described these papers below.

Z. Su et al. performed an extensive analysis of the mouse knockout phenotype data and corroborated a strong effect of duplicate genes on mouse genetics robustness. The study suggested the potential correlation between the effect of genetic buffering and sequence conservation as well as protein-protein interactivity.

Y. Guo et al. developed a software package, O18Quant, which calculated the peptide/protein relative ratio and provided a friendly graphical user interface (GUI). The software greatly enhanced user's visualization and understanding in quantitative proteomics data analysis.

S. Zhao et al. developed an R package, "heatmap3," which significantly improved the original "heatmap" function by adding several more powerful and convenient features, including highly customizable legends and side annotation, a wider range of color selections, and new labeling features.

W. Chen et al. developed a newly developed software package, Genepleio, to estimate the effective gene pleiotropy from phylogenetic analysis of protein sequences. This work would facilitate the understanding of how gene pleiotropy affected the pattern of genotype-phenotype map and the consequence of organismal evolution.

S. Lee at al. developed the BLAST-like alignment tool (BLAT) based comparative analysis for transposable elements (BLATCAT) program and compared specific regions of representative primate genome sequences.

H. Zhang et al. developed a new method for tumor-gene selection, the chi-square test-based integrated gene rank and direct classifier. The informative genes selected significantly improved the independent test precision of other classifiers.

L. Zhong et al. analyzed miRNA-mRNA paired variations (MMPVs) comprehensively and demonstrated that the existence of MMPVs is general and widespread but that there is a general unbalance in the distribution of MMPVs among the different pathological features.

N. Jin et al. systematically evaluated frequently used methods using two types of integration strategies, empirical and machine, and provided an important basis for future network-based biological research learning methods.

H. Zhao et al. developed an integrated strategy to identify differential coexpression patterns of genes and probed the functional mechanisms of the modules. This approach was

able to robustly detect coexpression patterns in transcriptomes and to stratify patterns according to their relative differences.

A. V. Polonikov et al. comprehensively analyzed the associations between adult asthma and single nucleotide polymorphisms and found the epistatic interactions between ADE genes underlying asthma susceptibility and the genetic heterogeneity between allergic and nonallergic variants of the disease.

J. Shu et al. constructed a database to design vaccine that targeted the Chinese, and predicted 20 potential HIV epitopes. This work will facilitate the development of a CD4+ T cell vaccine especially catered for the Chinese.

Y. Li et al. built a link map between small molecules and pathways using gene expression profiles, pathways, and gene expression of cancer cell lines intervened by small molecules and provided a valuable reference for identifying drug candidates and targets in molecularly targeted therapy.

Y.-D. Gao et al. investigated the *E. coli* strains in the human gut microbiome by using deep sequencing data. The authors reconstructed genome-wide metabolic networks for the three most common *E. coli* strains and provided a systematic perspective on *E. coli* strains in the human gut microbiome.

T. Tan et al. compared small RNA signatures to address small RNA transition during mouse spermatogenesis and provided insight into the mechanisms involved in the regulation of spermatogonial stem cells activities.

H. Hu et al. performed RNA-seq to investigate the mice testicular transcriptome to elucidate the mechanism of male reproductive toxicity and found several transcriptional signatures closely related to the biological processes of regulation of hormone, gamete generation, and sexual reproduction.

By launching this issue, we wish to stimulate the continuing efforts in computational genomics for more efficient analysis of big genomics data.

Acknowledgments

We would like to acknowledge the anonymous reviewers for their critical comments that significantly improved the quality of the papers in this special issue. Z. Su was supported by the Shanghai Pujiang Program (13PJD005). Y. Guo was supported by CCSG (P30CA068485).

Leng Han
Yan Guo
Zhixi Su
Siyuan Zheng
Zhixiang Lu

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 269150, 6 pages
http://dx.doi.org/10.1155/2015/269150

Research Article

Genepleio Software for Effective Estimation of Gene Pleiotropy from Protein Sequences

Wenhai Chen,[1,2] Dandan Chen,[2] Ming Zhao,[3] Yangyun Zou,[2] Yanwu Zeng,[2,4] and Xun Gu[2,5]

[1] *College of Mathematics & Information Science, Wenzhou University, Wenzhou, China*
[2] *State Key Laboratory of Genetic Engineering and MOE Key Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University, Shanghai 200433, China*
[3] *College of Life & Environmental Science, Wenzhou University, Wenzhou 325035, China*
[4] *Shanghai Stem Cell Institute, Institutes of Medical Sciences, School of Medicine, Shanghai Jiao Tong University, Shanghai 200240, China*
[5] *Department of Genetics, Developmental and Cell Biology, Iowa State University, Ames, IA 50011, USA*

Correspondence should be addressed to Yanwu Zeng; zengyanwu@gmail.com and Xun Gu; xungufudan@gmail.com

Received 3 June 2014; Revised 15 July 2014; Accepted 26 July 2014

Academic Editor: Siyuan Zheng

Though pleiotropy, which refers to the phenomenon of a gene affecting multiple traits, has long played a central role in genetics, development, and evolution, estimation of the number of pleiotropy components remains a hard mission to accomplish. In this paper, we report a newly developed software package, *Genepleio,* to estimate the effective gene pleiotropy from phylogenetic analysis of protein sequences. Since this estimate can be interpreted as the minimum pleiotropy of a gene, it is used to play a role of reference for many empirical pleiotropy measures. This work would facilitate our understanding of how gene pleiotropy affects the pattern of genotype-phenotype map and the consequence of organismal evolution.

1. Introduction

Understanding the role of gene pleiotropy in the map from genotypes to phenotypes has been one of the central topics for biologists, which refers to the phenomenon of a gene affecting multiple traits As a major measure for the functional importance of a gene, this concept has played a fundamental role in genetics, development, and evolution (see [1–3] for recent reviews and comments). However, the degree of gene pleiotropy remains largely unknown. Historically, proposed the concept of universal pleiotropy; that is, a single mutation can potentially affect all phenotypic traits. Though Fisher's model has been widely used as a theoretical basis for exploring the evolutionary interplay between the genotype and phenotype, the notion of universal pleiotropy has not been well tested.

Indeed, compared with the wide availability of genomics data, the whole-range phenotype recourses, or "phenomics," are highly limited. Nevertheless, recent advances have been able to bring high throughput data to bear on the nature and extent of pleiotropy [4–6]. These experiments showed that the number of phenotypic traits that may be affected by a gene may be actually limited implying the role of modularity in shaping the degree of gene pleiotropy. Many controversial issues such as the problem of arbitrary number of correlated traits may directly affect the count of phenotypes that are predicted to be affected by a gene. On the other hand, a new approach has emerged in the past decade, to estimate the gene pleiotropy from genetics or sequence data, rather than from the affected phenotypes [7–13] (Chen et al., 2013). In particular, Gu [8] developed a practically feasible approach. It proposed that molecular evolution of a gene occurs in a multidimensional space corresponding to the same canonical number of molecular phenotypes. Random mutations of the gene could affect these molecular phenotypes constrained by the stabilizing selection. Moreover, Gu [8] developed a statistical method to estimate the "effective pleiotropy" (K_e) of a gene from the multiple sequence alignment of protein sequences. Most genes have K_e in the range between 1 and 20 [11], with the medium of $K_e = 6.5$ of these estimates that is

comparable to some empirical pleiotropy measures [1, 3]. Yet the relationship between these two approaches remains complex. As the degree of gene pleiotropy is a basic parameter for understanding the complexity of genotype-phenotype map, to facilitate this line of research we develop a software package *Genepleio* (freely available at http://www.xungulab.com) to estimate the effective gene pleiotropy from the protein sequences.

2. Material and Methods

2.1. Sequences. Three groups of datasets include eight vertebrates, twelve fruit flies, and seven yeast species. Each dataset contains 300 random selected orthologous sets. Eight vertebrates are fugu, zebrafish, xenopus, chicken, dog, cow, mouse, and human. Twelve *Drosophila* species are *D. melanogaster, D. pseudoobscura, D. sechellia*, *D. simulans*, *D. yakuba*, *D. erecta*, *D. ananassae*, *D. persimilis*, *D. willistoni*, *D. mojavensis, D. virilis,* and *D. grimshawi.* Seven yeast species are *Candida glabrata, Debaryomyces hansenii, Kluyveromyces lactis, Saccharomyces bayanus, Yarrowia lipolytica, Saccharomyces cerevisiae,* and *Schizosaccharomyces pombe.* Vertebrate CDS and protein sequences were extracted from Ensmart, Drosophila CDS and protein sequences were extracted from FlyBase, and yeast CDS was extracted from ORNA Man's dataset (corresponding protein sequences were translated by Bioperl).

2.2. Sequences Alignment. Multiple protein sequence alignment for each orthologous group was obtained by ClustalW at default settings.

2.3. Estimation of d_N/d_S. The number of synonymous substitutions per synonymous site (d_S) and the number of nonsynonymous substitutions per nonsynonymous site (d_N) between human and mouse orthologs were calculated by CODEML of PAML package [14]. When calculating the variance of d_N and d_S, we changed "getSE = 1" in CODEML control file; otherwise, we used the default CODEML parameters. We used the estimates of d_N/d_S between human and mouse for vertebrates, those between *D. melanogaster (dmel) and D. sechellia (dsec)* for *Drosophila,* and those between *S. bayanus and S. cerevisiae* for yeasts, respectively.

3. Results and Discussion

3.1. Estimation of Effective Gene Pleiotropy. Gu [8] analyzed the pleiotropy model of molecular evolution under the following assumptions. (i) K-dimensional molecular phenotypes ($\mathbf{y}$) of the gene are under Gaussian-like stabilizing selection, indicating a single fitness optima for multiple functions. Any deviation from the optima is under the purifying selection. (ii) The fitness optima of $\mathbf{y}$ may shift randomly during the course of evolution, according to a multivariate normal distribution. It generates the process of *microadaptation* that could be caused by the external (environmental) or internal (physiological) perturbations or the functional compensation for the previously fixed slightly deleterious mutation. (iii) And the distribution of mutational effects, $p(\mathbf{y})$, follows a multivariate normal distribution.

The estimation procedure implemented in the software *Genepleio* is summarized in Figure 1. We address several key issues concisely to help in understanding how the software works. One may see Gu [8], Su et al. [11], and Gu (2014) for technical details.

3.1.1. Calculation of H-Measure. Calculation of H is the key step to estimate K_e. Biologically, H measures the strength of rate variation among sites: $H = 0$ when $\text{var}(\lambda) = 0$, and $H = 1$ when $\text{var}(\lambda) = \infty$. After the gene phylogeny is given or inferred by the NJ option, the software implemented the methods of Gu and Zhang [15] to infer the number of amino acid changes along the phylogeny at each site, after correcting the multiple hits. The next step is to calculate the mean (M) and variance (V) of the estimated number of changes over sites. Under the Poisson-based evolutionary model, H can be estimated by $H = (V - M)/[V + M(M - 1)]$.

3.1.2. Estimation of Effective Gene Pleiotropy (K_e). Genepleio has implemented the following procedure to estimate K. (i) Calculate the d_N/d_S ratio (the ratio of nonsynonymous to synonymous rates) used as an empirical measure for the mean sequence conservation. (ii) Calculate the g-function $g = d_N/d_S/(1-H)$. (iii) And the effective gene pleiotropy (K_e) can be estimated by numerically solving the following equation:

$$\frac{d_N/d_S}{1-H} = 2^{-K_e/2}\left[1 + \phi\left(K_e\right)\right], \tag{1}$$

where $\phi(K_e) = 0.0208K_e(K_e + 2)/(1 + 0.289K_e)$.

3.1.3. Estimation of Selection Intensities. There are two types of selection intensity measures. The first one is the (overall) selection intensity of the gene under study, S, for the overall strength of purifying selection imposed on the protein sequence; the negative sign indicates the negative (purifying) selection. The second one is the baseline selection intensity, B_0, which is a scaled measure for the contribution of a single pleiotropy component. The relationship between B_0 and S is $S = -K \times B_0$. When K_e is obtained, the software estimates the effective selection intensity S_e for S and the effective baseline selection intensity (B_e) for the baseline selection intensity B_0.

3.1.4. Calculation of Sampling Variance of K_e. The sampling variance of K_e can be approximately calculated by the delta method. Numerical analysis of (1) found that the following formula is accurate enough in practice:

$$\begin{aligned}\text{Var}\left(K_e\right) &\approx 11.037\left[\frac{\text{Var}\left(d_N/d_S\right)}{\left(d_N/d_S\right)^2} + \frac{\text{Var}\left(H\right)}{(1-H)^2}\right] \\ &\approx 11.037\left[\frac{\text{Var}\left(d_N\right)}{d_N^2} + \frac{\text{Var}\left(d_S\right)}{d_S^2} + \frac{\text{Var}\left(H\right)}{(1-H)^2}\right].\end{aligned} \tag{2}$$

In (2), $\text{Var}(d_N/d_S)$ can be estimated by the delta method so that $\text{Var}(d_N/d_S) \approx \text{Var}(d_N)/(d_N)^2 + (d_N)^2 *$

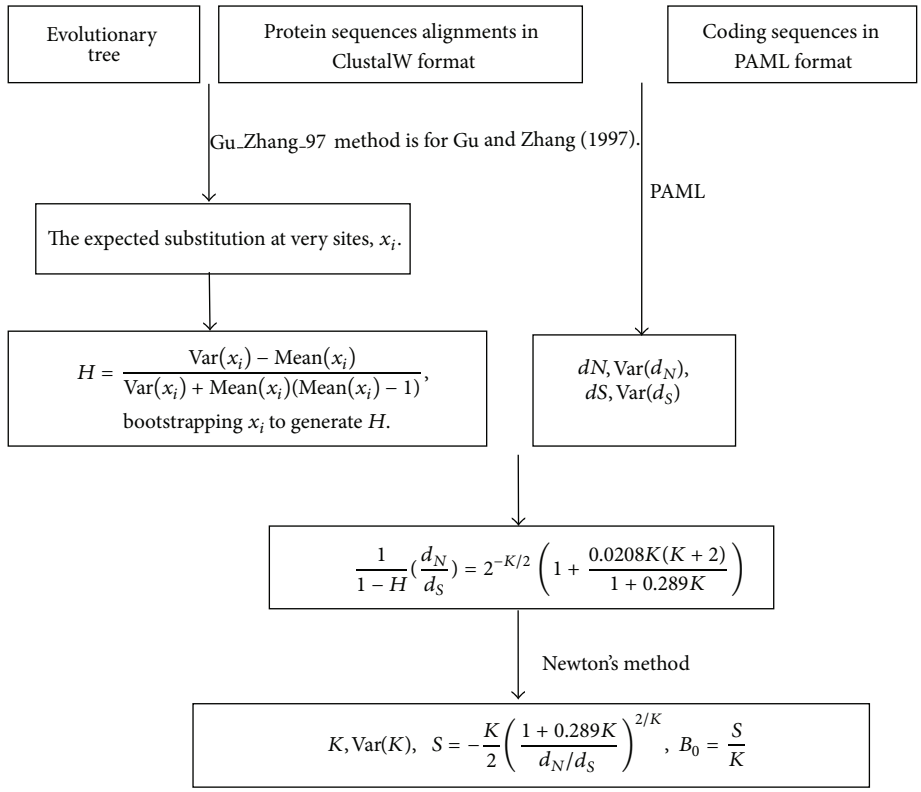


Figure 1: A flow chart to outline the computational pipeline implemented in software *Genepleio*.

$\mathrm{Var}(d_S)/(d_S)^4$, where $\mathrm{Var}(d_N)$, $\mathrm{Var}(d_S)$ are the variances of d_N and d_S, respectively. The sampling variance H is difficult to compute analytically. *Genepleio* has implemented a bootstrapping approach to calculate the sampling variance of H, $\mathrm{Var}(H)$, whereas sampling variances of d_N and d_S, $\mathrm{Var}(d_N)$ and $\mathrm{Var}(d_S)$, depend on the users' input; their default values are set to be zero. Using this method, we can bootstrap 100 times within 1~2 minutes.

We use triosephosphate isomerase gene (TPI1, SWISSPROT P60174) for illustration. (i) Infer the phylogenetic tree from the multiple alignment of vertebrate homologous TPI1 protein sequences (human, mouse, dog, cow, chicken, xenopus, fugu, and zebrafish), which is consistent with the known vertebrate phylogeny. (ii) Estimate $d_N/d_S = 0.045$ between the human and mouse genes by the likelihood method using PAML. (iii) Estimate H-index for the rate variation among sites; $H = 0.614$ for TPI1 gene. (iv) And we estimated $K_e = 7.29$ and the mean selection intensity $S = -11.65$. Then, the baseline selection intensity is given by $B_0 = 11.65/7.29 \approx 1.60$.

The estimated effective gene pleiotropy varies among different treatments but the scale of variation is small. On the other hand, we found that when the number of changes at each site is estimated by the parsimony method without any correction, gene pleiotropy tends to be overestimated. At any rate, we conclude that these 5–10% estimation differences should not affect the general pattern about the degree of gene pleiotropy.

3.2. Biological Interpretation of K_e. The key question is how one can acquire the number of pleiotropy components of a gene without biologically knowing each component (Su et al., 2010) [12, 16]. Gu (2014) [17] addressed this issue, showing that the method of Gu [8] actually aims to estimate the rank (K) of genotype-phenotype map. The main result can be concisely represented by the following simple formula: $K = \min(r, P_{\min})$, where $P_{\min}$ is the minimum pleiotropy among all legitimate pleiotropy measures and r is the rank of mutational effects. In short, the meaning of "*effective gene pleiotropy*" (K_e) estimated by Gu-2007 method is as follows. (i) K_e is an estimate of $K = \min(r, P_{\min})$, the rank of genotype-phenotype map. (ii) K_e is an estimate for the minimum pleiotropy $P_{\min}$ only if $P_{\min} < r$. (iii) Gu-2007 method attempted to estimate the pleiotropy of amino acid sites, a conserved proxy to the true minimum pleiotropy. (iv) With a sufficiently large phylogeny such that the rank of mutational effect at an amino acid site is $r \to 19$ (the number of amino acid types minus one), one can estimate $P_{\min}$ in the range from 1 to 19 by this method. And (v) K_e is a conserved estimate of K because those pleiotropy components that have small effects on fitness would be effectively removed by the estimation procedure.

3.3. Software Overview. We have developed *Genepleio*, a GUI-based software package that estimates the effective gene pleiotropy from the phylogenetic sequence analysis of amino acids. *Genepleio* has three inputs. (i) Input the file of multisequence alignment (MSA) of protein sequences:

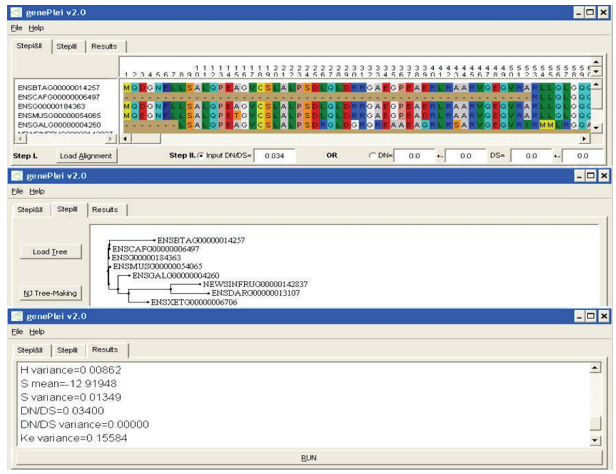


Figure 2: Screen illustration of the software *Genepleio*.

Table 1: Simulation results of K_e-estimation by *Genepleio*.

K	B_0	K_e (model (a))	K_e (model (b))	K_e (model (c))
2	0.5	0.98 ± 0.03	0.94 ± 0.02	0.95 ± 0.02
4	0.5	1.98 ± 0.03	1.96 ± 0.03	1.58 ± 0.03
8	0.5	4.05 ± 0.05	3.67 ± 0.05	2.49 ± 0.03
12	0.5	6.16 ± 0.06	4.76 ± 0.06	3.89 ± 0.05
16	0.5	8.29 ± 0.13	6.65 ± 0.10	4.36 ± 0.06
2	1.0	1.34 ± 0.04	1.31 ± 0.04	1.31 ± 0.04
4	1.0	2.71 ± 0.06	2.13 ± 0.05	2.13 ± 0.05
8	1.0	5.44 ± 0.13	4.99 ± 0.10	3.26 ± 0.07
12	1.0	8.17 ± 0.29	6.56 ± 0.16	5.22 ± 0.10
16	1.0	10.9 ± 0.84	9.10 ± 0.37	5.85 ± 0.17
2	2.0	1.64 ± 0.06	1.60 ± 0.03	1.61 ± 0.06
4	2.0	3.28 ± 0.12	3.25 ± 0.05	2.61 ± 0.08
8	2.0	6.50 ± 0.40	6.12 ± 0.07	4.04 ± 0.15
12	2.0	9.67 ± 0.65	8.27 ± 0.07	6.59 ± 0.32
16	2.0	13.02 ± 0.67	11.30 ± 0.22	7.45 ± 0.46

Genepleio supports the alignment format of CLUSTALW. As required by the method, the multiple alignment file should contain at least four sequences with reasonably large sequence divergences between them. (ii) Input two values d_N (nonsynonymous distance) and d_S (synonymous distance). Several methods such as PAML can be used to obtain these estimates. We suggest choosing closely related sequences, say, $d_S < 1$, to avoid large sampling variance when calculating the d_N/d_S ratio. Note that the d_N/d_S ratio should be less than 1; otherwise, the gene may not be suitable to do this type of analysis. (iii) Input the tree file in the Phylip format. Alternatively, one may use the neighbor-joining (NJ) method implemented in the software to infer the gene phylogeny. As illustrated in Figure 2, the interface of *Genepleio* includes three main tab pages: the first page is for the MSA input and d_N/d_S values, the second page is for the input or the inference of the phylogenetic tree, and the third page will output the results of estimation.

We have conducted a preliminary analysis and found that the 75% quantile of estimated K_e is typically within $K_e \pm 2$, suggesting that K_e estimation as a measure of gene pleiotropy is statistically reliable. Besides, we notice that the contributions from $\mathrm{Var}(d_N)$ and $\mathrm{Var}(d_S)$ are nontrivial. In other words, the sampling variance of K_e would be severely underestimated if the user has no input for the sampling variances of d_N and d_S.

There are some notices about usage of *Genepleio*. First, the multiple alignment file should contain more than four sequences; second, the d_N/d_S value should be within $(0, 1)$; third, the sequences similarity >90% should be cautious because of the lack of statistical power; fourth, in order to shorten the time consumed, we do not give the mean of K_e value through bootstrapping in *Genepleio*. Nevertheless, according to much simulation, the mean value is close to the estimated K_e value, so we only give the estimated K_e value.

3.4. Computer Simulations. We have carried computer simulations to evaluate the software performance. We set K to vary from 1 to 20, with the fixed baseline selection intensity $B_0 =$ 0.5, 1.0, or 2.0, respectively. In particular, we consider three simulation models. (a) Independent-equal model: pleiotropy components are identical and independent of each other.

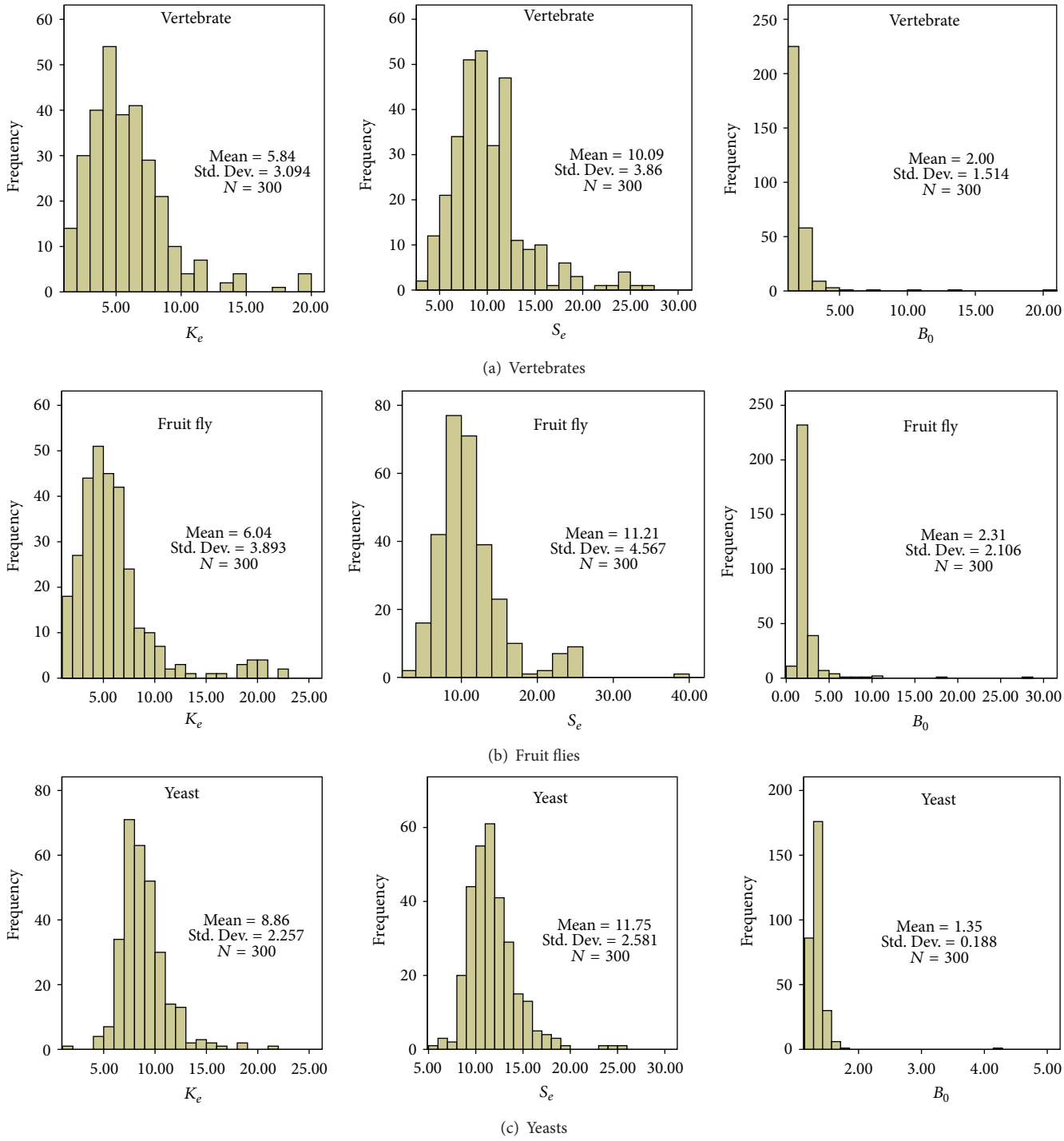


FIGURE 3: Estimation of K_e from three datasets: eight vertebrates, twelve fruit flies, or seven yeast species, respectively. Each dataset contains 300 random selected (one-to-one) orthologous sets.

(b) Independent-unequal model: pleiotropy components are independent of each other but have different strengths. (c) Random-matrix model: the strengths and correlations between pleiotropy components are randomly drawn from a specified random matrix model. Our main results are summarized in Table 1. In general, the estimation bias of K_e decreases with the increasing of the baseline selection intensity B_0. For instance, in model (a), underestimation of K_e is considerable only when B_0 is very small, say, 0.5 or less. The estimation bias becomes intermediate when $B_0 > 1$ and becomes negligible when $B_0 > 3$ (not shown). Moreover, the estimation bias of K_e may increase when the simulation model becomes more complex. Indeed, in model (c), K_e only describes the canonical number of pleiotropy components that could be much less than the number of pleiotropy components used in the simulation model.

3.5. Case Studies. To validate the performance of the newly developed software for the estimation of K_e, we analyzed three datasets, each of which includes eight vertebrates, twelve fruit flies, or seven yeast species, respectively (Figure 3). Each dataset contains 300 random selected (one-to-one) orthologous sets. We calculated K_e, S_e (selection intensity), and B_0 (the baseline selection intensity). Our analysis shows that all K_e estimates are in a reasonable range with only a few outliners. Interestingly, the distribution of K_e estimates is similar across these distantly related species. The underlying reason to explain this similarity remains unclear. K_e is an important parameter for evolutionary analysis. Indeed, the square of coefficient correlation (r^2) between K_e and S_e is 0.64 in vertebrates, 0.94 in yeasts, and 0.55 in fruit flies, suggesting that gene pleiotropy may be an important evolutionary constraint in molecular evolution. In short, in this paper, we have reported a new software package *Genepleio* and demonstrated the steps of gene pleiotropy (K) estimation. We also examined the extent to which the statistical properties of d_N/d_S and H affect the estimation efficiency of K and S. Comparison among three different groups of species validates the stability of K estimation procedure.

Conflict of Interests

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

Authors' Contribution

Wenhai Chen and Dandan Chen have equal contributions.

Acknowledgments

The authors are grateful to Yong Huang for the involvement of early work and to Zhixi Su and Wei Huang for critical comments on the early version of the paper.

References

[1] G. P. Wagner and J. Zhang, "The pleiotropic structure of the genotype-phenotype map: the evolvability of complex organisms," *Nature Reviews Genetics*, vol. 12, no. 3, pp. 204–213, 2011.

[2] W. G. Hill and X. S. Zhang, "On the pleiotropic structure of the genotype-phenotype map and the evolvability of complex organisms," *Genetics*, vol. 190, no. 3, pp. 1131–1137, 2012.

[3] A. B. Paaby and M. V. Rockman, "The many faces of pleiotropy," *Trends in Genetics*, vol. 29, no. 2, pp. 66–73, 2013.

[4] A. M. Dudley, D. M. Janse, A. Tanay, R. Shamir, and G. M. Church, "A global view of pleiotropy and phenotypically derived gene function in yeast," *Molecular Systems Biology*, vol. 1, no. 1, Article ID 2005.0001, 2005.

[5] Y. Ohya, J. Sese, M. Yukawa et al., "High-dimensional and large-scale phenotyping of yeast mutants," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 52, pp. 19015–19020, 2005.

[6] C. Pál, B. Papp, and M. J. Lercher, "An integrated view of protein evolution," *Nature Reviews Genetics*, vol. 7, no. 5, pp. 337–348, 2006.

[7] G. Martin and T. Lenormand, "A general multivariate extension of fisher's geometrical model and the distribution of mutation fitness effects across species," *Evolution*, vol. 60, no. 5, pp. 893–907, 2006.

[8] X. Gu, "Evolutionary framework for protein sequence evolution and gene pleiotropy," *Genetics*, vol. 175, no. 4, pp. 1813–1822, 2007.

[9] X. Gu, "Stabilizing selection of protein function and distribution of selection coefficient among sites," *Genetica*, vol. 130, no. 1, pp. 93–97, 2007.

[10] L. Chevin, G. Martin, and T. Lenormand, "Fisher's model and the genomics of adaptation: restricted pleiotropy, heterogenous mutation, and parallel evolution," *Evolution*, vol. 64, no. 11, pp. 3213–3231, 2010.

[11] Z. Su, Y. Zeng, and X. Gu, "A preliminary analysis of gene pleiotropy estimated from protein sequences," *Journal of Experimental Zoology Part B*, vol. 314, pp. 115–122, 2010.

[12] Y. Zeng and X. Gu, "Genome factor and gene pleiotropy hypotheses in protein evolution," *Biology Direct*, vol. 5, article 37, 2010.

[13] P. Razeto-Barry, J. Díaz, and R. A. Vásquez, "The nearly neutral and selection theories of molecular evolution under the fisher geometrical framework: substitution rate, population size, and complexity," *Genetics*, vol. 191, no. 2, pp. 523–534, 2012.

[14] Z. Yang, "PAML: a program package for phylogenetic analysis by maximum likelihood," *Computer Applications in the Biosciences*, vol. 13, no. 5, pp. 555–556, 1997.

[15] X. Gu and J. Zhang, "A simple method for estimating the parameter of substitution rate variation among sites," *Molecular Biology and Evolution*, vol. 14, no. 11, pp. 1106–1113, 1997.

[16] W. Chen, Z. X. Su, and X. Gu, "A note on gene pleiotropy estimation from phylogenetic analysis of protein sequences," *Journal of Systematics and Evolution*, vol. 51, pp. 365–369, 2012.

[17] X. Gu, "Pleiotropy can be effectively estimated without counting phenotypes through the rank of a genotype-phenotype map," *Genetics*, vol. 197, pp. 1357–1363, 2014.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 296349, 13 pages
http://dx.doi.org/10.1155/2014/296349

Research Article

Integration Strategy Is a Key Step in Network-Based Analysis and Dramatically Affects Network Topological Properties and Inferring Outcomes

Nana Jin,[1,2] Deng Wu,[2] Yonghui Gong,[2] Xiaoman Bi,[2] Hong Jiang,[1] Kongning Li,[2] and Qianghu Wang[1,2]

[1] *Bioinformatics Department of School of Basic Medical Sciences, Nanjing Medical University, Nanjing 210029, China*
[2] *College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, China*

Correspondence should be addressed to Kongning Li; kongningli@hotmail.com and Qianghu Wang; wangqh@njmu.edu.cn

Received 22 May 2014; Revised 14 July 2014; Accepted 17 July 2014; Published 27 August 2014

Academic Editor: Siyuan Zheng

An increasing number of experiments have been designed to detect intracellular and intercellular molecular interactions. Based on these molecular interactions (especially protein interactions), molecular networks have been built for using in several typical applications, such as the discovery of new disease genes and the identification of drug targets and molecular complexes. Because the data are incomplete and a considerable number of false-positive interactions exist, protein interactions from different sources are commonly integrated in network analyses to build a stable molecular network. Although various types of integration strategies are being applied in current studies, the topological properties of the networks from these different integration strategies, especially typical applications based on these network integration strategies, have not been rigorously evaluated. In this paper, systematic analyses were performed to evaluate 11 frequently used methods using two types of integration strategies: empirical and machine learning methods. The topological properties of the networks of these different integration strategies were found to significantly differ. Moreover, these networks were found to dramatically affect the outcomes of typical applications, such as disease gene predictions, drug target detections, and molecular complex identifications. The analysis presented in this paper could provide an important basis for future network-based biological researches.

1. Introduction

Molecular interactions, such as protein-DNA interactions [1], protein-RNA interactions [2], DNA-DNA interactions [3], RNA-RNA interactions [4], and protein-protein interactions [5], facilitate various organismal functions, including the process of transcription [6], multiple long-range interactions between promoters and distal elements [7], and the regulation of gene expression [8]. Therefore, many experiments have been designed to detect intercellular and intracellular molecular interactions [9, 10]. Of these molecular interactions, protein-protein interactions have especially been found to play a crucial role in defining most of the molecular functions [11].

Consequently, molecular networks based on these interactions have been built to elucidate their underlying roles in biology [12]. Interactions have been utilised to build many protein-protein interaction network databases, such as DIP [13], HPRD [14], BIND [15], BioGRID [16], IntAct [17], and MINT [18], yielding more than 150,000 binary interactions. Researchers have used the interaction network of these databases to perform many studies and applications [11, 19, 20]. The interactions reported in these databases are derived from sources including yeast two-hybrid, anti-tag coimmunoprecipitation, mass-spectrometric, and literature mining experiments.

Traditional protein-protein interactions have been detected in a high-quality manner based on top-down- and hypothesis-based methods supported by experimental data [11]. Further, recent protein-protein interaction data have been generated in large numbers based on high-throughput methods, thus reconfiguring the biological network from

a different point of view. However, two major shortages cannot be ignored. The coverage of current protein-protein interaction networks is less than 50%, and the accuracy of these data ranges from 10% to 50% [21–23]. Consider an example in which, by utilising various methodologies, researchers identify 80,000 protein interaction pairs in yeast; however, they only confirm interactions for approximately 2,400 pairs using more than two methods [22]. Several reasons can cause this situation. First, some data sources are not completely annotated. Second, each method has its own bias, meaning that each method can identify a subset of specific interactions. Third, large portions of the resulting dataset suffer from a high false-positive rate [22, 24, 25].

Comprehensive consideration of these data sources by the use of integration algorithms can solve the data bias inherently when using only a single data source and can also effectively increase the coverage of the interactome and decrease the false-positive rate [26]. Therefore, the development of new statistics and computational methods for integrating data from different databases is urgently needed and is a subject of concern in the present study [27]. Previous studies have directly utilised integration strategies that have not been properly evaluated, such as the intersection set of different networks (Intersection), the union set of different networks (Union), voting (which is a choice made by a network, Vote) [26], and the integration strategies based on Naive Bayes [28], Bayesian Networks [29], Logistic Regression [30], SVM [21], and decision trees, including Random Trees [31], Random Forest, and J48 [24]. For example, Lin and Chen applied a tree-augmented naive Bayesian (TAN) classifier to integrate heterogeneous data sources and generated fair results [28]; Wu et al. used SVM and Bayesian classifiers to detect whether a protein-protein interaction was reliable [21]; Gerstein et al. considered that voting did not take full advantage of the data source information in the process, and therefore, cannot generally obtain good results [32]; Ben-Hur and Noble deemed that SVM adopted different kernel functions depending on different integration tasks [33]; Jansen et al. and Rhodes et al. regarded that the premise of Naive Bayes was that the conditional probability of each attribute was independent [29, 34]; Sprinzak et al. thought that Logistic Regression actually was a generalised linear statistical model [30]; Chen and Liu believed that Random Forest combined many decision trees to enhance the correct rate of classification [35]; and by evaluating the precision, recall and area under the curve (AUC) scores of Support Vector Machine (SVM), Naive Bayes, Logistic Regression, Decision Tree, and Random Forest when predicting interactions, Qi et al. determined that Random Forest ranked as the top classifier for integration [24].

Although various types of integration strategies have been applied to the current research, the method of choice has not been considered. Although some researchers have simply evaluated some integration results, the comprehensive topological properties of the networks for different integration strategies and the impact of the outcomes of the typical applications based on these networks have not been rigorously evaluated.

In this paper, we combined 37 features representing 10 distinct groups of biological data sources based on former studies [24, 36, 37], including gene expression, physical interactions, domain interactions, HMS_PCI mass, TAP mass, yeast two-hybrid, genetic interactions, gene ontology (GO) annotations, and gene context analysis, to predict the more reliable protein-protein interactions. Our method utilised gold standard data sets and 11 commonly used methods (Union, Intersection, 2-Vote, 3-Vote, Naive Bayes, Bayesian Networks, Logistic Regression, SVM, Random Tree, Random Forest, and J48) from two types of integration strategies (empirical and machine learning) to integrate all of the interactions in previously mentioned databases; 2-Vote and 3-Vote indicate interactions that were supported by two and three databases, respectively. For seven machine-learning methods, we systematically evaluated the accuracy of correct classification, the area under the receiver operating characteristic (ROC) curve, the precision rate, recall rate, and the true-positive to false-positive ratio. To gain a more detailed understanding of the differences between these 11 new networks, we also compared the differences among their topological properties. For these integration strategies, topological properties, such as the number of proteins and interactions, the clustering coefficient, network density, average degree, and average path length, differed significantly between the different networks. Moreover, by analysing the ranks when predicting disease genes, searching for differences in detecting drug targets, and researching the modules for identifying molecular complexes, we found that the networks dramatically affected the outcomes of these typical applications. For example, when using phenotype similarity to detect disease genes, we obtained four different genes that were ranked as the top candidate in each of the 11 integration strategies. Compared to previous studies, the present study focuses more on the influence of different network integration strategies on typical biological applications, providing a novel perspective from which protein networks are studied from different viewpoints and an important basis for future network-based biological research.

2. Materials and Methods

2.1. DIP Database. DIP records experimentally detected protein interactions. Because the CORE set in the DIP database had been widely used to develop the prediction methods by the high-quality, high-throughput protein interaction data of it, and to study the properties of protein interaction networks, we selected the CORE set as the positive set for a gold standard database.

2.2. NEGATOME Database. The NEGATOME collects proteins that are experimentally supported noninteracting protein pairs via manual literature mining and analysing protein complexes from the RCSB Protein Data Bank (PDB). Because the manual dataset in the NEGATOME database does not contain high-throughput data and describes the unlikely direct physical interactions circumscribed only to mammalian proteins, most of which in this database are

Homo sapiens, we selected the manual dataset as the negative set for a gold standard database.

2.3. Testing Datasets. We used five sources (HPRD, BIND, MINT, IntAct, and BioGRID) for protein-protein interaction data, representing most of the authoritative databases. These databases contain data derived almost from high-throughput experiments based on literature mining, yeast two-hybrid, mass spectrometric, and anti-tag coimmunoprecipitation experiments. However, approximately half of the interactions obtained from high-throughput experiments may represent false-positives as estimated by Von Mering et al. Therefore, it is critical to determine whether the interactions are authentic or pseudo.

2.4. Gene Expression. Genes that are mRNA coexpressed typically indicate protein interactions [38]. We collected 28 gene expression profiles, including more than 5000 samples of different tissues from the Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/), as previously described by Xu et al. [36]. Each gene containing a missing value was deleted, and all of the expression values were log-2 transformed. We combined any probes containing the same Gene Identifier. We then calculated the Pearson correlation coefficient (PCC) between each pair of genes to obtain a correlation coefficient matrix.

2.5. Domain-Domain Interactions. A domain is a structural or functional protein subunit, and the interaction between two proteins often involves binding between pairs of their constituent domains. Therefore, the selection of domains as characteristics is credible. We obtained domain information from the PFAM database, which is a large collection of protein families and is authoritative about domains. Additionally, domain-domain interaction information was obtained from the DOMINE database, which is a database of known and predicted protein domain (domain-domain) interactions. The database contains interactions inferred from PDB entries and those that were predicted by 13 different computational approaches using PFAM domain definitions. It contains 26,219 domain-domain interactions among 5,410 domains.

2.6. Physical Protein-Protein Interactions. We collected all of the interactions from the BioGRID, BIND, IntAct, HPRD, and MINT databases. All of the interactions that were not mapped to homologous human interaction proteins by HomoloGene (http://www.ncbi.nlm.nih.gov/homologene) in NCBI [39] were deleted. The physical protein-protein interaction scores ranged from 0 to 5; 0 meant that the interaction was from none of these databases, while 5 meant that the interaction was supported by each of these databases.

2.7. High-Throughput Direct PPI Dataset. The high-throughput direct PPI dataset contains two types: (1) derived from mass spectrometry and (2) derived from Y2H. In the mass spectrometry dataset, two subdatasets, TAP [40] and HMS-PCI [41], utilised two different protocols for this technique. We used the high-throughput PPI dataset provided by Qi et al. [24].

2.8. Human Phenotype. Function deletion in interactions or functionally related proteins frequently resulted in similar phenotypes [41–43]. We mapped the interactions that Han et al. attributed to homologous human interactions to obtain more accurate results.

2.9. Genetic Interactions. A synthetic genetic analysis (SGA) was used to reveal genetic interactions in *Saccharomyces cerevisiae* [44, 45]. Some reports have demonstrated a significant overlap between protein-protein interactions and genetic interactions [46]. Therefore, most neighbours of genetic interaction genes can be used to predict protein-protein interactions [45].

2.10. Biological Functional Annotation. Compared to interactions of different biological functions, protein-protein interactions are more likely to occur in proteins with similar biological functions. Moreover, proteins sharing a more specific annotation tend to interact with each other compared to those that share a more common annotation.

2.11. Gene Context Analysis. The gene context is based on genome sequences to infer *in silico* protein-protein interactions [22]. The gene context includes three types: gene fusion, gene cooccurrence, and gene neighbourhood.

Human phenotype, genetic interaction, biological functional annotation, and gene context analyses have been previously performed by Xia et al. [37] to predict interactions from model organisms.

To avoid the impact of human factors on the results analysis, we constructed all the machine learning integration strategies with a unified software platform, WEKA (Waikato Environment for Knowledge Analysis), which is widely used in classification. WEKA, a public data-mining platform, collects a large number of machine learning algorithms that are used to undertake the task of data mining, including preprocessing, classifying, clustering, associating, attribute selecting, and visualising.

2.12. Seven Machine Learning Classifiers Constructed by Using the Gold Standard Datasets. All of the protein data that we obtained were derived from different databases, and each database has its own presentation pattern, such as Gene Identifier, Gene Symbol, Accession Number, and UniProtKB Number. To unify the data, we converted all of the protein presentation patterns into Gene Identifiers. Then, we deleted any interactions that were not mapped to the Gene Identifiers or homologous human interactions by NCBI HomoloGene. After obtaining the gold standard protein networks, we constructed seven different machine learning classifiers using seven integration strategies (Naive Bayes, Bayesian Networks, Logistic Regression, SVM, Random Tree, Random Forest, and J48) (Figure 1).

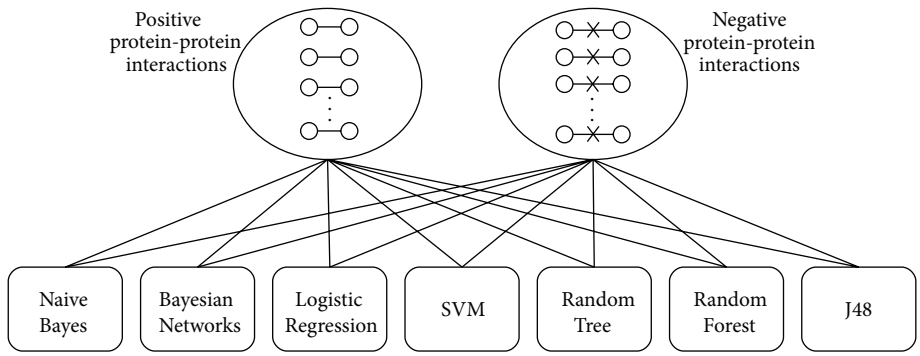


Figure 1: Seven machine learning classifiers constructed by using the gold standard datasets. The gold standard datasets, positive protein-protein interactions in the DIP database, and negative protein-protein interactions in the NEGATOME database were used to construct the seven machine learning classifiers based on the following methods: Naive Bayes, Bayesian Networks, Logistic Regression, Support Vector Machine (SVM), Random Tree, Random Forest, and J48.

Table 1: Performance of the classifiers constructed by seven machine learning integration strategies.

Strategy	ACC	AUC	Precision	Recall	FP rate	TP/FP
Naive Bayes	0.5391	0.62	0.524	0.539	0.518	1.041
Bayesian Networks	0.6325	0.736	0.683	0.632	0.418	1.512
Logistic Regression	0.7188	0.772	0.724	0.719	0.275	2.615
SVM	0.7144	0.723	**0.738**	0.714	**0.267**	**2.674**
Random Tree	0.6568	0.648	0.656	0.657	0.35	1.877
Random Forest	**0.7196**	**0.787**	0.72	**0.72**	0.292	2.466
J48	0.6808	0.671	0.681	0.681	0.323	2.108

Note: ACC stands for the accuracy of the correctly classified items (after a 10-fold cross-validation). AUC indicates the area under the ROC curve. Precision is the number of true positives divided by the total number of elements labelled as belonging to the positive class. Recall (also referred to as the True Positive Rate) represents the number of true positives divided by the total number of elements that actually belong to the positive class. The FP rate indicates the false positive rate. TP/FP reveals the true positive to the false positive ratio. Bold type indicates the maximum value in the ACC, AUC, Precision, Recall, and TP/FP columns and indicates the minimum value in the FP rate column.

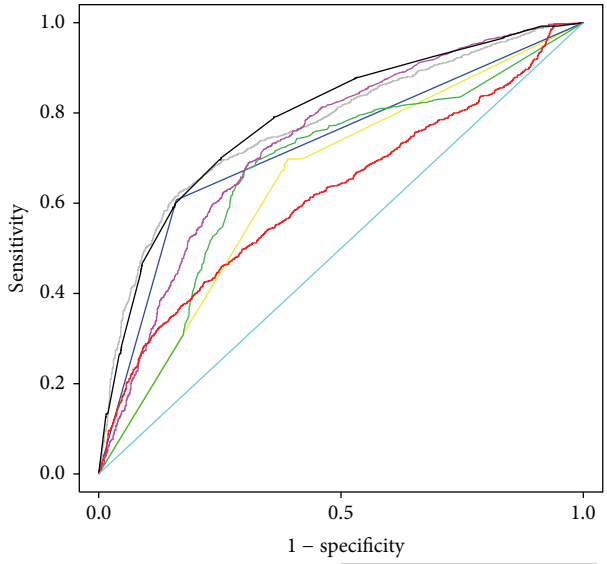
3. Results

In this study, we used four empirical integration strategies and seven machine learning classifiers constructed by the reliable positive and negative gold standard sets from DIP and NEGATOME, respectively, to integrate the protein-protein interaction networks. Some indicators, such as the accuracy of those correctly classified, the area under the ROC curve, and the precision and recall rates, are typically used to evaluate a supervised machine learning method; therefore, we initially evaluated the performance of these seven machine-learning classifiers in these ways.

3.1. Performance of the Classifiers Constructed by Seven Machine Learning Integration Strategies. From the seven different integration strategies, the seven classifiers showed quite different classification results. The ACC score, AUC score, precision and recall rates, and TP/FP score of each integration strategy were significantly distinct (Figure 2, Table 1). For example, the ACC score ranged from 0.5391 in Naive Bayes to 0.7196 in Random Forest, and the area under the ROC curve ranged from 0.62 in Naive Bayes to 0.787 in Random Forest. Therefore, different integration strategies affect the outcome of the classification.

3.2. Eleven New Networks Built by Empirical and Machine Learning Integration Strategies. After inputting 145,534 interaction pairs from the five databases into these funnel-like classifiers, we obtained 11 different new networks (Figure 3). As shown in Table 2, the 11 new networks are significantly different from each other. Although the input network was constant, the ratio of the number of predicted protein pairs to the originally considered protein pairs was remarkably discrepant. It is clear that the differences between the seven machine learning networks are not significant; in other words, the machine learning strategies are somewhat stable. For example, the coverage of each network range from 0.7874 (Random Tree) to 0.9773 (SVM); however, most of the coverage in the machine learning strategies were approximately 95%. However, a remarkable distinction was present in four empirical strategies. For example, Intersection considered only 0.34% of the interactions to be true interactions, 2-Vote considered 28.01% of the interactions to be true interactions, and Union considered all of the interactions to be true interactions.

Seven machine learning networks in Table 3 show that although a certain ratio of repeats was observed, the interactions in each machine learning network were not the same. For example, compared with the original network,



Area under the ROC curve	
Integration methods	Area
Random Forest	0.787
Bayesian Networks	0.736
Logistic Regression	0.772
SVM	0.723
J48	0.671
Random Tree	0.648
Naive Bayes	0.62
Random case	0.5

Figure 2: ROC curves for seven machine learning integration strategies using 10-fold cross-validation against the gold standard datasets. Each point on the ROC curves of the seven integration strategies is created by the unique sensitivity and specificity against a specific likelihood ratio cut-off. Each name of the curve derived from the different integration strategies is shown in the legend. The different colours stand for the different curves for the different strategies. The area under the curve is also presented in the figure. Sensitivity and specificity are calculated during the 10-fold cross-validations.

Table 2: The coverage of each network built by 11 integration strategies.

Strategy	Number	Coverage
Union	145534	1
Intersection	497	0.0034
2-Vote	40766	0.2801
3-Vote	12891	0.0886
Naive Bayes	134095	0.9214
Bayesian Networks	140956	0.9685
Logistic Regression	140746	0.9671
SVM	142226	0.9773
Random Tree	114598	0.7874
Random Forest	139082	0.9557
J48	120541	0.8283

Note: Number stands for the number of the predicted interaction pairs by each integration strategy. Percentage represents the ratio of the number of the predicted interaction pairs to the number of total interaction pairs in the five databases.

Table 3: The duplication of seven machine learning networks and all 11 integration networks.

Seven machine learning networks			All 11 integration networks		
DT	Number	Percentage	DT	Number	Percentage
0	1808	1.24%	1	1683	1.16%
1	1277	0.88%	2	134	0.09%
2	79	0.05%	3	718	0.49%
3	299	0.21%	4	777	0.53%
4	1410	0.97%	5	1234	0.85%
5	9434	6.48%	6	7653	5.26%
6	41487	28.51%	7	32465	22.31%
7	89740	61.66%	8	71478	49.11%
			9	20680	14.21%
			10	8400	5.77%
			11	312	0.22%

Note: DT stands for the number of times in which all of the interactions were duplicated. Number represents the number of such interactions. Percentage reveals the ratio of the number of such interactions to the total number in the original network.

the number of interactions that did not appear in any of the machine learning classifier outputs was 1,808 (1.24% of the total), while the number of interactions in all of the classifiers' output was 89,740 (61.66% of the total). As indicated in all 11 integration networks in Table 3, in all of the networks, including empirical and machine learning strategies, the number of interactions varies among 11 networks generated by different integration strategies. For example, the number of interactions that appeared in eight networks was 71,478 (49.11%), and the number of interactions that appeared in 11 networks was 312 (only 0.22%).

3.3. Topological Properties of the 11 Empirical and Machine Learning Networks. We first analysed the network topological properties of each integration network; the two most critical attributes in the network are the distance and the number of connections [47]. Almost all of the other topological properties are based on these two properties. We calculated the number of proteins and interactions, network diameter, average degree, network density, average path length, and global clustering coefficient for each network (Table 4).

The network diameter is the maximum eccentricity of any point in the protein network. It represents the greatest distance between protein pairs. The density of a protein network is the total number of interactions divided by the total number of possible interactions. The average path length represents the average distance of the shortest path between all of the node pairs. Additionally, it provides the overall efficiency of information or mass transport in a network. The global clustering coefficient represents the degree to which the proteins in a protein network tend to cluster together.

Tremendous differences were found between the 11 networks of empirical and machine learning strategies. For example, the number of proteins in the networks integrated

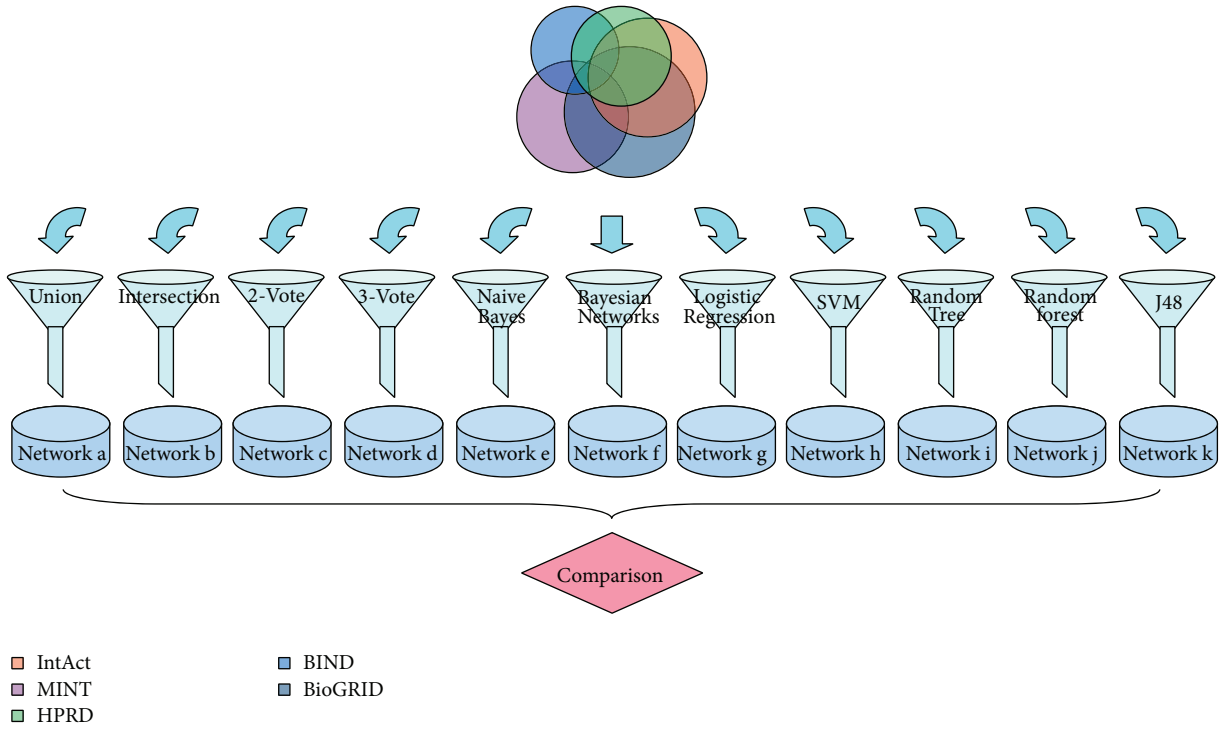


Figure 3: Eleven new networks built by empirical and machine learning integration strategies. Eleven new networks constructed by funnel-like empirical and machine learning integration strategies, namely, Union, Intersection, 2-Vote, 3-Vote, Naive Bayes, Bayesian Networks, Logistic Regression, SVM, Random Tree, Random Forest, and J48, from the entire set of data in the IntAct, MINT, HPRD, BIND, and BioGRID databases.

Table 4: The topological properties of the 11 new empirical and machine learning networks.

	Empirical				Machine learning						
	Union	Intersection	2-Vote	3-Vote	Naive Bayes	Bayesian Networks	Logistic Regression	SVM	Random Tree	Random Forest	J48
Proteins	**14936**	507	9548	5558	14840	14869	14890	**14895**	14486	14860	14570
Interactions	**145534**	497	40766	12891	134095	140956	140746	**142226**	114598	139082	120541
Diameter	15	6	**16**	15	15	15	**16**	**16**	**16**	15	**16**
Degree	**19.49**	1.96	8.54	4.64	18.07	18.96	18.90	**19.10**	15.82	18.72	16.55
Density	0.00130	**0.00387**	0.00089	0.00083	0.00122	**0.00128**	0.00127	**0.00128**	0.00109	0.00126	0.00114
ASP	2.9216	**1.9164**	4.4487	4.7394	2.9372	2.9245	2.9256	**2.9217**	3.0447	2.9280	3.0103
CC	0.0206	**0.1262**	0.0471	0.0340	0.0161	**0.0204**	0.0197	**0.0204**	0.0156	0.0194	0.0170

Note: Proteins, Interactions, Diameter, Degree, and Density indicate the number of proteins, the number of interactions, the network diameter, the average degree and the network density, respectively. ASP and CC are the average path length and clustering coefficient, respectively. Bold type indicates the minimum value for average path length and the maximum value for the other topological properties of the empirical and machine learning methods.

by machine learning strategies ranged from 507 to more than 14,000, and the average degree ranged from 1.96 to more than 15. Additionally, the average path length and clustering coefficient were dramatically varied.

If we account only for the networks that were integrated by the empirical strategies, almost all of the properties were dramatically varied; for example, the range of changes in the number of proteins and the interactions were especially large. Dramatic variation was also observed in the remaining properties, such as the network diameter, average degree, network density, average path length, and clustering coefficient.

Considering only the networks integrated by the machine learning strategies, although some properties, such as the number of proteins, the network diameter, the network density and the shortest path length, did not vary, other properties varied dramatically. For example, the number of interactions ranged from 114,598 in the Random Tree to 142,226 in the SVM network; the average degree of the networks ranged from 15.82 in the Random Tree network to 19.10 in the SVM network; and the clustering coefficient ranged from 0.0156 in the Random Tree network to 0.0204 in the Bayesian Networks network and the SVM network.

Table 5: Description of the top genes in 11 integration networks from the detection of disease genes based on a phenotype similarity study.

Strategy	Gene symbol	Official full name
Union	ATP2B2	ATPase, Ca++ transporting, plasma membrane 2
Intersection	MYD88	Myeloid differentiation primary response 88
2-Vote	TGFBR2	Transforming growth factor, beta receptor II (70/80 kDa)
3-Vote	TGFBR2	Transforming growth factor, beta receptor II (70/80 kDa)
SVM	ATP2B2	ATPase, Ca++ transporting, plasma membrane 2
Naive Bayes	ATP2B2	ATPase, Ca++ transporting, plasma membrane 2
Random Tree	GRM7	Glutamate receptor, metabotropic 7
J48	ATP2B2	ATPase, Ca++ transporting, plasma membrane 2
Logistic Regression	ATP2B2	ATPase, Ca++ transporting, plasma membrane 2
Random Forest	ATP2B2	ATPase, Ca++ transporting, plasma membrane 2
Bayesian Networks	ATP2B2	ATPase, Ca++ transporting, plasma membrane 2

Therefore, the alteration in the topological properties reveals that the different integration strategies dramatically affect the outcomes. These integration strategies affect the number of proteins and interactions in the networks, consequently affecting the network aggregation, mass transport, and connectivity.

3.4. Detection of Disease Genes Using Phenotype Similarity Based on Networks. Protein interaction data are ultimately integrated to facilitate actual applications. The advancement of biotechnology enables the proteome scope of protein interaction networks, making the networks become more and more attractive to researchers studying systems biology [48]. Typically, researchers tend to use protein interaction networks to identify disease candidate genes [49, 50], drug targets [51, 52], and functional modules [53].

The prediction of disease genes based on a protein network is an important typical application of biological networks [54] and is also vital to the development of physianthropy [20]. To detect disease genes, we utilised the method of Lage et al. [54], which is based on phenotype similarity. In this method, disease gene prediction is accomplished based on the assumption that proteins that are directly connected to disease proteins tend to have the same disease phenotype as the disease protein [55–58].

We used epithelial ovarian cancer (EOC) as a specific case. After downloading all 301 genes in the linkage interval on 3p25-22 [59] from GENE of NCBI [60], we identified the subnetworks of these 301 genes in the 11 networks obtained from the 11 integration strategies. Then, a score was obtained for every interaction that was an edge in the subnetworks via Kasper's scoring rules. Next, according to the method described by van Driel et al. [61], which is based on OMIM [62] and MeSH [63], we calculated the similarity between each phenotype and EOC as the score of the protein that was the node in the subnetworks [64]. According to the following formula, we obtained the final score for each candidate gene:

$$\text{Score} = \sum_{i=1}^{N} S_i P_i, \tag{1}$$

where N is the number of partners connected to the candidate gene, S_i is the interaction score, and P_i is the protein score.

Finally, we sorted the entire candidate genes to identify the one that had the highest score. Table 5 lists the highest scoring candidate genes for each integration network.

These findings clearly revealed that four different genes, ATP2B2, MYD88, TGFBR2, and GRM7, were ranked as the top genes in each of 11 integration strategies. This finding indicates that the empirical and machine learning strategies dramatically affected the overall outcomes. Separately, seven machine learning strategies mainly identified ATP2B2 as the top gene; however, Random Tree identified GRM7 as the top gene; overall, this result was relatively stable. However, four empirical strategies yielded three different top genes, indicating that the empirical strategy was quite unstable and seriously impacted the reliability of the results.

3.5. Detection of Disease Genes Using a Network-Based Random Walk with Restart (RWR). Based on an early disease-gene screening method based on phenotype similarity or network topological properties and the advances of genome sequencing, gene expression analysis and other parallel technologies, it is clear that new disease-gene screening methods are emerging [54, 65, 66]. Well-known studies have demonstrated that the RWR method is superior to other methods, such as methods based on clustering or based on neighbouring nodes [66]. Therefore, we used the RWR method to screen for causative disease genes.

RWR refers to a process in which a given node in a network is used as a starting point upon which iterations are performed; at each iteration, the current node is used as a starting point for a transfer to a randomly selected adjacent node as follows:

$$p^{t+1} = (1 - r) W p^t + r p^0, \tag{2}$$

where p^t is a vector that represents the probability of a certain node being the random walk node at time t in the network, r represents the probability of the random walk node returning to the starting node at any moment, and W represents the adjacency matrix after the column standardisation of the network.

Table 6: The performance of the detection of disease genes using RWR.

Strategy	Rank	Nodes	Rank ratio
Naive Bayes	2065.86	14840	0.1392
Logistic Regression	2113.31	14890	0.1419
SVM	2133.86	14895	0.1433
Union	2146.21	14936	0.1437
Random Forest	2136.90	14860	0.1438
Bayesian Networks	2139.83	14869	0.1439
J48	2370.10	14570	0.1627
Random Tree	2697.93	14486	0.1862
3-Vote	1306.61	5558	0.2351
2-Vote	2245.89	9548	0.2352
Intersection	123.5	507	0.2436

Note: Rank indicates the average rank of the nonseed genes in several repeated experiments; the number of repetitions depended on the number of remaining genes. The rank ratio reveals the average rank divided by the total number of nodes in each network. The rank ratio was used to evaluate whether the performance of the integration strategy was outstanding. The smaller the scale is, the better the integration strategy is.

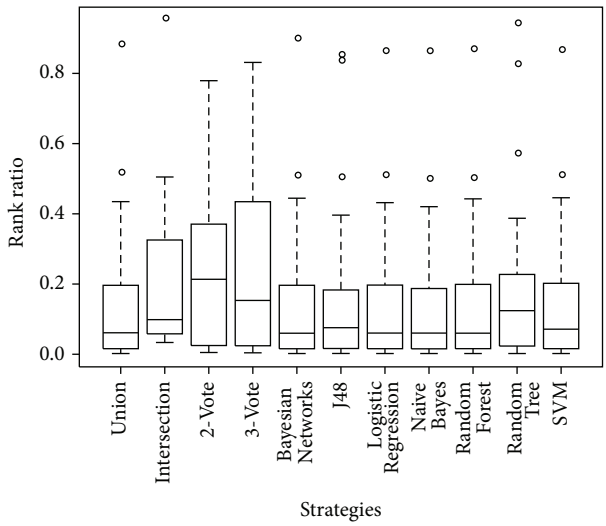


Figure 4: The performance of detecting disease gene using RWR. We used a box-plot to show the rank difference between each of the 11 integration strategies. Apparent distinctions exist between the different networks by different integration strategies.

We selected 29 disease genes of brain tumours, including neurofibroma, glioma, glioblastoma, and astrocytoma, from the Cancer Gene Census (CGC) [67, 68]. One gene was randomly selected to verify the prediction efficiency, and the remaining genes were used as seeds for the RWR algorithm. The number of repetitions depended on the number of remaining genes. We considered whether the integration strategy was outstanding based on the ratio of the average position of a nonseed gene in these repeated experiments to the total number of nodes (Table 6, Figure 4).

By comparing the rank ratios in Table 6 and Figure 4, it is clear that (A) the average rank of the remaining (single) disease gene was approximately 14% of the total number of nodes, indicating a satisfactory performance at discovering disease genes by the RWR method, and (B) although some of the machine learning strategies were stable in the rank ratio (e.g., the rank ratio of Naïve Bayes, Logistic Regression, SVM, Random Forest, and Bayesian Networks were approximately 0.14), the rank ratios of the other two machine learning strategies were 0.1627 in J48 and 0.1862 in Random Tree; these values were significantly different from the former five strategies. Furthermore, the rank ratios of four empirical strategies were remarkably distinct from the machine learning strategies; for example, the ratio of 3-Vote, 2-Vote and Intersection were approximately 0.24. Therefore, the different strategies greatly impacted the rank of the nonseed gene by the RWR method.

We next selected all of the disease genes as seeds to obtain the top 10 genes for all of the generated networks except for the seeds (Table 7). Table 7 indicates that the gene lists discovered by the empirical methods are significantly different from the gene lists discovered by the machine learning methods. For example, the genes identified by empirical strategies, such as YWHAB, RAD50, YWHAZ, YWHAE, ERBB2, and RB1, were not identified by any of the machine learning strategies. The top 10 genes detected by the four empirical strategies were also remarkably distinct; for example, UBC, TAF1, MYC, and HNF4A were identified by Union but were not identified by any of the other empirical strategies. Although the genes that were identified by the machine learning methods shared some overlap, different methods also identified different genes; for example, DTNBP1 was only identified by J48. Therefore, the different strategies dramatically impacted the top 10 genes identified by the RWR method.

3.6. The Approach of Discovering Drug Targets Based on Network Topology Properties. As mentioned above, the identification of drug targets is one typical use of a protein interaction network. Similar to Zhu et al. [69], we applied the original protein network topology-based approach to identify drug targets.

Previous studies have shown that compared to general network proteins, drug target proteins are significantly different with respect to their topological properties. For example, the degree of a drug target is larger [70], the average distance and the shortest length between two drug targets are shorter than between a drug target and a general protein, the proportion of the target proteins in the neighbouring nodes of a target protein is significantly higher than the proportion of the target proteins in the neighbouring nodes of a general protein, and the clustering coefficient of a drug target is significantly lower than for a general protein [69].

We obtained the drug target information from DrugBank [71] on March 16, 2013; we then mapped these target genes to each protein interaction network. We next selected five measures to identify drug target proteins; these measures included the degree, 1N index, clustering coefficient, and the average distance and shortest path length between a protein and a drug target protein. The 1N index was the proportion of target proteins in the neighbouring nodes of a protein.

Table 7: The top 10 genes of all of the genes, except for the seed genes, from 11 integration networks in the detection of disease genes using RWR.

Strategy	The symbols of the top 10 genes
Union	UBC, TAF1, MYC, HNF4A, SMARCA4, ELAVL1, CDK2, FASLG, XRCC6, and SDHA
Intersection	YWHAB, RAD50, CTNNB1, GRB2, SHC1, ABL1, YWHAZ, YWHAE, ERBB2, and RB1
2-Vote	MLH1, PTPN6, XRCC6, EXO1, ARHGDIA, VAV3, HRAS, FASLG, APP, and TNIK
3-Vote	PTPN6, MAX, ZHX1, CCDC90B, MLH1, EXO1, IMMT, VIM, ASF1B, and ASF1A
SVM	UBC, TAF1, MYC, HNF4A, SMARCA4, ELAVL1, CDK2, FASLG, XRCC6, and SDHA
Naive Bayes	UBC, TAF1, MYC, HNF4A, SMARCA4, ELAVL1, CDK2, XRCC6, FASLG, and SDHA
Random Tree	UBC, MYC, XRCC6, SMARCA4, ARHGDIA, TAF1, ABL1, ELAVL1, FASLG, and CDK2
J48	UBC, TAF1, MYC, HNF4A, SMARCA4, XRCC6, ELAVL1, FASLG, CDK2, and DTNBP1
Logistic Regression	UBC, TAF1, MYC, HNF4A, SMARCA4, XRCC6, ELAVL1, CDK2, FASLG, ARHGDIA
Random Forest	UBC, TAF1, MYC, HNF4A, SMARCA4, ELAVL1, CDK2, FASLG, XRCC6, and SDHA
Bayesian Networks	UBC, TAF1, MYC, HNF4A, SMARCA4, ELAVL1, CDK2, FASLG, XRCC6, and SDHA

Note: the description of these genes was listed in Supplementary Table S1 of the Supplementary Material available online at http://dx.doi.org/10.1155/2014/296349.

Table 8: Performance of each network built by integration strategies for the discovery of drug targets based on topological properties.

Strategy	1N	Target	Node	Target ratio
Union	82	**1984**	**14936**	0.132833
Intersection	40	133	507	**0.262327**
2-Vote	79	1464	9548	0.153331
3-Vote	55	885	5558	0.15923
SVM	83	1974	14895	0.132528
Naive Bayes	83	1969	14840	0.132682
Random Tree	81	1941	14486	0.133991
J48	81	1945	14570	0.133493
Logistic Regression	**85**	1981	14890	0.133042
Random Forest	83	1972	14860	0.132705
Bayesian Networks	83	1973	14869	0.132692

Note: 1N indicates the number of targets included in the top 100 proteins. Target indicates the number of targets in the network. The target ratio reveals the percentage of targets in a network. The bold type indicates the maximum values in the 1N, Target, Node, and Target ratio columns.

We ranked each protein based on these five measures in each of the networks. We then determined the number of drug targets that were included in the top 100 proteins of each network (1N) and the proportion of target proteins in all of the proteins (Table 8).

Table 9: The duplication of targets in the top 100 in each network built by all 11 integration strategies.

DT	Number	Percentage
1	58	0.3391
2	21	0.1228
3	5	0.0292
4	4	0.0234
5	6	0.0351
6	3	0.0175
7	6	0.0351
8	27	0.1579
9	21	0.1228
10	11	0.0643
11	9	0.0526

Note: DT indicates the duplication times of the targets that appear in the top 100 of each network. Number represents the number of targets. Percentage reveals the ratio of the number of targets to the total of all of the targets that appear in the top 100 of each network.

Due to the characteristics of the drug targets, based on the machine learning strategies, the proportion of drug target proteins in the top-ranked 100 proteins was approximately 83%, and the target ratio was approximately 0.13. Additionally, there was only a small change between the different machine learning strategies; for example, Random Tree and J48 each identified 81 targets, while Logistic Regression identified 85 targets. However, the results obtained from the different empirical integration strategies were significantly different. For example, Union identified 82 drug targets in the top 100 proteins from a network containing 1,984 drug targets, while Intersection identified 40 drug targets in the top 100 proteins from a network containing 133 drug targets.

As shown in Table 9, the numbers of targets were different. For example, 58 drug targets were uniquely detected by one integration strategy, accounting for 33.91% of the 171 targets that appear in the top 100 of each network; however, only four drug targets were simultaneously detected by four integration strategies, accounting for 2.34% of the 171 targets that appear in the top 100 of each network. Therefore, the outcomes of the drug target discovery process were dramatically affected by the different strategies.

3.7. Identification of Molecular Complexes Based on the MCODE Clustering Algorithm. Molecular complexes are key elements in molecular function. Human disease is closely correlated with human molecular complexes, and molecular complexes are widely applied in molecular functional annotation and disease prediction. Therefore, it is critical to identify molecular complexes [72]. Because the protein interaction network contains functional annotation data, it is important to identify molecular complexes from protein interaction networks. Because a subunit of the protein exercises a biological function, the prediction of the function of unknown proteins has been demonstrated to be of great significance [73].

The identification of molecular complexes is an important application in biological networks. For example, Wu et al.

Table 10: The topology properties of the molecular complexes found by 11 networks built by integration strategies based on the MCODE clustering algorithm.

	Empirical				Machine learning						
	Union	Intersection	2-Vote	3-Vote	Naive Bayes	Bayesian Networks	Logistic Regression	SVM	Random Tree	Random Forest	J48
Proteins	**63**	5	26	29	55	**65**	64	61	40	59	48
Interactions	**1721**	9	169	92	438	**1787**	1761	1628	702	1497	1028
Diameter	2	2	2	**7**	**5**	2	2	2	2	2	2
Degree	**54.127**	3.6	12.769	5.241	15.927	**54.985**	54.844	53.377	35.1	50.746	42.833
Density	**0.873**	0.9	0.511	0.187	0.295	0.859	0.871	0.890	0.9	0.875	**0.911**
ASP	1.127	**1.1**	1.489	3.264	2.773	1.141	1.129	1.110	1.1	1.125	**1.089**
CC	0.903	0.9	**0.940**	0.816	0.851	0.894	0.899	0.913	0.904	0.895	**0.928**

Note: Proteins, Interactions, Diameter, Degree, and Density indicate the number of proteins, the number of interactions, network diameter, average degree, and network density, respectively. ASP and CC are the average path length and clustering coefficient, respectively. Bold type indicates the minimum value on an average path length and the maximum value in the other topological properties of empirical and machine learning methods.

compiled the redundant human complexes to build a comprehensive catalogue and then investigated the relationship between protein complexes and drug-related systems [72]. Song and Singh analysed proteins, complexes, and processes and considered physical interactions within and across complexes and biological processes to understand the protein essentiality [74]. Zhang and Shen analysed functional modules based on a protein-protein network analysis in ankylosing spondylitis [75].

In this paper, we utilised a Cytoscape [76] plug-in called MCODE, which is based on the MCODE [73] clustering algorithm; this plug-in mines tightly connected regions in protein interaction networks that represent molecular complexes. Cytoscape is free software program that graphically displays, edits, and analyses networks. It supports a variety of network description formats, and the user can add rich annotation information to the networks. a large number of functional plug-ins that were developed by developers and third parties can be used for in-depth analysis of network problems. We analysed the topological properties of each single top molecular complex in each network and compared their intersections (Table 10).

Table 10 reveals that the different networks obtained from different integration strategies affected the finding on the effect of molecular complexes. Overall, even though the diameters of the networks and the clustering coefficient were nearly identical, the number of proteins and interactions differed greatly in both the empirical and machine learning strategies. For example, only five proteins and nine interactions were identified in the molecular complexes mined by Intersection, and only 40 proteins and 702 interactions were identified by Random Tree; however, Union identified 63 proteins and 1,721 interactions, and Bayesian Networks identified 65 proteins and 1,787 interactions when identifying molecular complexes. Additionally, the average degree of each network ranged from 3.6 in the Intersection network to 54.127 in the Union network and from 35.1 in the Random Tree network to 54.985 in the Bayesian Network. The network density and average path length also varied in both the empirical and machine learning strategies.

Table 11: Gene symbol and degree of the proteins that have the largest degree in every molecular complex of each network.

Strategies	Gene symbol	Degree
Union	RPL5, UBC	64
Intersection	IRAK1, IRAK2, and IRAK3	4
2-Vote	UCHL5	25
3-Vote	IKBKG	10
SVM	RPS8, RPS2, RPL5, RPL11, RPL18, RPS16, RPS6, RPL19, RPS13, RPL21, RPL6, RPL10A, UBC, RPS4X, RPL4, and RPS3	60
Naive Bayes	MED26, MED29	27
Random Tree	RPL5, UBC, and RPL4	39
J48	RPS2, RPL5, RPL11, RPS6, RPL19, RPL21, RPL6, RPL10A, UBC, RPS4X, RPL14, and RPL4	47
Logistic Regression	RPL5	65
Random Forest	RPL11, RPS6, RPL14, and RPL4	58
Bayesian Networks	RPL18, RPS16, RPS6, RPS4X, RPS8, RPS2, RPL5, RPL21, UBC, and RPL4	64

Table 11 lists the proteins that displayed the largest degree in each molecular complex of each network. It is clear that every molecular complex is different from the others because the proteins with the largest degree are different (Table 11).

The proteins displaying the smallest degree (4) were IRAK1, IRAK2, and IRAK3 (based on Intersection) in the molecular complex identification, while the proteins with the largest degree (64) were RPL5 and UBC by Union in molecular complex finding by empirical strategies. Additionally, Logistic Regression revealed that RPL5 had the largest degree (65), while Naive Bayes revealed that MED26 and MED29 displayed the smallest degree (27) by machine learning strategies. Therefore, large distinctions exist between the empirical and machine learning strategies when identifying molecular complexes.

4. Discussion

Protein-protein interaction studies act as new method for improving our understanding of molecular physiological processes. With the growing number of in-depth studies on protein-protein interaction networks, scientists are gaining knowledge of the interactions from various methods. Therefore, the key to network analyses is determining which integration strategy should be implemented. In this study, we analysed and evaluated the networks integrated by 11 commonly used strategies of two types of integration strategies, empirical and machine learning, including Union, Intersection, 2-Vote, 3-Vote, Bayesian Network, Support Vector Machine, Naive Bayes, Random Tree, J48, Logistic Regression, and Random Forest. By comparing the scores and the ranks, these strategies detected disease genes based on phenotype similarity and the RWR algorithm. Based on rank, the networks identified drug targets based on five measures, including average degree, 1N index, clustering coefficient, average path length, and shortest path; the topological properties of the molecular complexes that were identified were based on a Cytoscape plug-in called MCODE. Thus, we conclude that different integration strategies can obtain extremely different outcomes for these typical applications.

Most of the methods of the existing studies are to evaluate the character of the network itself. For example, Qi et al. found that Random Forest performed best of the six methods that they analysed [24]. Although Random Forest performed better based on ACC and AUC, with scores of 0.7196 and 0.787, respectively, subsequent evaluations confirmed that it is insufficient to determine the best integration strategy based solely on accuracy. Ultimately, one must also consider the comprehensive applications. Nevertheless, we did not only analyse the quality of the networks based simply on the integration of a wide range of data. In other words, although we analysed the AUC, accuracy, and topological properties, we also focused on typical practical applications, such as disease gene discovery, drug target detection, and molecular complex identification. We then compared the differences between the various networks in these applications. Therefore, this study is more biologically significant than previous studies, and it provides a novel perspective from which scholars can study protein networks.

It should be emphasised that a substantial amount of in-depth exploration of this topic remains. First, the integration strategies can be combined with other methods for further improvement. For example, the Naive Bayesian method used by Lin and Chen [28] is a tree-like Naive Bayesian method. Alternatively, a variety of integration strategies may be combined in a manner that emphasises the advantages of each integration strategy to improve the results of the integration. Second, because some features, such as phenotype similarity, genetic interaction, and shared GO annotation, which were utilised in IntNetDB described by Xia et al. [37], and TAP, HMS-PCI, and Y2H, which were utilised by Qi et al. [24], do not consider current data, deviations may exist in the results. However, our results are reliable because the same input data were used for all of the integration strategies; therefore, these deviations were not significant.

Although the processes of disease gene discovery and drug target detection revealed the stability of the seven machine learning strategies, these supervised machine learning strategies should have been similar; any difference between them warrants further examination. However, some properties used to identify molecular complexes have revealed the instability of several machine learning strategies. Almost all of the typical applications indicate that empirical strategies are quite unstable; however, these empirical strategies are applied in a substantial number of studies. Consequently, if these strategies are not evaluated, the resulting data will be unreliable, strongly influencing the studies.

Integration strategies are the key step in the network analysis, and they severely affect the outcomes of the various applications. Therefore, because technological advancement dictates the subsequent update of data and the integration strategies, the integration of the updated data becomes even more important. Software and websites that can rapidly integrate these updated data should be developed so that researchers can gain access to more reliable data and complete protein-protein interaction networks.

Conflict of Interests

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

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant no. 31100948), the Natural Science Foundation of Jiangsu Province (Grant no. BK20131385), and the Natural Science Foundation of Heilongjiang Province (Grant no. D201114).

References

[1] A. S. Coates, E. K. A. Millar, S. A. O'Toole et al., "Prognostic interaction between expression of p53 and estrogen receptor in patients with node-negative breast cancer: results from IBCSG Trials VIII and IX," *Breast Cancer Research*, vol. 14, article R143, no. 6, 2012.

[2] C. Wostenberg, J. W. Lary, D. Sahu et al., "The role of human Dicer-dsRBD in processing small regulatory RNAs," *PLoS ONE*, vol. 7, no. 12, Article ID e51829, 2012.

[3] S. Lee, M. Kwon, J. M. Oh, and T. Park, "Gene-gene interaction analysis for the survival phenotype based on the cox model," *Bioinformatics*, vol. 28, no. 18, pp. i582–i588, 2012.

[4] S. W. Chi, G. J. Hannon, and R. B. Darnell, "An alternative mode of microRNA target recognition," *Nature Structural and Molecular Biology*, vol. 19, no. 3, pp. 321–327, 2012.

[5] P. Braun, A. R. Carvunis, B. Charloteaux et al., "Evidence for network evolution in an Arabidopsis interactome map," *Science*, vol. 333, no. 6042, pp. 601–607, 2011.

[6] M. Li, X. Xu, and Y. Liu, "The set2-RPB1 interaction domain of human RECQ5 is important for transcription-associated genome stability," *Molecular and Cellular Biology*, vol. 31, no. 10, pp. 2090–2099, 2011.

[7] A. Sanyal, B. R. Lajoie, G. Jain, and J. Dekker, "The long-range interaction landscape of gene promoters," *Nature*, vol. 489, no. 7414, pp. 109–113, 2012.

[8] K. Dahlman-Wright, Y. Qiao, P. Jonsson, J. Gustafsson, C. Williams, and C. Zhao, "Interplay between AP-1 and estrogen receptor α in regulating gene expression and proliferation networks in breast cancer cells," *Carcinogenesis*, vol. 33, no. 9, pp. 1684–1691, 2012.

[9] S. Lalonde, D. W. Ehrhardt, D. Loqué, J. Chen, S. Y. Rhee, and W. B. Frommer, "Molecular and cellular approaches for the detection of protein-protein interactions: latest techniques and current limitations," *Plant Journal*, vol. 53, no. 4, pp. 610–635, 2008.

[10] P. Ulrichts, I. Lemmens, D. Lavens, R. Beyaert, and J. Tavernier, "MAPPIT (Mammalian protein-protein interaction trap) analysis of early steps in toll-like receptor signalling," *Methods in Molecular Biology*, vol. 517, pp. 133–144, 2009.

[11] D. Li, W. Liu, Z. Liu et al., "PRINCESS, a protein interaction confidence evaluation system with multiple data sources," *Molecular and Cellular Proteomics*, vol. 7, no. 6, pp. 1043–1052, 2008.

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

[13] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg, "The database of interacting proteins: 2004 update," *Nucleic Acids Research*, vol. 32, pp. D449–D451, 2004.

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

[15] C. Alfarano, C. E. Andrade, K. Anthony et al., "The Biomolecular Interaction Network Database and related tools 2005 update," *Nucleic Acids Research*, vol. 33, pp. D418–D424, 2005.

[16] C. Stark, B. J. Breitkreutz, A. Chatr-Aryamontri et al., "The BioGRID interaction database: 2011 update," *Nucleic Acids Research*, vol. 39, pp. D698–D704, 2011.

[17] S. Kerrien, B. Aranda, L. Breuza et al., "The IntAct molecular interaction database in 2012," *Nucleic Acids Research*, vol. 40, no. 1, pp. D841–D846, 2012.

[18] L. Licata, L. Briganti, D. Peluso et al., "MINT, the molecular interaction database: 2012 update," *Nucleic Acids Research*, vol. 40, no. 1, pp. D857–D861, 2012.

[19] R. Goel, H. C. Harsha, A. Pandey, and T. S. K. Prasad, "Human protein reference database and human proteinpedia as resources for phosphoproteome analysis," *Molecular BioSystems*, vol. 8, no. 2, pp. 453–463, 2012.

[20] A. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: a network-based approach to human disease," *Nature Reviews Genetics*, vol. 12, no. 1, pp. 56–68, 2011.

[21] M. Wu, X. Li, H. N. Chua, C. Kwoh, and S. Ng, "Integrating diverse biological and computational sources for reliable protein-protein interactions," *BMC Bioinformatics*, vol. 11, article S8, no. 7, 2010.

[22] C. Von Mering, R. Krause, B. Snel et al., "Comparative assessment of large-scale data sets of protein-protein interactions," *Nature*, vol. 417, no. 6887, pp. 399–403, 2002.

[23] R. Gentleman and W. Huber, "Making the most of high-throughput protein-interaction data," *Genome Biology*, vol. 8, no. 10, article 112, 2007.

[24] Y. Qi, Z. Bar-Joseph, and J. Klein-Seetharaman, "Evaluation of different biological data and computational classification methods for use in protein interaction prediction," *Proteins*, vol. 63, no. 3, pp. 490–500, 2006.

[25] K. Tarassov, V. Messier, C. R. Landry et al., "An in vivo map of the yeast protein interactome," *Science*, vol. 320, no. 5882, pp. 1465–1470, 2008.

[26] L. J. Lu, Y. Xia, A. Paccanaro, H. Yu, and M. Gerstein, "Assessing the limits of genomic data integration for predicting protein networks," *Genome Research*, vol. 15, no. 7, pp. 945–953, 2005.

[27] P. Kirk, J. E. Griffin, R. S. Savage, Z. Ghahramani, and D. L. Wild, "Bayesian correlated clustering to integrate multiple datasets," *Bioinformatics*, vol. 28, no. 24, pp. 3290–3297, 2012.

[28] X. Lin and X. W. Chen, "Heterogeneous data integration by tree-augmented naïve Bayes for protein-protein interactions prediction," *Proteomics*, vol. 13, no. 2, pp. 261–268, 2013.

[29] R. Jansen, H. Yu, D. Greenbaum et al., "A Bayesian networks approach for predicting protein-protein interactions from genomic data," *Science*, vol. 302, no. 5644, pp. 449–453, 2003.

[30] E. Sprinzak, Y. Altuvia, and H. Margalit, "Characterization and prediction of protein-protein interactions within and between complexes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 40, pp. 14718–14723, 2006.

[31] L. V. Zhang, S. L. Wong, O. D. King, and F. P. Roth, "Predicting co-complexed protein pairs using genomic and proteomic data integration," *BMC Bioinformatics*, vol. 5, article 38, 2004.

[32] M. Gerstein, N. Lan, and R. Jansen, "Proteomics. Integrating interactomes," *Science*, vol. 295, no. 5553, pp. 284–287, 2002.

[33] A. Ben-Hur and W. S. Noble, "Kernel methods for predicting protein-protein interactions," *Bioinformatics*, vol. 21, no. 1, pp. i38–i46, 2005.

[34] D. R. Rhodes, S. A. Tomlins, S. Varambally et al., "Probabilistic model of the human protein-protein interaction network," *Nature Biotechnology*, vol. 23, no. 8, pp. 951–959, 2005.

[35] X. W. Chen and M. Liu, "Prediction of protein-protein interactions using random decision forest framework," *Bioinformatics*, vol. 21, no. 24, pp. 4394–4400, 2005.

[36] Y. Xu, W. Hu, Z. Chang et al., "Prediction of human protein-protein interaction by a mixed Bayesian model and its application to exploring underlying cancer-related pathway crosstalk," *Journal of the Royal Society Interface*, vol. 8, no. 57, pp. 555–567, 2011.

[37] K. Xia, D. Dong, and J. J. Han, "IntNetDB v1.0: an integrated protein-protein interaction network database generated by a probabilistic model," *BMC Bioinformatics*, vol. 7, article 508, 2006.

[38] H. Ge, Z. Liu, G. M. Church, and M. Vidal, "Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae," *Nature Genetics*, vol. 29, no. 4, pp. 482–486, 2001.

[39] A. Acland, R. Agarwala, T. Barrett et al., "Database resources of the National Center for Biotechnology Information," *Nucleic Acids Research*, vol. 42, no. D1, pp. D7–D17, 2014.

[40] A. Gavin, M. Bösche, R. Krause et al., "Functional organization of the yeast proteome by systematic analysis of protein complexes," *Nature*, vol. 415, no. 6868, pp. 141–147, 2002.

[41] Y. Ho, A. Gruhler, and A. Heilbut, "Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry," *Nature*, vol. 415, no. 6868, pp. 180–183, 2002.

[42] J. A. Brown, G. Sherlock, C. L. Myers et al., "Global analysis of gene function in yeast by quantitative phenotypic profiling," *Molecular Systems Biology*, vol. 2, p. 2006.0001, 2006.

[43] A. M. Dudley, D. M. Janse, A. Tanay, R. Shamir, and G. M. Church, "A global view of pleiotropy and phenotypically derived gene function in yeast.," *Molecular Systems Biology*, vol. 1, p. 2005.0001, 2005.

[44] A. H. Y. Tong, M. Evangelista, A. B. Parsons et al., "Systematic genetic analysis with ordered arrays of yeast deletion mutants," *Science*, vol. 294, no. 5550, pp. 2364–2368, 2001.

[45] A. H. Tong, G. Lesage, G. D. Bader et al., "Global mapping of the yeast genetic interaction network," *Science*, vol. 303, no. 5659, pp. 808–813, 2004.

[46] T. Shlomi, D. Segal, E. Ruppin, and R. Sharan, "QPath: a method for querying pathways in a protein-protein interaction network," *BMC Bioinformatics*, vol. 7, article 199, 2006.

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

[48] J. Chen, B. J. Aronow, and A. G. Jegga, "Disease candidate gene identification and prioritization using protein interaction networks," *BMC Bioinformatics*, vol. 10, article 73, 2009.

[49] L. Sam, Y. Liu, J. Li, C. Friedman, and Y. A. Lussier, "Discovery of protein interaction networks shared by diseases," in *Proceedings of the Pacific Symposium on Biocomputing (PSB '07)*, pp. 76–87, January 2007.

[50] H. Goehler, M. Lalowski, U. Stelzl et al., "A protein interaction network links GIT1, an enhancer of huntingtin aggregation, to Huntington's disease," *Molecular Cell*, vol. 15, no. 6, pp. 853–865, 2004.

[51] H. Ruffner, A. Bauer, and T. Bouwmeester, "Human protein-protein interaction networks and the value for drug discovery," *Drug Discovery Today*, vol. 12, no. 17-18, pp. 709–716, 2007.

[52] V. Neduva, R. Linding, I. Su-Angrand et al., "Systematic discovery of new recognition peptides mediating protein interaction networks," *PLoS Biology*, vol. 3, no. 12, article e405, 2005.

[53] Z. Lubovac, J. Gamalielsson, and B. Olsson, "Combining functional and topological properties to identify core modules in protein interaction networks," *Proteins: Structure, Function and Genetics*, vol. 64, no. 4, pp. 948–959, 2006.

[54] K. Lage, E. O. Karlberg, Z. M. Størling et al., "A human phenome-interactome network of protein complexes implicated in genetic disorders," *Nature Biotechnology*, vol. 25, no. 3, pp. 309–316, 2007.

[55] M. Oti, B. Snel, M. A. Huynen, and H. G. Brunner, "Predicting disease genes using protein-protein interactions," *Journal of Medical Genetics*, vol. 43, no. 8, pp. 691–698, 2006.

[56] M. Krauthammer, C. A. Kaufmann, T. C. Gilliam, and A. Rzhetsky, "Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's desease," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 42, pp. 15148–15153, 2004.

[57] L. Franke, H. van Bakel, L. Fokkens, E. D. de Jong, M. Egmont-Petersen, and C. Wijmenga, "Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes," *The American Journal of Human Genetics*, vol. 78, no. 6, pp. 1011–1025, 2006.

[58] I. Iossifov, T. Zheng, M. Baron, T. C. Gilliam, and A. Rzhetsky, "Genetic-linkage mapping of complex hereditary disorders to a whole-genome molecular-interaction network," *Genome Research*, vol. 18, no. 7, pp. 1150–1162, 2008.

[59] M. Sekine, H. Nagata, S. Tsuji et al., "Localization of a novel susceptibility gene for familial ovarian cancer to chromosome 3p22-p25," *Human Molecular Genetics*, vol. 10, no. 13, pp. 1421–1429, 2001.

[60] D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova, "Entrez Gene: gene-centered information at NCBI," *Nucleic Acids Research*, vol. 39, no. 1, pp. D52–D57, 2011.

[61] M. A. van Driel, J. Bruggeman, G. Vriend, H. G. Brunner, and J. A. M. Leunissen, "A text-mining analysis of the human phenome," *European Journal of Human Genetics*, vol. 14, no. 5, pp. 535–542, 2006.

[62] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick, "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders," *Nucleic Acids Research*, vol. 33, pp. D514–D517, 2005.

[63] C. E. Lipscomb, "Medical Subject Headings (MeSH)," *Bulletin of the Medical Library Association*, vol. 88, no. 3, pp. 265–266, 2000.

[64] S. Zhang, Z. Chang, Z. Li et al., "Calculating phenotypic similarity between genes using hierarchical structure data based on semantic similarity," *Gene*, vol. 497, no. 1, pp. 58–65, 2012.

[65] 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.

[66] S. Navlakha and C. Kingsford, "The power of protein interaction networks for associating genes with diseases," *Bioinformatics*, vol. 26, no. 8, pp. 1057–1063, 2010.

[67] P. A. Futreal, L. Coin, M. Marshall et al., "A census of human cancer genes," *Nature Reviews Cancer*, vol. 4, no. 3, pp. 177–183, 2004.

[68] T. Santarius, J. Shipley, D. Brewer, M. R. Stratton, and C. S. Cooper, "A census of amplified and overexpressed human cancer genes," *Nature Reviews Cancer*, vol. 10, no. 1, pp. 59–64, 2010.

[69] M. Zhu, L. Gao, X. Li et al., "The analysis of the drug-targets based on the topological properties in the human protein-protein interaction network," *Journal of Drug Targeting*, vol. 17, no. 7, pp. 524–532, 2009.

[70] M. A. Yildirim, K. I. Goh, M. E. Cusick, A. L. Barabási, and M. Vidal, "Drug-target network," *Nature Biotechnology*, vol. 25, no. 10, pp. 1119–1126, 2007.

[71] 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.

[72] M. Wu, Q. Yu, X. Li, J. Zheng, J. Huang, and C. Kwoh, "Benchmarking human protein complexes to investigate drug-related systems and evaluate predicted protein complexes," *PLoS ONE*, vol. 8, no. 2, Article ID e53197, 2013.

[73] G. D. Bader and C. W. V. Hogue, "An automated method for finding molecular complexes in large protein interaction networks.," *BMC Bioinformatics*, vol. 4, no. 1, article 2, 2003.

[74] J. Song and M. Singh, "From hub proteins to hub modules: the relationship between essentiality and centrality in the yeast interactome at different scales of organization," *PLoS Computational Biology*, vol. 9, no. 2, Article ID e1002910, 2013.

[75] C. Zhang and L. Shen, "Functional modules analysis based on protein-protein network analysis in ankylosing spondylitis," *European Review for Medical and Pharmacological Sciences*, vol. 16, no. 13, pp. 1821–1827, 2012.

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

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 589290, 9 pages
http://dx.doi.org/10.1155/2014/589290

Research Article

Informative Gene Selection and Direct Classification of Tumor Based on Chi-Square Test of Pairwise Gene Interactions

Hongyan Zhang,[1,2,3] Lanzhi Li,[1,3] Chao Luo,[2] Congwei Sun,[1,3] Yuan Chen,[1,3] Zhijun Dai,[1,3] and Zheming Yuan[1,3]

[1] *Hunan Provincial Key Laboratory of Crop Germplasm Innovation and Utilization, Changsha 410128, China*
[2] *College of Information Science and Technology, Hunan Agricultural University, Changsha 410128, China*
[3] *Hunan Provincial Key Laboratory for Biology and Control of Plant Diseases and Insect Pests, Changsha 410128, China*

Correspondence should be addressed to Zheming Yuan; zhmyuan@sina.com

Received 28 May 2014; Accepted 10 July 2014; Published 23 July 2014

Academic Editor: Yan Guo

In efforts to discover disease mechanisms and improve clinical diagnosis of tumors, it is useful to mine profiles for informative genes with definite biological meanings and to build robust classifiers with high precision. In this study, we developed a new method for tumor-gene selection, the Chi-square test-based integrated rank gene and direct classifier (χ^2-IRG-DC). First, we obtained the weighted integrated rank of gene importance from chi-square tests of single and pairwise gene interactions. Then, we sequentially introduced the ranked genes and removed redundant genes by using leave-one-out cross-validation of the chi-square test-based Direct Classifier (χ^2-DC) within the training set to obtain informative genes. Finally, we determined the accuracy of independent test data by utilizing the genes obtained above with χ^2-DC. Furthermore, we analyzed the robustness of χ^2-IRG-DC by comparing the generalization performance of different models, the efficiency of different feature-selection methods, and the accuracy of different classifiers. An independent test of ten multiclass tumor gene-expression datasets showed that χ^2-IRG-DC could efficiently control overfitting and had higher generalization performance. The informative genes selected by χ^2-IRG-DC could dramatically improve the independent test precision of other classifiers; meanwhile, the informative genes selected by other feature selection methods also had good performance in χ^2-DC.

1. Introduction

Tumors are the consequences of interactions between multiple genes and the environment. The emergence and rapid development of large-scale gene-expression technology provide an entirely new platform for tumor investigation. Tumor gene-expression data has the following features: high dimensionality, small or relatively small sample size, large differences in sample backgrounds, presence of nonrandom noise (e.g., batch effects), high redundancy, and nonlinearity. Mining of tumor-informative genes with definite biological meanings and building of robust classifiers with high precision are important goals in the context of clinical diagnosis of tumors and discovery of disease mechanisms.

Informative gene selection is a key issue in tumor recognition. Theoretically, there are 2^m possibilities in selecting the optimal informative gene subset from m genes, which is an N-P hard problem. Available high-dimensional feature-selection methods often fall into one of the following three categories: (i) filter methods, which simply rank all genes according to the inherent features of the microarray data, and their algorithm complexities are low. However, redundant phenomena are usually present among the selected informative genes, which may result in low classification precision. Univariate filter methods include t-test [1], correlation coefficient [2], Chi-square statistics [3], information gain [4], relief [5], signal-to-noise ratio [6], Wilcoxon rank sum [7], and entropy [8]. Multivariable filter methods include mRMR [9], correlation-based feature selection [10], and Markov blanket filter [11]; (ii) wrapper methods, which search for an optimal feature set that maximizes the classification performance, defined in terms of an evaluation function

(such as cross-validation accuracy). Their training precision and algorithm complexity are high; consequently, it is easy for over-fitting to occur. Search strategies include sequential forward selection [12], sequential backward selection [12], sequential floating selection [13], particle swarm optimization algorithm [14], genetic algorithm [15], ant colony algorithm [16], and breadth-first search [17]. SVM and ANN are usually used for feature subset evaluation; (iii) embedded methods, which use internal information about the classification model to perform feature selection. These methods include SVM-RFE [18], support vector machine with RBF kernel based on recursive feature elimination (SVM-RBF-RFE) [19], support vector machine and T statistics recursive feature elimination (SVM-T-RFE) [20], and random forest [21].

Classifier is another key issue in tumor recognition. Traditional classification algorithms include Fisher linear discriminator, Naive bayes (NB) [22], K-nearest neighbor (KNN) [23], DT [24], support vector machine (SVM) [18], and artificial neural network (ANN) [25]. There are dominant expressions in parametric models (e.g., Fisher linear discriminator) based on induction inference. The first goal for parametric models is to obtain general rules through training-sample learning, after which these rules are utilized to judge the testing sample. However, this is not the case for nonparametric models (e.g., SVM) based on transduction inference, which predict special testing samples through observation of special training samples, but classifiers needed for training. Training is the major reason for model over-fitting [3]. Therefore, it is important to determine whether it is feasible to develop a direct classifier based on transduction interference that has no demand for training.

In recent years, several methods have been developed to perform both feature-selection and classification for the analysis of microarray data as follows: prediction analysis for microarrays (PAM), based on nearest shrunken centroids [26]; top scoring pair (TSP), based entirely on relative gene expression values [27]; refined TSP algorithms, such as k disjoint Top Scoring Pairs (k-TSP) for binary classification and the HC-TSP, HC-k-TSP for multiclass classification [28]; an extended version of TSP, the top-scoring triplet (TST) [29]; an extended version of TST, top-scoring "N" (TSN) [30]. A remarkable advantage of the TSP family is that they can effectively control experimental system deviations, such as background differences and batch effects between samples. However, TSP, k-TSP, TST, and TSN are only suitable for binary data, and the HC-TSP/HC-TSP calculation process for conversion from multiclass to binary classification is tedious. The gene score Δ_{ij} [27] cannot reflect size differences among samples, and k-TSPs may introduce redundancy and undiscriminating voting weights.

Chi-square-statistic-based top scoring genes (TSG) [31], an improved version of TSP family we proposed before, introduces Chi-square value as the score for each marker set so that the sample size information is fully utilized. TSG proposes a new gene selection method based on joint effects of multiple genes, and the informative genes number is allowed both even and odd. Moreover, TSG gives a new classification method with no demand for training, and it is in a simple unified form for both binary and multiclass cases. In TSG paper, we did not name the classification method alone. Here we called it the chi-square test-based direct classifier (χ^2-DC). To predict the class information for each sample in the test data, χ^2-DC use the selected marker set and calculate the scores of this sample belonging to each class. The predicted class is set to be the one that has the largest score. Although TSG has many merits, it also has the following disadvantages: (i) for $k \geq 3$, in order to find the top scoring k genes (TS_k), all the combined scores between TS_{k-1} and each of remaining gene need to be calculated. It needs a large amount of calculation; (ii) if there are multiple TS_ks with identical maximum Chi-square value, TSG should further calculate the LOOCV accuracy of these TS_ks using the training data and record those TS_ks that yield the highest LOOCV accuracy. If there is still more than one TS_k, the computational complexity will be much higher to find TS_{k+1}; (iii) in TSG, an upper bound B should be set and find TS_B. However, the number of information genes is often less than B. The termination condition of feature selection is not objective enough.

Emphasizing interactions between genes or biological marks is a developing trend in cancer classification and informative gene selection. The TSP family, mRMR, doublets [32], nonlinear integrated selection [33], binary matrix shuffling filter (BMSF) [34], and TSG all take interactions into consideration. In genome-wide association studies, ignorance of interactions between SNPs or genes will cause the loss of inheritability [35]. Therefore, we developed a novel high-dimensional feature-selection algorithm called a Chi-square test-based integrated rank gene and direct classifier (χ^2-IRG-DC), which inherits the advantages of TSG while overcoming the disadvantages documented above in feature selection. First, this algorithm obtains the weighted integrated rank of gene importance on the basis of chi-square tests of single and pairwise gene interactions. Then, the algorithm sequentially forward introduces ranked genes and removes redundant parts using leave-one-out cross validation (LOOCV) of χ^2-DC within the training set to obtain the final informative gene subset of tumor.

A large number of feature-selection methods and classifiers currently exist. Informative gene subsets obtained by different feature-selection methods are very minute overlap [36]. However, different models combined with a certain feature-selection method and a suitable classifier can get a close prediction precision [37]. It is difficult to determine which feature-selection method is better. Therefore, evaluation of the robustness of feature-selection methods deserves more attention [32]. In this paper, we analyzed the robustness of χ^2-IRG-DC by comparing the generalization performance of different models, the efficiency of different feature-selection methods, and the precision of different classifiers.

2. Data and Methods

2.1. Data. Because nine common binary-class tumor-genomics datasets [28] did not offer independent test sets, we simply selected ten multiclass tumor-genomics datasets with

independent test sets (Table 1) for analysis in this study. It should be noted that the method proposed in this paper could also be applied to binary-class datasets.

2.2. Weighted Integrated Rank of Genes. Assume the training dataset has p markers and n samples. The data can be denoted as (y_i, x_{ij}) $(i = 1, \ldots, n; j = 1, \ldots, p)$. x_{ij} represents the expression value of the jth marker in the ith sample; y_i represents the label of ith sample, where $y_i \in C = \{C_1, \ldots, C_m\}$, the set of possible labels; m stands for the total number of labels in the data.

(1) *Chi-Square Values of Single Genes*. For any single gene G_j, $\overline{x}_{\cdot j}$ denotes the mean expression value of all samples. Sf_{k1} and Sf_{k2} $(k = 1, \ldots, m)$ represent the frequency counts of samples in class C_k when $x_{ij} > \overline{x}_{\cdot j}$ and $x_{ij} < \overline{x}_{\cdot j}$, respectively. These frequencies can be presented as an $m \times 2$ contingency table, as shown in Table 2. Record the frequency counts of samples in class C_k as Sf_{k3} When x_{ij} equals $\overline{x}_{\cdot j}$ in class C_k, then both Sf_{k1} and Sf_{k2} should be incremented by $0.5 * Sf_{k3}$ separately; thus, the chi-square value χ_j^2 of gene G_j can be calculated according to (1)

$$\chi_j^2 = SN\left(\sum_{k=1}^{m}\sum_{q=1}^{2}\frac{Sf_{kq}^2}{Sn_k ST_q} - 1\right). \quad (1)$$

(2) *Chi-Square Values of Pairwise Genes*. For any two genes G_j and G_l $(j = 1, \ldots, p; l = 1, \ldots, p; l \neq j)$, Pf_{k1} and Pf_{k2} $(k = 1, \ldots, m)$ represent the frequency counts of samples in class C_k when $x_{ij} > x_{il}$ and $x_{ij} < x_{il}$, respectively. x_{ij} and x_{il} are expression values of the ith sample in genes G_j and G_l, respectively. These frequencies can be presented as an $m \times 2$ contingency table (Table 3). Record the frequency counts of samples in class C_k as Pf_{k3} When x_{ij} equals x_{il} in class C_k, then both Pf_{k1} and Pf_{k2} should be incremented by $0.5 * Pf_{k3}$ separately. The Chi-square value $\chi_{j,l}^2$ of pairwise genes (G_j, G_l) can be calculated according to (2)

$$\chi_{j,l}^2 = PN\left(\sum_{k=1}^{m}\sum_{q=1}^{2}\frac{Pf_{kq}^2}{Pn_k PT_q} - 1\right). \quad (2)$$

(3) *Rank Genes according to Integrated Weighted Score*. Judging whether a gene is important not only should take main effect of gene into account, but also consider the interaction between it and other genes. Therefore, we integrated the Chi-square value of single gene and the Chi-square values of pairwise genes to define an integrated weighted score of each gene S_j as shown in (3). S_j is the integrated weighted score of gene G_j $(j = 1, \ldots, p)$, χ_j^2 is the chi-square value of single gene G_j, and $\chi_{j,l}^2$ is the chi-square value of pairwise genes G_j and G_l $(l = 1, \ldots, p; l \neq j)$. Genes are ranked by the integrated weighted score S_j to become a descending-range sequence. Consider

$$S_j = \chi_j^2 + \sum_{l=1}^{p}\left(\frac{\chi_j^2}{\chi_j^2 + \chi_l^2} \times \chi_{j,l}^2\right) \quad (3)$$

make an ordered list Θ of all the genes G_j in accordance with the descending values of the scores S_j.

2.3. Chi-Square Test-Based Direct Classifier (χ^2-DC). When the training set has n samples and m labels, with r $(r \geq 2)$ selected genes, there are $r \times (r-1)/2$ contingency tables included, each of which has m rows and 2 columns (Table 2). If the testing sample belongs to class C_k $(k = 1, \ldots, m)$, $r \times (r-1)/2$ chi-square values of pairwise genes with $n + 1$ samples (i.e., including n training samples and a testing sample) can be worked out. The sum of $r \times (r-1)/2$ chi-square values was set as $\chi_{(C_k)}^2$ $(k = 1, \ldots, m)$. We assign the test sample to the class with the largest chi-square value: class of testing sample = arg $\max_{k=1,\ldots,m}\chi_{(C_k)}^2$ [31].

2.4. Introduce Ranked Genes Sequentially and Remove Redundant Parts to Obtain Informative Genes. Take the top two genes from the ordered list Θ and extract their expression values from the training dataset to form the initial training set. Next, compute the LOOCV accuracy of the initial training data based on χ^2-DC and denote it as LOOCV_2. Record m chi-square values $\chi_{(C_1)}^2, \chi_{(C_2)}^2, \ldots, \chi_{(C_m)}^2$ of every sample taken as a measured sample. Finally, introduce parameter h, as shown in (4)

$$h = \sum_{k=1}^{m}\frac{\chi_{(C_t)}^2 - \chi_{(C_k)}^2}{\chi_{(C_t)}^2} \quad k \neq t, \quad (4)$$

where C_t is the true label of the measured sample. The average value of every training sample is denoted as $\overline{h_2}$.

Now import the third gene from the ordered list Θ and extract its expression values from the training dataset to update the initial training set. Following the steps documented above, obtain LOOCV_3 and $\overline{h_3}$ of the updated training set. If $\text{LOOCV}_3 > \text{LOOCV}_2$, or $\text{LOOCV}_3 = \text{LOOCV}_2$ and $\overline{h_3} > \overline{h_2}$, the third gene is selected as an informative gene; Otherwise, it is deemed as a redundant gene.

Similarly, informative gene subsets will be obtained by sequentially introducing the top 2% genes from the ordered list Θ.

2.5. Independent Prediction. With the informative gene subsets, independent prediction based on χ^2-DC was conducted individually on the testing sample to obtain the test accuracy.

2.6. Models Used for Comparison. In this paper, a model is considered as a combination of a specific feature-selection method and a specific classifier. Some feature-selection methods are also classifiers (HC-TSP, HC-k-TSP, TSG, DT, PAM, etc.). We selected mRMR-SVM, SVM-RFE-SVM, HC-k-TSP and TSG as comparative models for χ^2-IRG-DC; NB, KNN, and SVM as the comparative classifiers of χ^2-DC; mRMR, SVM-RFE, HC-k-TSP and TSG as the comparative feature-selection approaches of χ^2-IRG-DC.

mRMR conducts minimum redundancy maximum relevance feature selection. Mutual information difference (MID) and mutual information quotient (MIQ) are two versions of mRMR. MIQ was better than MID in general [9], so the evaluation criterion in this paper is mRMR-MIQ. SVM-RFE is a simple and efficient algorithm which conducts gene selection

Table 1: Multiclass gene-expression datasets.

Dataset	Platform	No. of classes	No. of genes	No. of samples in training	No. of samples in test	Source
Leuk1	Affy	3	7,129	38	34	[6]
Lung1	Affy	3	7,129	64	32	[43]
Leuk2	Affy	3	12,582	57	15	[44]
SRBCT	cDNA	4	2,308	63	20	[45]
Breast	Affy	5	9,216	54	30	[46]
Lung2	Affy	5	12,600	136	67	[47]
DLBCL	cDNA	6	4,026	58	30	[48]
Leukemia3	Affy	7	12,558	215	112	[49]
Cancers	Affy	11	12,533	100	74	[50]
GCM	Affy	14	16,063	144	46	[51]

Table 2: Frequency counts of samples in each class for single genes.

Class	$x_{ij} > \overline{x}_{\bullet j}$	$x_{ij} < \overline{x}_{\bullet j}$	Total
C_1	Sf_{11}	Sf_{12}	$Sn_1 = Sf_{11} + Sf_{12}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
C_m	Sf_{m1}	Sf_{m2}	$Sn_m = Sf_{m1} + Sf_{m2}$
Total	$ST_1 = \sum_{k=1}^{m} Sf_{k1}$	$ST_2 = \sum_{k=1}^{m} Sf_{k2}$	$SN = \sum_{k=1}^{m} Sn_k$

Table 3: Frequency counts of samples in each class for pairwise genes.

Class	$x_{ij} > x_{il}$	$x_{ij} < x_{il}$	Total
C_1	Pf_{11}	Pf_{12}	$Pn_1 = Pf_{11} + Pf_{12}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
C_m	Pf_{m1}	Pf_{m2}	$Pn_m = Pf_{m1} + Pf_{m2}$
Total	$PT_1 = \sum_{k=1}^{m} Pf_{k1}$	$PT_2 = \sum_{k=1}^{m} Pf_{k2}$	$PN = \sum_{k=1}^{m} Pn_k$

in a backward elimination procedure. The mRMR and SVM-RFE have been widely applied in analyzing high-dimensional biological data. They only provide a list of ranked genes; a classification algorithm needs to be used to choose the set of variables that minimize cross validation error. In this paper, SVM was selected as the classification algorithm, and our SVM implementation is based on LIBSVM which supports 1-versus-1 multiclass classification. For SVM-RFE-SVM and mRMR-SVM models, informative genes were selected by the following methods: (i) rank the genes separately by mRMR or SVM-RFE; (ii) select the top genes from 1 to s, which is equal to approximately 2% of the total gene number, and conduct 10-fold cross-validation (CV10) for the training sets based on SVM. Accuracy was denoted as $\mathrm{CV10}_w$ ($w = 1, \ldots, s$); (iii) with the highest CV10 accuracy, the genes were selected as informative genes.

3. Results and Discussion

3.1. Comparison of Independent Test Accuracy and the Number of Informative Genes Used in Different Models. In order to evaluate the performance of model in this study, we used the eight different models to perform independent test on ten multiclass datasets. The test accuracy and informative gene number are presented in Table 4. In this case, the classification accuracy of each dataset is the ratio of the number of the correctly classified samples to the total number of samples in that dataset. The best model based on average accuracy of the ten multiclass datasets used in this study is χ^2-IRG-DC (90.81%), followed by TSG (89.2%), PAM (88.5%), SVM-RFE-SVM (86.72%) and HC-k-TSP (85.12%). We do not consider these differences in accuracy as noteworthy and conclude that all five methods perform similarly. However, in terms of efficiency, decision rule and the number of informative genes, one can argue that the χ^2-IRG-DC method is superior. Recall that the χ^2-IRG-DC, TSG and PAM have easy interpretation and can directly handle multiclass case, but HC-k-TSP and SVM-RFE-SVM need a tedious process to covert multiclass case into binclass case. For the ten multiclass datasets, χ^2-IRG-DC selected 37.2 (range, 20–64 in ten datasets) informative genes on average. It clearly uses less number of genes than PAM (1638.8) and TSG (51). Moreover, the algorithm complexities of χ^2-IRG-DC is far less than TSG. χ^2-IRG-DC ranked all genes according to integrated weighted score firstly and sequentially introduced the ranked genes based on LOOCV accuracy of training data. In fact, χ^2-IRG-DC is a hybrid filter-wrapper models that take advantage of the simplicity of the filter approach for initial gene screening and then make use of the wrapper approach to optimize classification accuracy in final gene selection [38].

3.2. Robustness Analysis—Evaluating Generalization Performance of Different Models. As shown in Table 4, the five models (mRMR-SVM, SVM-RFE-SVM, HC-k-TSP, TSG, and χ^2-IRG-DC) exhibited high independent test accuracy and similar informative gene numbers. We further compared the LOOCV accuracy for the training data and the independent test accuracy for the test data from these four models. The results are shown in Figures 1, 2, 3, 4, and 5. Obviously, overfitting occurred in all five models. Among them, χ^2-IRG-DC had higher generalization performance. The test accuracy of mRMR-SVM and SVM-RFE-SVM was no greater than their training accuracy for all ten datasets. However, the test accuracy of χ^2-IRG-DC was superior to the training accuracy

Table 4: Independent test accuracy and informative gene number used indifferent models (in parentheses) for multiclass gene-expression datasets.

Model	Leuk1	Lung1	Leuk2	SRBCT	Breast	Lung2	DLBCL	Leuk3	Cancers	GCM	Aver ± std
HC-TSP*	**97.06**	71.88	80	95	66.67	83.58	83.33	77.68	74.32	52.17	78.17 ± 13.17
	(4)	(4)	(4)	(6)	(8)	(8)	(10)	(12)	(20)	(26)	(10.2)
HC-K-TSP*	**97.06**	78.13	**100**	**100**	66.67	94.03	83.33	82.14	82.43	**67.39**	85.12 ± 12.42
	(36)	(20)	(24)	(30)	(24)	(28)	(46)	(64)	(128)	(134)	(53.4)
DT*	85.29	78.13	80	75	73.33	88.06	86.67	75.89	68.92	52.17	76.35 ± 10.49
	(2)	(4)	(2)	(3)	(4)	(5)	(5)	(16)	(10)	(18)	(6.9)
PAM*	**97.06**	78.13	93.33	95	**93.33**	**100**	90	**93.75**	87.84	56.52	88.5 ± 12.71
	(44)	(13)	(62)	(285)	(4,822)	(614)	(3,949)	(3,338)	(2,008)	(1,253)	(1,638.8)
mRMR-SVM	76.47	78.13	100.00	75.00	96.67	95.52	96.67	91.96	71.62	45.65	82.77 ± 16.85
	(7)	(13)	(19)	(9)	(97)	(120)	(16)	(119)	(89)	(57)	(54.6)
SVM-RFE-SVM	85.29	78.13	93.33	95.00	90.00	88.06	90.00	91.07	93.24	63.04	86.72 ± 9.62
	(5)	(9)	(8)	(3)	(7)	(9)	(13)	(35)	(29)	(199)	(31.7)
TSG	97.06	81.25	100	100	86.67	95.52	93.33	91.07	79.73	67.39	89.20 ± 10.5
	(6)	(20)	(44)	(13)	(63)	(60)	(16)	(95)	(81)	(112)	(51)
χ^2-IRG-DC	**97.06**	84.38	**100**	**100**	90	97.01	93.33	**93.75**	85.14	**67.39**	**90.81** ± 9.91
	(29)	(23)	(20)	(23)	(31)	(52)	(37)	(46)	(47)	(64)	(37.2)

*Results reported in [28].

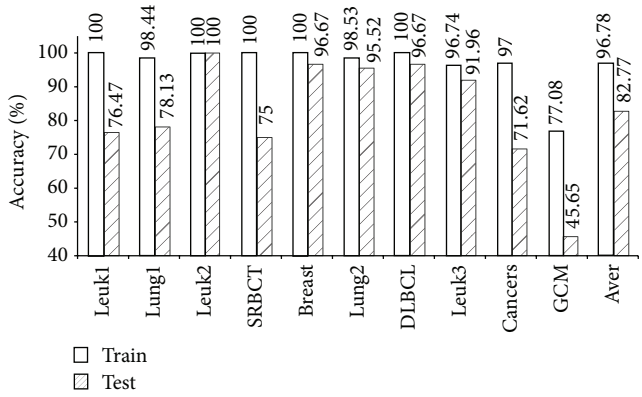


Figure 1: Accuracy of mRMR-SVM for training and test data.

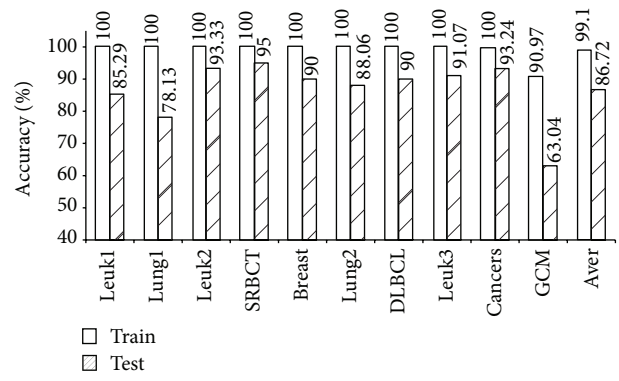


Figure 2: Accuracy of SVM-RFE-SVM for training and test data.

for the Leuk2, Lung2, and Leuk3 datasets, and the test accuracy of TSG was superior to the training accuracy for the Lung1, Lung2, Leuk2, and Leuk3 datasets. For another direct classifier, HC-k-TSP, the test accuracy was also higher than the training accuracy for the SRBCT and cancers datasets. These results indicated that the special direct classification algorithm of χ^2-IRG-DC, TSG and HC-k-TSP can effectively control over-fitting, and exhibiting a better generalization performance.

3.3. Robustness Analysis—Evaluating Different Feature-Selection Methods. As shown in Table 5, with the informative genes selected by the five feature-selection methods, the classification performances of NB and KNN were significantly improved. However, the performance of SVM was improved only with the genes selected by our method, χ^2-IRG-DC. This observation indicated, on the one hand, that SVM is not sensitive to feature dimensions [39], and on the other hand, that χ^2-IRG-DC was more robust than the other four feature-selection methods.

With the genes selected by χ^2-IRG-DC, four classifiers (NB, KNN, SVM, and χ^2-DC) performed very well, with average accuracies of 84.23%, 85.54%, 89.54%, and 90.81%, respectively, across ten datasets; the overall average accuracy was 87.53%. Similarly, we calculated the overall average accuracy of other feature-selection methods: 87.53% (χ^2-IRG-DC) > 85.99% (HC-k-TSP) > 84.45% (TSG) > 81.93% (SVM-RFE) > 80.16% (mRMR), once again confirming the robustness and effectiveness of χ^2-IRG-DC.

3.4. Robustness Analysis—Comparison of Classifiers. The overall average accuracies of the four classifiers with informative genes selected by five feature-selection methods across

Table 5: Test accuracy of different classifiers with informative genes selected by different feature-selection methods.

Classifier	Feature-selection method	Leuk1	Lung1	Leuk2	SRBCT	Breast	Lung2	DLBCL	Leuk3	Cancers	GCM	Aver-F
NB	ALL*	85.29	81.25	100.00	60.00	66.67	88.06	86.67	32.14	79.73	52.17	73.20
	χ^2-IRG-DC	97.06	81.25	100.00	85.00	86.67	92.54	96.67	59.82	82.43	60.87	84.23
	mRMR	79.41	68.75	100.00	90.00	93.33	97.01	96.67	74.11	70.27	45.65	81.52
	SVM-RFE	67.65	81.25	80.00	95.00	80.00	89.55	90.00	95.00	77.03	63.04	81.85
	HC-K-TSP	91.18	81.25	100.00	80.00	80.00	95.52	86.67	100.00	77.03	65.22	85.69
	TSG	91.18	84.38	93.33	100	86.67	94.03	100	51.79	71.62	65.22	83.82
	Aver-C†	85.30	79.38	94.67	90.00	85.33	93.73	94	76.14	75.68	60.00	**83.42**
KNN	ALL*	67.65	75.00	86.67	70.00‡	63.33	88.06	93.33	75.89	64.86	34.78	71.96
	χ^2-IRG-DC	97.06	71.88	86.67	100.00	86.67	85.07	96.67	87.50	85.14	58.70	85.54
	mRMR	70.59	68.75	80.00	80.00	96.67	86.57	100.00	91.07	54.05	36.96	76.47
	SVM-RFE	76.47	68.75	86.67	100.00	90.00	86.57	90.00	91.96	58.11	45.65	79.42
	HC-K-TSP	88.24	87.50	86.67	85.00	83.33	94.03	93.33	88.39	64.86	52.17	82.35
	TSG	91.18	75	93.33	100	80	88.06	96.67	86.6	74.32	39.13	82.43
	Aver-C†	84.71	74.38	86.67	93.00	87.33	88.06	95.33	89.10	67.30	46.52	**81.24**
SVM	ALL*	79.41	87.50	100.00	100.00	83.33	97.01	100.00	84.82	83.78	65.22	88.11
	χ^2-IRG-DC	97.06	87.50	93.33	100.00	93.33	92.54	96.67	86.61	91.89	56.52	89.54
	mRMR	76.47	78.13	100.00	75.00	96.67	95.52	96.67	91.96	71.62	45.65	82.77
	SVM-RFE	85.29	78.13	93.33	95.00	90.00	88.06	90.00	91.07	93.24	63.04	86.72
	HC-K-TSP	85.29	84.38	100.00	90.00	86.67	98.51	96.67	94.64	82.43	60.87	87.95
	TSG	91.18	81.25	93.33	80	80	94.03	100	80.36	68.92	54.35	82.34
	Aver-C†	87.06	81.88	96.00	88.00	89.33	93.73	96.00	88.93	81.62	56.09	**85.86**
χ^2-DC	χ^2-IRG-DC	97.06	84.38	100.00	100.00	90.00	97.01	93.33	93.75	85.14	67.39	90.81
	mRMR	82.35	65.63	100.00	90.00	90.00	95.52	70.00	96.43	60.81	47.83	79.86
	SVM-RFE	79.41	56.25	66.67	85.00	76.67	92.54	80.00	96.43	94.59	69.57	79.71
	HC-K-TSP	97.06	84.38	100.00	95.00	76.67	97.01	93.33	88.39	78.38	69.57	87.98
	TSG	97.06	81.25	100	100	86.67	95.52	93.33	91.07	79.73	67.39	89.20
	Aver-C†	90.59	74.38	93.33	94.00	84.00	95.52	86.00	93.21	79.73	64.35	**85.51**

*Results reported in [28]; ‡30 in original paper, whereas the actual number was 70 after validation; †Aver-C was the average accuracy of a classifier with informative genes selected by four feature-selection methods.

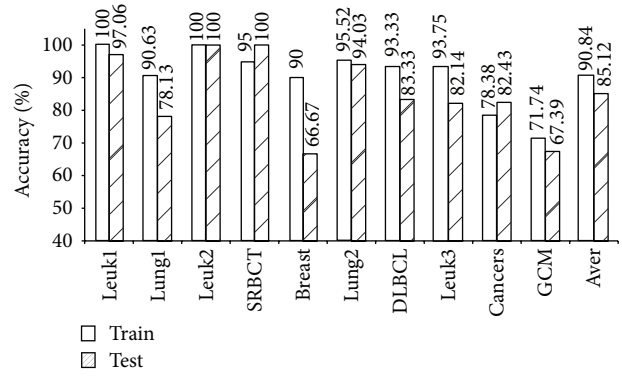


Figure 3: Accuracy of HC-k-TSP for training and test data.

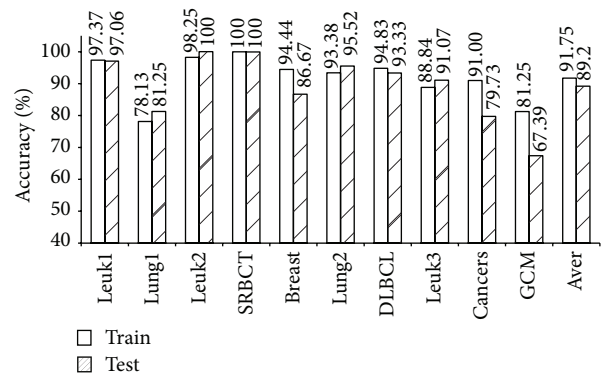


Figure 4: Accuracy of TSG for training and test data.

ten datasets are highlighted in bold in Table 5. The order is as follows: 85.86% (SVM) > 85.51% (χ^2-DC) > 83.42% (NB) > 81.24% (KNN). This result revealed that SVM is an excellent classifier; at the same time, the χ^2-DC classifier also performed well.

4. Conclusion

Informative gene subsets selected by different feature-selection methods often differ greatly. As we can see, genes number selected by the three different models (mRMRSVM, SVM-RFE-SVM) in are listed in Table S1. The numbers

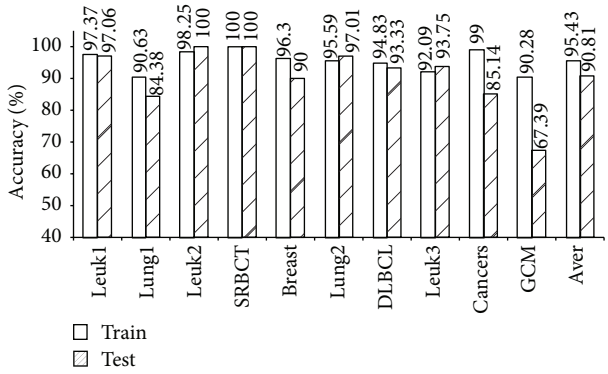


Figure 5: Accuracy of χ^2-IRG-DC for training and test data.

of overlapped gene selected by different models are listed in Table S2. Results showed that there are few overlaps of genes selected by the three models (see supplementary Tables S1 and S2 in supplementary materials available online at http://dx.doi.org/10.1155/2014/589290). However, different models combined with a certain feature-selection method and a suitable classifier can get a close prediction precision. Evaluations of robustness of feature-selection methods and classifiers should include the following aspects: (i) models should have good generalization performance, that is, a model should not only have high accuracy in training sets, but should also have high and stable test accuracy across many datasets (average accuracy ± standard deviation); (ii) with informative genes selected by an excellent feature-selection method, should improve varies classifiers performance; (iii) similarly, a good classifier should perform well with different informative genes selected by different excellent feature-selection approaches.

The results of this study illustrate that pairwise interaction is the fundamental type of interaction. Theoretically, the complexity of the algorithm could be controlled within $O(n^2)$ with pairwise interactions. When three or more genes connect to each other, the complex combination of three or more genes could be represented by the pairwise interactions. Based on this assumption, this paper proposes a novel algorithm, χ^2-IRG-DC, used for informative gene selection and classification based on chi-square tests of pairwise gene interactions. The proposed method was applied to ten multiclass gene-expression datasets; the independent test accuracy and generalization performance were obviously better than those of mainstream comparative algorithms. The informative genes selected by χ^2-IRG-DC were able to significantly improve the independent test accuracy of other classifiers. The average extent of test accuracy raised by χ^2-IRG-DC is superior to those of comparable feature-selection algorithms. Meanwhile, informative genes selected by other feature-selection methods also performed well on χ^2-DC.

Currently, integrated analysis of multisource heterogeneous data is a key challenge in cancer classification and informative gene selection. This includes the integration of repeated measurements from different assays for the same disease on the same platform [40], as well as the integration of gene chips, protein mass spectrometry, DNA methylation, and GWAS-SNP data collected on different platforms for the study of the same disease [41], and so forth. In future, we will apply χ^2-IRG-DC to the integrated analysis of multisource heterogeneous data. Combining this method with the GO database, biological pathways, disease databases, and relevant literature, we will conduct a further assessment of the relevance of the biological functions of selected informative genes to the mechanisms of disease [42].

Conflict of Interests

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

Authors' Contribution

Hongyan Zhang and Lanzhi Li contributed equally to this work. Hongyan Zhang and Lanzhi Li are joint senior authors on this work.

Acknowledgments

The research was supported by a Grant from the National Natural Science Foundation of China (no. 61300130), the Doctoral Foundation of the Ministry of Education of China (no. 20124320110002), the Postdoctoral Science Foundation of Hunan Province (no. 2012RS4039), and the Science Research Foundation of the National Science and Technology Major Project (no. 2012BAD35B05).

References

[1] I. Hedenfalk, D. Duggan, Y. D. Chen et al., "Gene-expression profiles in hereditary breast cancer," *New England Journal of Medicine*, vol. 344, no. 8, pp. 539–548, 2001.

[2] V. R. Lyer, M. B. Eisen, D. T. Ross et al., "The transcriptional program in the response of human fibroblasts to serum," *Science*, vol. 283, pp. 83–87, 1999.

[3] X. Jin, A. Xu, R. Bie, and P. Guo, "Machine learning techniques and chi-square feature selection for cancer classification using SAGE gene expression profiles," in *Data Mining for Biomedical Applications*, vol. 3916 of *Lecture Notes in Computer Science*, pp. 106–115, Springer, Berlin, Germany, 2006.

[4] M. Dash and H. Liu, "Feature selection for classification," *Intelligent Data Analysis*, vol. 1, no. 1–4, pp. 131–156, 1997.

[5] K. Kenji and A. R. Larry, "The feature selection problem: traditional methods and a new algorithm," in *Proceedings of the 10th National Conference on Artificial Intelligence*, W. Swartout, Ed., pp. 129–134, AAAI Press/The MIT Press, Cambridge, Mass, USA, 1992.

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

[7] Z. Fang, R. Du, and X. Cui, "Uniform approximation is more appropriate for wilcoxon rank-sum test in gene set analysis," *PLoS ONE*, vol. 7, no. 2, Article ID e31505, 2012.

[8] S. Zhu, D. Wang, K. Yu, T. Li, and Y. Gong, "Feature selection for gene expression using model-based entropy," *IEEE Transactions*

on Computational Biology and Bioinformatics, vol. 7, no. 1, pp. 25–36, 2010.

[9] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[10] Y. Wang, I. V. Tetko, M. A. Hall et al., "Gene selection from microarray data for cancer classification—a machine learning approach," *Computational Biology and Chemistry*, vol. 29, no. 1, pp. 37–46, 2005.

[11] M. Han and X. Liu, "Forward feature selection based on approximate Markov blanket," in *Advances in Neural Networks-ISNN 2012*, vol. 7368 of *Lecture Notes in Computer Science*, pp. 64–72, Springer, Berlin, Germany, 2012.

[12] J. Kittler, "Feature set search algorithms," in *Pattern Recognition and Signal Processing*, C. H. Chen, Ed., pp. 41–60, Sijthoff and Noordhoff, Alphen aan den Rijn, The Netherlands, 1978.

[13] P. Pudil, J. Novovičová, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognition Letters*, vol. 15, no. 11, pp. 1119–1125, 1994.

[14] L.-Y. Chuang, H.-W. Chang, C.-J. Tu, and C.-H. Yang, "Improved binary PSO for feature selection using gene expression data," *Computational Biology and Chemistry*, vol. 32, no. 1, pp. 29–38, 2008.

[15] B. Q. Hu, R. Chen, D. X. Zhang, G. Jiang, and C. Y. Pang, "Ant Colony Optimization Vs Genetic Algorithm to calculate gene order of gene expression level of Alzheimer's disease," in *Proceedings of the IEEE International Conference on Granular Computing (GrC '12)*, pp. 169–172, Hangzhou, China, August 2012.

[16] L. J. Cai, L. B. Jiang, and Y. Q. Yi, "Gene selection based on ACO algorithm," *Application Research of Computers*, vol. 25, no. 9, pp. 2754–2757, 2008.

[17] S. Wang, J. Wang, H. Chen, S. Li, and B. Zhang, "Heuristic breadth-first search algorithm for informative gene selection based on gene expression profiles," *Chinese Journal of Computers*, vol. 31, no. 4, pp. 636–649, 2008.

[18] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1–3, pp. 389–422, 2002.

[19] Q. Liu, A. H. Sung, Z. Chen et al., "Gene selection and classification for cancer microarray data based on machine learning and similarity measures," *BMC Genomics*, vol. 12, no. 5, article S1, 2011.

[20] X. Li, S. Peng, J. Chen, B. Lü, H. Zhang, and M. Lai, "SVM-T-RFE: a novel gene selection algorithm for identifying metastasis-related genes in colorectal cancer using gene expression profiles," *Biochemical and Biophysical Research Communications*, vol. 419, no. 2, pp. 148–153, 2012.

[21] K. K. Kandaswamy, K. Chou, T. Martinetz et al., "AFP-Pred: a random forest approach for predicting antifreeze proteins from sequence-derived properties," *Journal of Theoretical Biology*, vol. 270, no. 1, pp. 56–62, 2011.

[22] W. Wei, S. Visweswaran, and G. F. Cooper, "The application of naive Bayes model averaging to predict Alzheimer's disease from genome-wide data," *Journal of the American Medical Informatics Association*, vol. 18, no. 4, pp. 370–375, 2011.

[23] R. M. Parry, W. Jones, T. H. Stokes et al., "K-Nearest neighbor models for microarray gene expression analysis and clinical outcome prediction," *Pharmacogenomics Journal*, vol. 10, no. 4, pp. 292–309, 2010.

[24] T. Mehenni and A. Moussaoui, "Data mining from multiple heterogeneous relational databases using decision tree classification," *Pattern Recognition Letters*, vol. 33, no. 13, pp. 1768–1775, 2012.

[25] T. K. Wu, S. C. Huang, Y. L. Lin, H. Chang, and Y. R. Meng, "On the parallelization and optimization of the genentic-based ANN classifier for the diagnosis of students with learning disabilities," in *Proceedings of the IEEE International Conference on Systems Man and Cybernetics*, pp. 4263–4269, Istanbul, Turkey, 2010.

[26] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, "Diagnosis of multiple cancer types by shrunken centroids of gene expression," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 10, pp. 6567–6572, 2002.

[27] D. Geman, C. d'Avignon, D. Q. Naiman, and R. L. Winslow, "Classifying gene expression profiles from pairwise mRNA comparisons," *Statistical Applications in Genetics and Molecular Biology*, vol. 3, no. 1, 2004.

[28] A. C. Tan, D. Q. Naiman, L. Xu, R. L. Winslow, and D. Geman, "Simple decision rules for classifying human cancers from gene expression profiles," *Bioinformatics*, vol. 21, no. 20, pp. 3896–3904, 2005.

[29] X. Lin, B. Afsari, L. Marchionni et al., "The ordering of expression among a few genes can provide simple cancer biomarkers and signal BRCA1 mutations," *BMC Bioinformatics*, vol. 10, article 256, 2009.

[30] A. T. Magis and N. D. Price, "The top-scoring "N" algorithm: a generalized relative expression classification method from small numbers of biomolecules," *BMC Bioinformatics*, vol. 13, article 227, no. 1, 2012.

[31] H. Wang, H. Zhang, Z. Dai, M. Chen, and Z. Yuan, "TSG: a new algorithm for binary and multi-class cancer classification and informative genes selection," *BMC Medical Genomics*, vol. 6, supplement 1, article S3, 2013.

[32] P. Chopra, J. Lee, J. Kang, and S. Lee, "Improving cancer classification accuracy using gene pairs," *PLoS ONE*, vol. 5, no. 12, Article ID e14305, 2010.

[33] H. Wang, S.-H. Lo, T. Zheng, and I. Hu, "Interaction-based feature selection and classification for high-dimensional biological data," *Bioinformatics*, vol. 28, no. 21, pp. 2834–2842, 2012.

[34] H. Zhang, H. Wang, Z. Dai, M. S. Chen, and Z. Yuan, "Improving accuracy for cancer classification with a new algorithm for genes selection," *BMC Bioinformatics*, vol. 13, article 298, 2012.

[35] C. Kooperberg, M. LeBlanc, and J. Y. a. Dai, "Structures and assumptions: strategies to harness gene × gene and gene × environment interactions in GWAS," *Statistical Science*, vol. 24, no. 4, pp. 472–488, 2009.

[36] G. Mohana Lakshmi and K. Mythili, "Survey of gene-expression-based cancer subtypes prediction," *International Journal of Advances in Computer Science and Technology*, vol. 3, no. 3, pp. 207–211, 2014.

[37] K.-J. Kim and S.-B. Cho, "Meta-classifiers for high-dimensional, small sample classification for gene expression analysis," *Pattern Analysis and Applications*, 2014.

[38] Y. Leung and Y. Hung, "A multiple-filter-multiple-wrapper approach to gene selection and microarray data classification," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 7, no. 1, pp. 108–117, 2010.

[39] L. S. Wang, O. U. ZY, and Y. C. Zhu, "Classifying images with SVM method," *Computer Applications and Software*, vol. 22, no. 5, pp. 98–102, 2005.

[40] B. Liquet, K. L. Cao, H. Hocini, and R. Thiébaut, "A novel approach for biomarker selection and the integration of repeated measures experiments from two assays," *BMC Bioinformatics*, vol. 13, no. 1, article 325, 2012.

[41] S. Wu, Y. Xu, Z. Feng, X. Yang, X. Wang, and X. Gao, "Multiple-platform data integration method with application to combined analysis of microarray and proteomic data," *BMC Bioinformatics*, vol. 13, no. 1, article 320, 2012.

[42] A. C. Haury, P. Gestraud, and J. P. Vert, "The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures," *PLoS ONE*, vol. 6, no. 12, Article ID e28210, 2011.

[43] D. G. Beer, S. L. R. Kardia, C. Huang et al., "Gene-expression profiles predict survival of patients with lung adenocarcinoma," *Nature Medicine*, vol. 8, no. 8, pp. 816–824, 2002.

[44] S. A. Armstrong, J. E. Staunton, L. B. Silverman et al., "MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia," *Nature Genetics*, vol. 30, no. 1, pp. 41–47, 2002.

[45] J. Khan, J. S. Wei, M. Ringnér et al., "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nature Medicine*, vol. 7, no. 6, pp. 673–679, 2001.

[46] C. M. Perou, T. Sørile, M. B. Eisen et al., "Molecular portraits of human breast tumours," *Nature*, vol. 406, no. 6797, pp. 747–752, 2000.

[47] A. Bhattacharjee, W. G. Richards, J. Staunton et al., "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 24, pp. 13790–13795, 2001.

[48] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, and I. S. Lossos, "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, pp. 503–511, 2000.

[49] E. J. Yeoh, M. E. Ross, S. A. Shurtleff et al., "Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling," *Cancer Cell*, vol. 1, no. 2, pp. 133–143, 2002.

[50] A. I. Su, J. B. Welsh, L. M. Sapinoso et al., "Molecular classification of human carcinomas by use of gene expression signatures," *Cancer Research*, vol. 61, no. 20, pp. 7388–7393, 2001.

[51] S. Ramaswamy, P. Tamayo, R. Rifkin et al., "Multiclass cancer diagnosis using tumor gene expression signatures," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 26, pp. 15149–15154, 2001.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 154251, 11 pages
http://dx.doi.org/10.1155/2014/154251

Research Article

miRNA Signature in Mouse Spermatogonial Stem Cells Revealed by High-Throughput Sequencing

Tao Tan,[1,2,3] Yanfeng Zhang,[1,4] Weizhi Ji,[1,3] and Ping Zheng[2]

[1] *Yunnan Key Laboratory of Primate Biomedical Research, No. 1 Boda Road, Yuhua Area, Chenggong District, Kunming, Yunnan 650500, China*
[2] *State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan 650223, China*
[3] *Kunming Biomed International and National Engineering Research Center of Biomedicine and Animal Science, Kunming, Yunnan 650500, China*
[4] *Department of Medicine, School of Medicine, Vanderbilt University, Nashville, TN 37203, USA*

Correspondence should be addressed to Weizhi Ji; wji@kbimed.com and Ping Zheng; zhengp@mail.kiz.ac.cn

Received 21 May 2014; Accepted 20 June 2014; Published 17 July 2014

Academic Editor: Zhixiang Lu

Spermatogonial stem cells (SSCs) play fundamental roles in spermatogenesis. Although a handful of genes have been discovered as key regulators of SSC self-renewal and differentiation, the regulatory network responsible for SSC function remains unclear. In particular, small RNA signatures during mouse spermatogenesis are not yet systematically investigated. Here, using next generation sequencing, we compared small RNA signatures of in vitro expanded SSCs, testis-derived somatic cells (Sertoli cells), developing germ cells, mouse embryonic stem cells (ESCs), and mouse mesenchymal stem cells among mouse embryonic stem cells (ESCs) to address small RNA transition during mouse spermatogenesis. The results manifest that small RNA transition during mouse spermatogenesis displays overall declined expression profiles of miRNAs and endo-siRNAs, in parallel with elevated expression profiles of piRNAs, resulting in the normal biogenesis of sperms. Meanwhile, several novel miRNAs were preferentially expressed in mouse SSCs, and further investigation of their functional annotation will allow insights into the mechanisms involved in the regulation of SSC activities. We also demonstrated the similarity of miRNA signatures between SSCs and ESCs, thereby providing a new clue to understanding the molecular basis underlying the easy conversion of SSCs to ESCs.

1. Introduction

Embryonic development in mice involves the migration of primordial germ cells (GCs) to the genital ridge and their subsequent differentiation into gonocytes. At about 6 days after birth, the gonocytes in male mice either undergo a transition to spermatogonia stem cells (SSCs), the foundation for continuous spermatogenesis throughout the reproductive lifetime, or develop directly into type A1 spermatogonia [1]. Spermatogenesis does not occur until puberty (about 3 weeks after birth), at which time SSCs undergo active self-renewal and differentiation to give rise to daughter cells for spermatogenesis [1]. SSCs thus play a fundamental role in spermatogenesis and male reproductive biology. Abnormalities in SSC function and regulation are closely related to male infertility, and SSC transplantation has potential clinical applications. Furthermore, unipotent SSCs have unique features in terms of their capacity to be easily reprogrammed into pluripotent embryonic stem cell- (ESC-) like cells in culture. These SSC-derived pluripotent cells are generally referred to as germline-derived pluripotent stem cells (gPSCs). When seeded at low density (<8000 SSCs per well in 24-well plate), SSCs undergo spontaneous conversion into gPSCs without modification of the culture medium [2]. gPSCs can also be derived from neonatal or adult murine or human testicular tissue [3–8]. These gPSCs

display morphological, functional, and molecular characteristics akin to ESCs [9, 10]. For example, they demonstrate pluripotent differentiation into cells forming all three germ layers and GCs [3, 4] and display similar gene, protein [11], and microRNA (miRNA) expression profiles [12], as well as epigenetic signatures, to ESCs. SSCs are therefore considered a potential source of pluripotent stem cells [13, 14].

The importance of SSCs means that numerous studies have investigated the regulation of self-renewal and differentiation activities of mouse and human SSCs in vivo or in vitro. Key genes, growth factors, and signaling pathways have been identified which are essential for SSC self-renewal and differentiation [15–22]. In addition, small noncoding RNAs also play essential roles in regulating SSC functions, such as spermatogenesis [23, 24]. Small RNAs are noncoding RNAs of 18–32 nt long, which can be further divided into three distinct classes: miRNAs, endogenous small interfering RNAs (endo-siRNAs), and Piwi-interacting RNAs (piRNAs). Comparisons of small noncoding RNA profiles in GCs at variable developmental stages and in testicular somatic cells identified a specific group of small noncoding RNAs important for SSC function. For instance, miR-34C is expressed specifically in mouse pachytene spermatocytes and in round spermatids and might play a role in regulating germ cell development [25]. miR-21 is highly expressed in mouse SSC populations and is important for self-renewal or homeostasis of SSCs [26].

Although SSCs are considered to be readily reprogrammed into gPSCs, the underlying mechanisms are poorly understood. Several pioneering studies have explored the possible mechanisms by comparing the molecular properties of SSCs and gPSCs (or ESCs). Kanatsu-Shinohara et al. examined the gene expression profiles of mouse SSCs and gPSCs by microarray analysis and revealed significant differences in mRNA expression patterns between these two cell types [27]. However, relatively fewer proteins were differentially expressed in gPSCs compared with SSCs in a proteomic assay [11]. Four transcription factors, including Oct4, Sox2, Klf4, and c-Myc, are widely used in reprogramming fibroblasts into pluripotent stem cells [28–30]. Although mRNAs of these Yamanaka factors were transcribed in SSCs, their expression levels were only 5–40% of those in pluripotent stem cells. Moreover, Sox2 protein expression was not detected and the protein levels of Oct4, Klf4, and c-Myc were extremely low in SSCs compared with ESCs or gPSCs [27]. Thus, the spontaneous conversion of SSCs to gPSCs cannot be explained simply by the rare transcription of reprogramming factors in SSCs. In this study, we investigated the whole-genome small noncoding RNA expression profiles of in vitro expanded mouse SSCs, developing GCs, mouse testis somatic cells (Sertoli cells (STs)), mouse ESCs, and mouse mesenchymal stem cells (MSCs). We compared their small noncoding RNA profiles and identified several highly expressed small RNAs in SSCs. Moreover, we found that ESCs and SSCs exhibited similar miRNA profiles, which could provide a new clue to understanding the molecular mechanisms underlying the spontaneous reprogramming of unipotent SSCs into multipotent gPSCs.

2. Materials and Methods

2.1. Ethics Statement. This study was carried out in strict accordance with the recommendations in the *Guide for the Care and Use of Laboratory Animals* of the National Research Council. The protocol was approved by the Institutional Animal Care and Use Committee (IACUC) of Kunming Institute of Zoology, Chinese Academy of Sciences. All surgery was performed under isoflurane anesthesia, and all efforts were made to minimize suffering.

2.2. Derivation and Expansion of Mouse SSCs. Testes were dissected from 4-5-week-old CD1 mice. Testicular tubules were isolated from the tunica albuginea and mechanically dissociated with forceps. The testicular tubules were then digested with 1 mg/mL collagenase IV for 15 min, washed twice with phosphate-buffered saline (PBS), and digested with 0.05% trypsin for 10 min. An equal volume of defined fetal bovine serum (FBS; Hyclone, Logan, UT, USA) was then added. The single-cell suspension was passed through a 30 μm filter and centrifuged at 300 g for 5 min at room temperature, and the supernatant was aspirated. The cells were incubated with Feeder Removal MicroBeads (Miltenyi Biotec, Auburn, CA, USA) and the negatively-labeled cells were collected, according to the manufacturer's instructions. The cell suspension was then centrifuged at 300 g for 5 min at room temperature and the supernatant was aspirated. The cells were plated onto MEF feeders from E13.5 mouse fetuses (CD1) in SSC medium.

Mouse SSCs were cultured as described previously [31], with minor modifications. Briefly, SSCs were seeded onto MEF feeders from E13.5 mouse fetuses (CD1) and cultured in StemPro-34 (Invitrogen, Carlsbad, CA, USA) supplemented with 2 mM glutamine (Invitrogen), 0.1 mM mercaptoethanol, 1× nonessential amino acids (Invitrogen), 1× penicillin-streptomycin (Invitrogen), 1× sodium pyruvate (Invitrogen), 40 ng/mL glial cell-derived neurotrophic factor (R&D Systems, Minneapolis, MN, USA), 10 ng/mL epidermal growth factor, 10^3 U/mL leukocyte migration inhibitory factor (Chemicon, Temecula, CA, USA), 10 ng/mL basic fibroblast growth factor (Chemicon), 60 μM putrescine, and 10% defined FBS (Hyclone) (subsequently referred to as SSC medium). The cells were passaged with 0.05% trypsin every 6-7 days. All chemicals were from Sigma Chemical (St. Louis, MO, USA) unless otherwise stated.

2.3. Isolation and Expansion of MSCs. MSCs were generated from bone marrow from tibias and femurs of 4-5-week-old CD1 mice, as described previously [32]. Established MSCs were cultured in low-glucose DMEM medium supplemented with 10% defined FBS (Hyclone), 2 mM glutamine (Invitrogen), 100 U/mL penicillin (Invitrogen), and 100 mg/mL streptomycin (Invitrogen).

2.4. Sertoli Cells and Developing Germ Cells Purification. Sertoli cells and developing germ cells were isolated from 4-5-week-old CD1 mouse testicles as previously described [33].

Isolated cells were resuspended in 1 mL of Trizol (Invitrogen) for subsequent use.

2.5. Immunofluorescence and Confocal Microscopy. Cells were fixed with 4% paraformaldehyde for 10–15 min at 25°C and then rinsed three times in PBS, followed by permeabilization with 0.2% Triton X-100 for 10–15 min. Cells were then blocked in 5% goat serum for 30 min at 25°C and incubated with primary and secondary antibodies (Table S1) before imaging under an LSM 510 META confocal microscope (Carl Zeiss, Jena, Germany) (see Supplementary Material available online at http://dx.doi.org/10.1155/2014/154251). Antibodies were obtained commercially and DNA was labeled with Hoechst 33342 or propidium iodide. An isotype-matched IgG was used as negative control in each experiment.

2.6. Reverse Transcription-Polymerase Chain Reaction. Total RNA was extracted from mouse MEF cells (negative control) and mouse SSCs using Trizol (Invitrogen). RNAs were subjected to treatment with DNase I (Fermentas, Vilnius, Lithuania) to remove possible genomic DNA contamination. Reverse transcription was carried out with approximately 2 μg of total RNA using a RevertAid H Minus First strand cDNA synthesis kit (Fermentas). One microliter of RT products was added to 1× Reaction Ready HotStart PCR master mix (Takara, Dalian, China) in a final volume of 25 μL and amplified under the following conditions: 1 cycle at 95°C for 5 min; 25–35 cycles at 95°C for 30 sec, 56–58°C for 30 sec, 72°C for 30 sec, and a full extension cycle at 72°C for 10 min. The sequences of the specific primer sets are provided in Table S2. The polymerase chain reaction products were separated on 2% agarose gels and visualized after staining with ethidium bromide.

2.7. Flow Cytometric Analysis of Mouse Mesenchymal Stem Cell Surface Antigens. 2×10^5 mouse MSCs were harvested and incubated with 1 μg of phycoerythrin (PE) conjugated antibodies or control isotype immunoglobulin Gs (IgGs) (Table S1) at 4°C for 30 minutes. Samples were analyzed using a FACS vantage SE (BD Biosciences, San Jose, CA, USA).

2.8. RNA Extraction and Small RNA Sequencing. Total RNA was isolated from mouse ESCs, SSCs, GCs, STs, and MSCs using Trizol (Invitrogen). Ten micrograms of total RNA from each sample was separated by 15% denaturing polyacrylamide gel electrophoresis and visualized by SYBR-gold staining. Small RNAs of 18–40 nt were gel-purified, and cDNAs were prepared using the Illumina small RNA preparation kit (Illumina, San Diego, CA, USA) and sequenced using the Illumina HiSeq 2000.

2.9. High-Throughput Sequencing Analysis and Annotation of Small RNAs. The Illumina base-calling pipeline was used for fluorescent image deconvolution, quality value calculation, and sequence conversion to obtain reads with a length of 50 nt. High-quality (clean) reads were obtained after trimming the 5′ and 3′ adaptors and eliminating contaminants and inadequate (<18 nt) and low-quality reads. The clean reads were then mapped to the mouse genome (mm9) using SOAP2 [34]. Perfectly matched reads were summarized and retained for further analyses. Read annotations were performed as described previously. Briefly, a hierarchical order that classified reads into specific RNA species was determined for annotation using the BLASTn (ftp://ftp.ncbi.nih.gov/blast/) program. The annotation order was miRNA > rRNA/snoRNA/tRNA/scRNA/snRNA > piRNA > endo-siRNA.

2.10. miRNA Profiling Analysis. Perfectly aligned reads annotating to miRNAs were initially counted; then miRNA expression levels were normalized using log2-RPM within each sample. An RPM value ≥1 for each mature miRNA was regarded as indicating expression. To identify miRNA signatures in mouse SSCs, each mature miRNA with ≥2-fold changes between SSCs and the other four cell types was regarded as an SSC-specific high expression miRNA.

2.11. piRNA Profiling Analysis. Because of clustering and repeat-derived characteristics of piRNAs, we analyzed piRNA expression using modified RPM normalization. Briefly, we used weighted #reads, ω_{piRNA} = #Reads/#Hits, instead of the number of reads (#Reads) to calculate RPM values. The genome-wide distributions of piRNA expression on both strands were compared among four samples.

2.12. Endo-siRNA Analysis. Endo-siRNA was identified on the basis of three stringent screening criteria analogous to those described previously [35]: (1) length of small RNA ranged from 18–23 nt; (2) reads of small RNAs perfectly matched to the mouse genome (mm9); (3) repeat-derived reads. As for piRNA analysis, the weighted expression was calculated and compared for putative endo-siRNA profiles.

3. Results

3.1. Derivation and Characterization of Mouse SSCs and MSCs. Mouse SSCs were derived and expanded according to the protocol developed by Kanatsu-Shinohara et al. [31]. The cells displayed typical germ stem cell morphology, expressed SSC markers including GFRα1, PLZF (ZBTB16), NGN3 (Neurog3), LIN28, and E-cadherin (CDH1) [36] (Supplementary Figure S1), and could be maintained in culture for more than 30 passages. Mouse MSCs were isolated and cultured as described previously [37]. The identity of the MSCs was verified by their spindle-shaped morphology, the expression of the MSC markers Sca-1 and CD44 and absence of hematopoietic markers CD45 and CD11b (Supplementary Figure S2), and their abilities to differentiate into adipocytes, osteocytes, and chondrocytes in culture (data not shown) [38].

3.2. Overview of Small Noncoding RNA High-Throughput Sequencing. Small RNAs were separated on 15% denaturing polyacrylamide gels and fragments of 18–40 nt were extracted and purified and used to construct a cDNA library, using the Illumina small RNA sample preparation kit (Illumina,

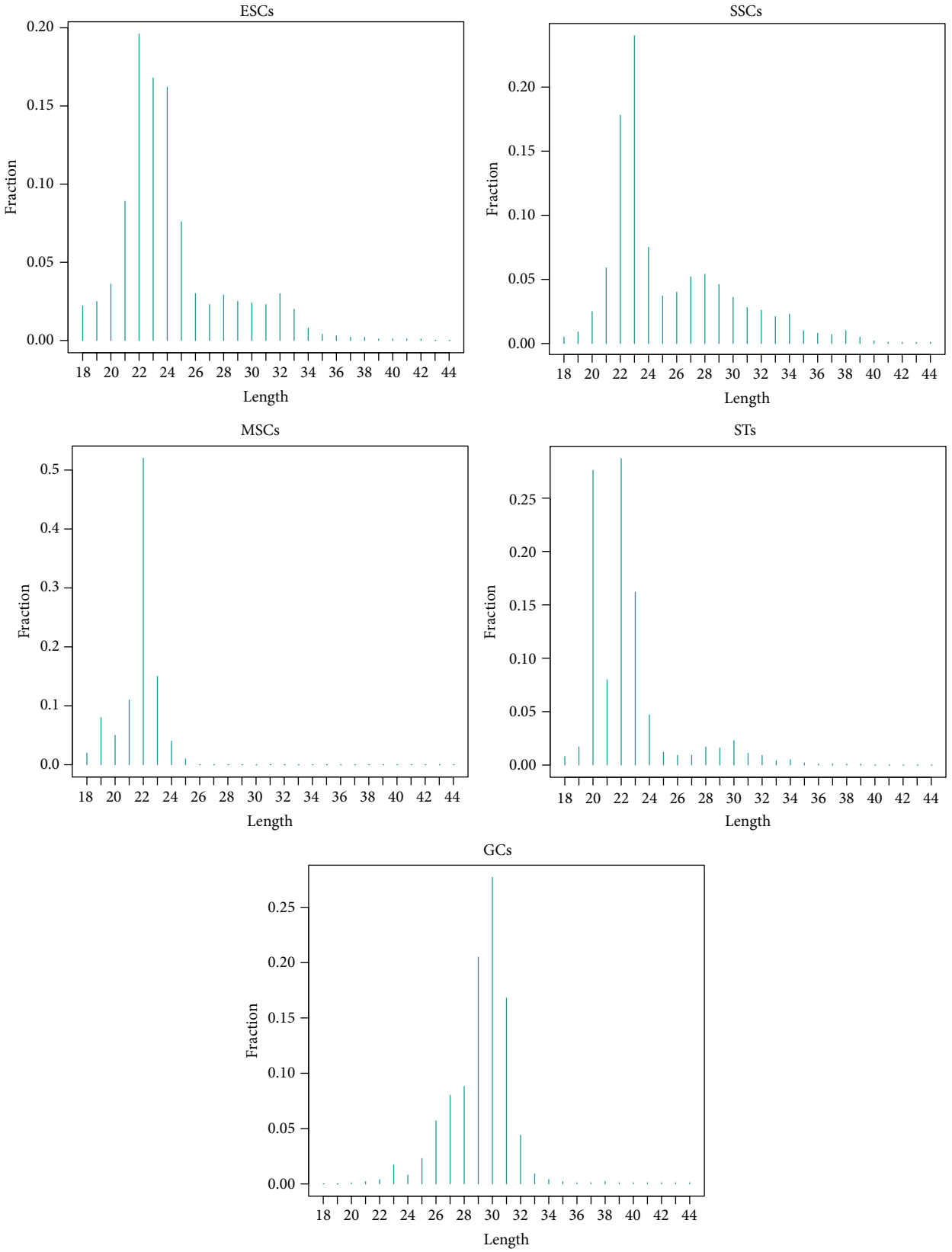


Figure 1: Size distribution of small RNAs in mouse ESCs, SSCs, MSCs, STs, and GCs. Perfectly mapped reads ≥18 nt long were densely plotted for each sample in this study.

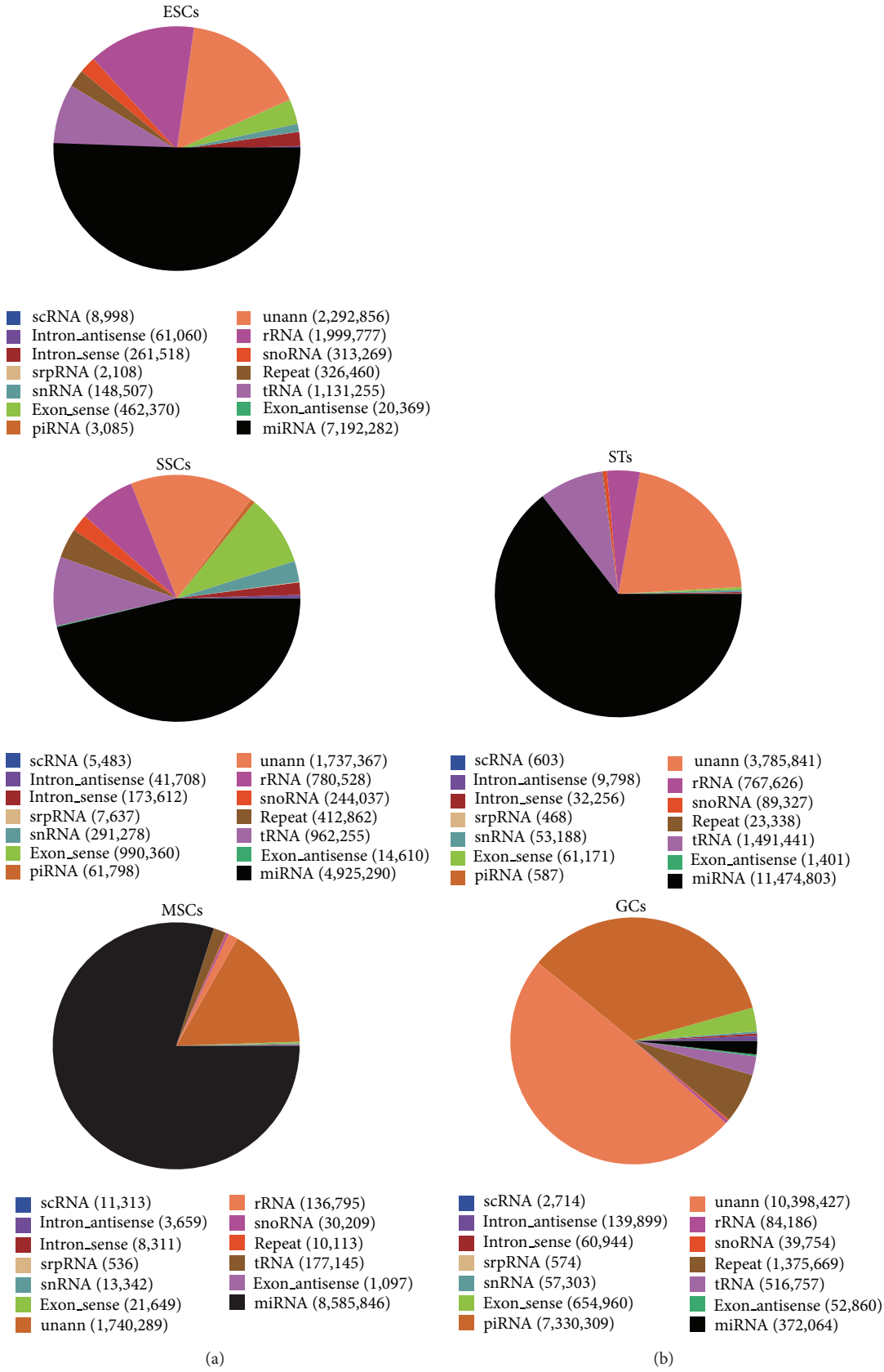


Figure 2: Small RNA annotations for mouse ESCs, SSCs, MSCs, STs, and GCs. The pie chart on the left for each cell type indicates the relative frequency of the annotated noncoding RNAs, and the right panel shows the absolute number of reads annotated.

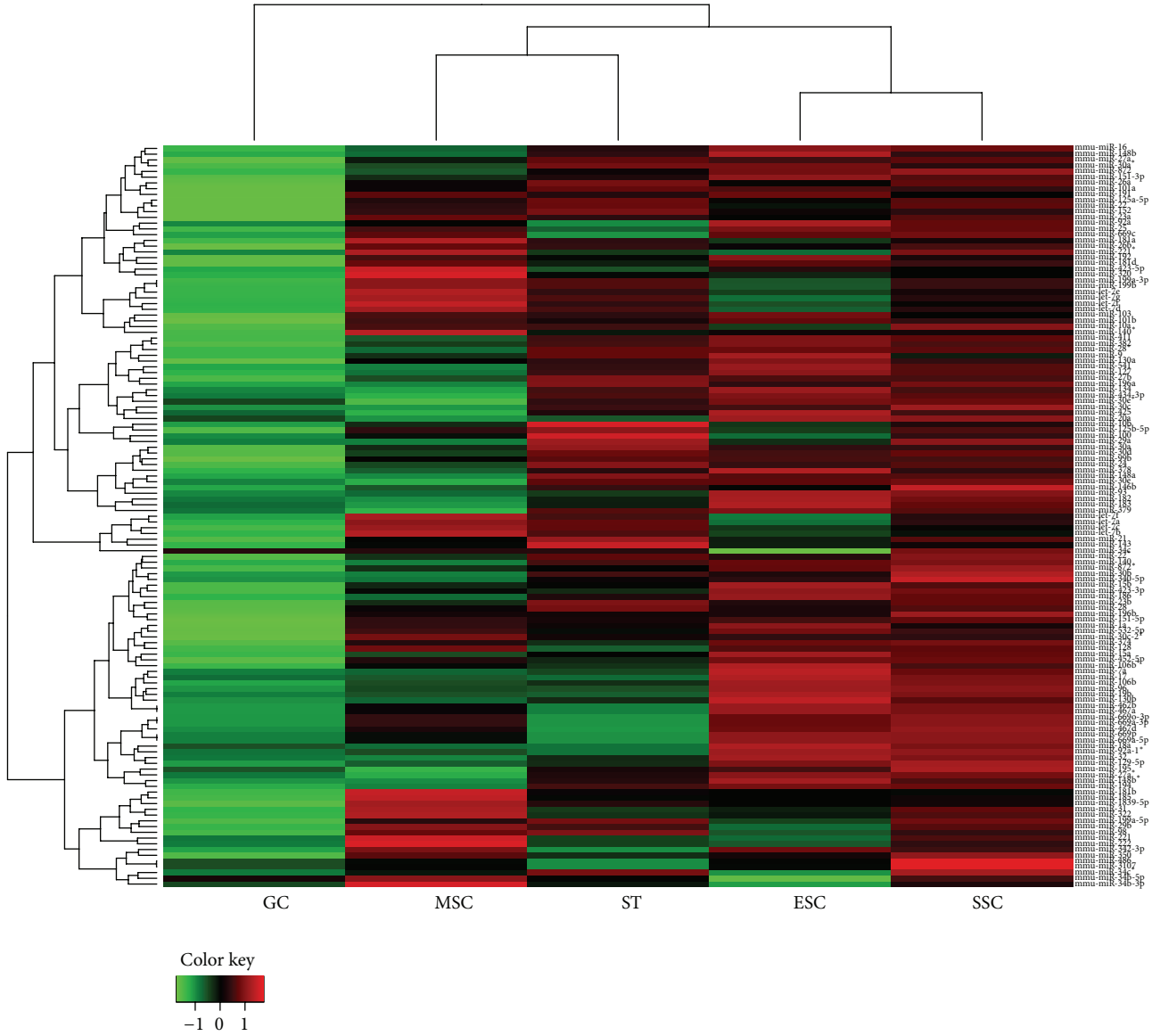


Figure 3: Heat map of five cell types determined by normalized miRNA expression. Each mature miRNA with log2-transformed expression level was used for clustering.

San Diego, CA, USA). Sequencing using an Illumina HiSeq 2000 produced 14.22×10^6, 10.65×10^6, 21.09×10^6, 17.79×10^6, and 10.74×10^6 clean reads from mouse ESCs, SSCs, developing GCs, STs, and MSCs, respectively. Around 75% of the total reads were perfectly matched to the mouse genome using the short oligonucleotide alignment program (SOAP2) [34] (Table 1). The length distributions of the mapped reads were compared among the five samples. Small RNAs in ESCs, SSCs, MSCs, and STs exhibited major length peaks at 22-23 nt, whereas those in GCs displayed a peak at 27–30 nt (Figure 1). The matched small RNA reads in each sample were further annotated and categorized. The major type of annotated small RNA in developing GCs was piRNA, whereas miRNAs accounted for about 60% of total annotated reads in the other four cell samples (Figure 2). This suggests that there is a special requirement for piRNAs in spermatogenesis, as reported by previous studies [39, 40].

Table 1: Summary of small RNA sequencing in mouse ESCs, SSCs, GCs, STs, and MSCs.

	Total clean reads	Mapping to genome	Percentage
ESC	14,223,914	10,477,211	73.66
SSC	10,648,825	8,379,103	78.69
GC	21,086,420	17,136,082	81.27
ST	17,791,848	13,255,304	74.50
MSC	10,740,304	8,832,696	82.24

3.3. miRNA Signature of Mouse SSCs. miRNAs represent the most significant class of small RNAs in many key biological

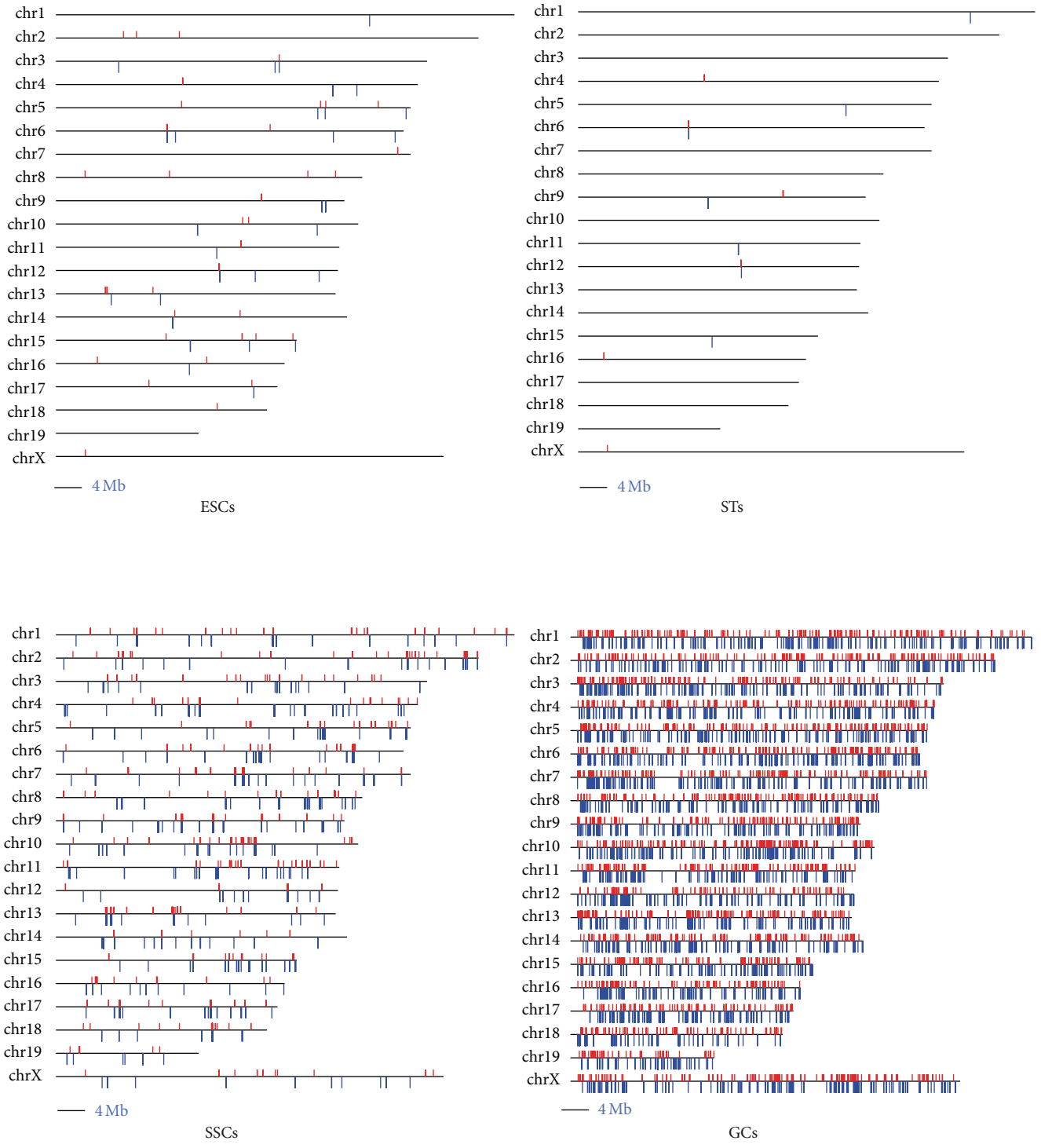


Figure 4: Genome-wide distribution of piRNA expression in four cell types. MSCs were excluded because of their low piRNA expression signal. Red and blue bars represent the plus and minus strands of expressed piRNA, respectively.

processes, including development, cell differentiation, the cell cycle, and apoptosis. To gain further insight into their functional roles in SSCs, we examined the miRNA expression profiles of these samples. After log2-read per million (RPM) normalization, mature miRNAs with RPM ≥ 1 were retained for further analysis. Heat map analysis (Figure 3) showed that SSCs were clustered with ESCs, whereas MSCs were clustered with STs. This clustering pattern was not influenced by the RPM threshold (data not shown), suggesting that SSCs resembled ESCs in terms of miRNA expression. This

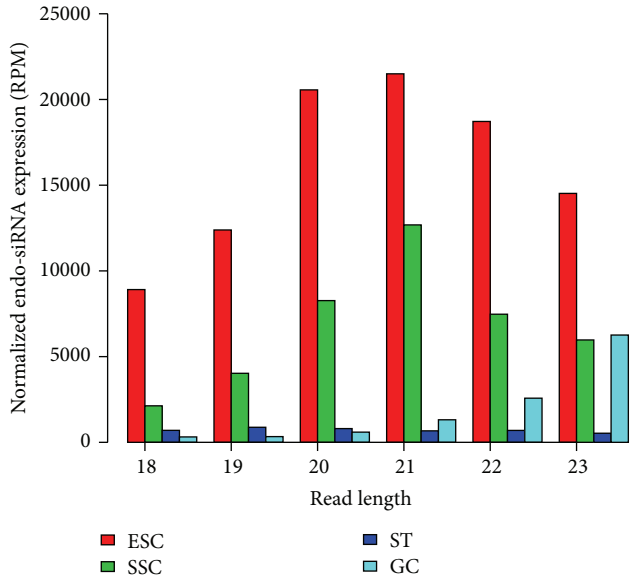


FIGURE 5: Relative endo-siRNA expression level as a function of read length. TPM denotes the total tag count per 10 million.

similarity in miRNA signatures might provide a molecular clue to understanding the spontaneous conversion of SSCs into gPSCs.

We identified a total of 128 mature miRNAs that were highly expressed specifically in mouse SSCs (Table S3), of which an X-linked miRNA cluster including numerous miRNAs was significantly expressed in SSCs. We also compared our data with previous study (raw data were obtained from Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) with the GSE29613 accession number GSE2564) to confirm the quality of our data. There were high correlations between our data and previous report (data not shown) [26]. Some components of this miRNA cluster (miR-883a, 883b) have previously been reported to be highly or specifically expressed in testes [41]. We also identified 232 miRNAs specifically expressed in ESCs (Table S4). One of the most notable miRNA features was the miR-302-367 cluster, which has been shown to be specifically expressed in ESCs and to play vital roles in reprogramming somatic cells into pluripotent stem cells [42–47]. Other miRNA clusters (miR-290 cluster, miR-200b-200a-429 cluster, and miR-106a-363 cluster) were also particularly abundant in ESCs compared with SSCs. The functional importance of the miR-290 cluster in stem cell pluripotency has been demonstrated previously [48].

3.4. piRNA and Endo-siRNA Profiles. piRNAs play important regulatory roles during spermatogenesis [39, 49]. Our genome-wide mapping of piRNAs consistently showed the highest enrichment of piRNA expression in developing GCs, followed by SSCs, ESCs, and STs (Figure 4 and Supplementary Figure S3). piRNA expression has been reported to cluster on one strand [50]. Because of the repeat-enriched property of piRNAs, we evaluated the frequency of this characteristic in piRNAs. Compared with ESCs and SSCs, endogenous retrovirus 1,2,3-derived repeats for piRNAs were significantly increased in developing GCs (Table S5), coincident with the suggestion that piRNAs are derived from retrotransposons [51]. We also examined the dynamics of endo-siRNAs during mouse spermatogenesis. Based on stringent criteria, we calculated weighted expression levels of endo-siRNAs and found that ESCs expressed the most abundant endo-siRNAs, followed by SSCs, with the lowest levels in developing GCs (Figure 5). This suggests that the trend for endo-siRNA expression profiles was similar to that for miRNAs.

4. Discussion

In this study, we used high-throughput sequencing to investigate the small RNA signatures and transitions during mouse spermatogenesis. Overall decreases in miRNAs and endo-siRNAs, in parallel with a gradual increase in piRNAs, were a feature of small RNA transition during mouse spermatogenesis (Figure 6). Although the mechanisms responsible for small RNA transitions remain unclear, these results provide insights into the interactive dynamic gene regulation at the posttranscriptional level during reproduction and development in mice. Moreover, the quantitative regulation by small RNAs further illuminates the spatiotemporal sophistication of gene expression in normal development, implying that spatiotemporal abnormalities of small RNAs may play roles in disease states.

Based on the opposite trends in small RNAs during mouse reproduction and development, focusing on any one type (or class) of small RNAs would only provide information on one aspect of gene regulation, and investigation of small RNAs at the system level is necessary to address posttranscriptional regulation and transition during mouse spermatogenesis in a quantitative manner. Further studies are also needed to determine if the pattern of small RNA transition during mouse spermatogenesis is conserved in humans. These results have potential implications for the reprogramming of SSCs to ESCs based on small RNAs.

The importance of SSCs in spermatogenesis and the ease of reprogramming them into gPSCs [2, 36, 52, 53] suggest that an understanding of the molecular properties of SSCs is essential. As noncanonical regulators of gene expression, small noncoding RNAs have been the subject of intensive studies over the past decade, and their roles in regulating SSC function and spermatogenesis have been investigated [26, 54–57]. In order to identify novel small noncoding RNAs potentially responsible for the functional properties of SSCs, we compared the genome-wide small RNA expression profiles of different mouse testis-derived cell populations, including somatic cells, developing GCs, and in vitro expanded SSCs. We also examined mouse ESCs and MSCs to investigate the similarities between SSCs and ESCs in terms of small noncoding RNA expression profiles. We identified a list of novel miRNAs that were specifically and abundantly expressed in mouse SSCs. Further investigation aimed at the functional dissection of these novel miRNAs could

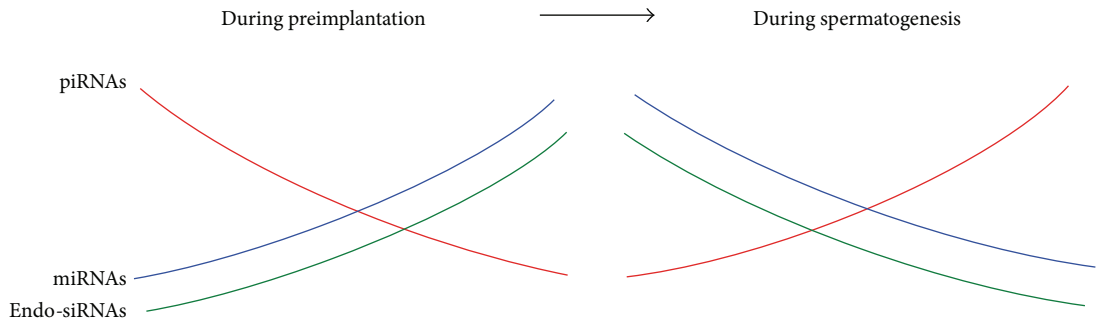


Figure 6: Schematic diagram of small RNA transition during mouse spermatogenesis.

generate new insights into the regulation of SSC activities. Interestingly, our data demonstrated that SSCs displayed similar miRNA expression profiles to ESCs. This similarity was unique to SSCs, and the miRNA signature of MSCs, the other somatic stem-cell type possessing multipotency, did not resemble that of ESCs. Overall, these results indicate that SSCs are primed to become ES-like cells, partially at the miRNA expression level, and this transition can occur easily under suitable culture conditions. miRNAs represent a higher regulatory layer of gene function and cell behavior, and the similarities in miRNA signatures between ESCs and SSCs provide new clues to understanding the molecular basis of the spontaneous reprogramming of unipotent SSCs into multipotent gPSCs. The results of this study may shed light on the mechanisms responsible for determining pluripotency and aid in the development of new ways to treat germline tumors.

5. Conclusion

Further investigation of SSC-specific miRNAs's functional annotation will allow insights into the mechanisms involved in the regulation of SSC activities. And the similarity of miRNA signatures between SSCs and ESCs will provide a new clue to understanding the molecular basis underlying the easy conversion of SSCs to ESCs.

Conflict of Interests

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

Authors' Contribution

Tao Tan and Yanfeng Zhang contributed equally to this work.

Acknowledgments

This work was supported by the National Program on Key Basic Research Project of China (973 Program; Grant no. 2012CBA01307), the National High-Tech R&D Program of China (863 Program; Grant no. 2012AA020701), the National Natural Science Fund for Young Scholars Grant (Grant no. 31101066), and the National Natural Science Fund (Grant no. 31371456). The authors thank Shao-Bin Xu for cluster computing support.

References

[1] M. Culty, "Gonocytes, the forgotten cells of the germ cell lineage," *Birth Defects Research C—Embryo Today: Reviews*, vol. 87, no. 1, pp. 1–26, 2009.

[2] K. Ko, M. J. Araúzo-Bravo, J. Kim, M. Stehling, and H. R. Schöler, "Conversion of adult mouse unipotent germline stem cells into pluripotent stem cells," *Nature Protocols*, vol. 5, no. 5, pp. 921–928, 2010.

[3] M. Kanatsu-Shinohara, K. Inoue, J. Lee et al., "Generation of pluripotent stem cells from neonatal mouse testis," *Cell*, vol. 119, no. 7, pp. 1001–1012, 2004.

[4] K. Guan, K. Nayernia, L. S. Maier et al., "Pluripotency of spermatogonial stem cells from adult mouse testis," *Nature*, vol. 440, no. 7088, pp. 1199–1203, 2006.

[5] M. Seandel, D. James, S. V. Shmelkov et al., "Generation of functional multipotent adult stem cells from GPR125$^+$ germline progenitors," *Nature*, vol. 449, no. 7160, pp. 346–350, 2007.

[6] S. Conrad, M. Renninger, J. Hennenlotter et al., "Generation of pluripotent stem cells from adult human testis," *Nature*, vol. 456, pp. 344–349, 2008.

[7] K. Ko, N. Tapia, G. Wu et al., "Induction of pluripotency in adult unipotent germline stem cells," *Cell Stem Cell*, vol. 5, no. 1, pp. 87–96, 2009.

[8] N. Kossack, J. Meneses, S. Shefi et al., "Isolation and characterization of pluripotent human spermatogonial stem cell-derived cells," *Stem Cells*, vol. 27, no. 1, pp. 138–149, 2009.

[9] C. J. Payne and R. E. Braun, "Human adult testis-derived pluripotent stem cells: revealing plasticity from the germline," *Cell Stem Cell*, vol. 3, no. 5, pp. 471–472, 2008.

[10] P. Sassone-Corsi, "Stem cells of the germline: the specialized facets of their differentiation program," *Cell Cycle*, vol. 7, no. 22, pp. 3491–3492, 2008.

[11] H. Kurosaki, Y. Kazuki, M. Hiratsuka et al., "A comparison study in the proteomic signatures of multipotent germline stem cells, embryonic stem cells, and germline stem cells," *Biochemical and Biophysical Research Communications*, vol. 353, no. 2, pp. 259–267, 2007.

[12] A. Zovoilis, J. Nolte, N. Drusenheimer et al., "Multipotent adult germline stem cells and embryonic stem cells have similar microRNA profiles," *Molecular Human Reproduction*, vol. 14, no. 9, pp. 521–529, 2008.

[13] H. Kubota and R. L. Brinster, "Technology insight: in vitro culture of spermatogonial stem cells and their potential therapeutic uses," *Nature Clinical Practice Endocrinology and Metabolism*, vol. 2, no. 2, pp. 99–108, 2006.

[14] T. Skutella, "Induced pluripotent stem cells from adult testis: a new source of stem cells?" *Regenerative Medicine*, vol. 4, no. 1, pp. 3–5, 2009.

[15] J. A. Costoya, R. M. Hobbs, M. Barna et al., "Essential role of Plzf in maintenance of spermatogonial stem cells," *Nature Genetics*, vol. 36, no. 6, pp. 653–659, 2004.

[16] H. Kubota, M. R. Avarbock, and R. L. Brinster, "Growth factors essential for self-renewal and expansion of mouse spermatogonial stem cells," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 47, pp. 16489–16494, 2004.

[17] C. Chen, W. Ouyang, V. Grigura et al., "ERM is required for transcriptional control of the spermatogonial stem cell niche," *Nature*, vol. 436, no. 7053, pp. 1030–1034, 2005.

[18] D. Ballow, M. L. Meistrich, M. Matzuk, and A. Rajkovic, "Sohlh1 is essential for spermatogonial differentiation," *Developmental Biology*, vol. 294, no. 1, pp. 161–167, 2006.

[19] J. M. Oatley, M. R. Avarbock, A. I. Telaranta, D. T. Fearon, and R. L. Brinster, "Identifying genes important for spermatogonial stem cell self-renewal and survival," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 25, pp. 9524–9529, 2006.

[20] B. Ryu, K. E. Orwig, J. M. Oatley, M. R. Avarbock, and R. L. Brinster, "Effects of aging and niche microenvironment on spermatogonial stem cell self-renewal," *Stem Cells*, vol. 24, no. 6, pp. 1505–1511, 2006.

[21] N. Vernet, C. Dennefeld, F. Guillou, P. Chambon, N. B. Ghyselinck, and M. Mark, "Prepubertal testis development relies on retinoic acid but not rexinoid receptors in Sertoli cells," *EMBO Journal*, vol. 25, no. 24, pp. 5816–5825, 2006.

[22] J. Lee, M. Kanatsu-Shinohara, K. Inoue et al., "Akt mediates self-renewal division of mouse spermatogonial stem cells," *Development*, vol. 134, no. 10, pp. 1853–1859, 2007.

[23] S. C. Mciver, S. D. Roman, B. Nixon, and E. A. Mclaughlin, "miRNA and mammalian male germ cells," *Human Reproduction Update*, vol. 18, no. 1, Article ID dmr041, pp. 44–59, 2012.

[24] R. P. Yadav and N. Kotaja, "Small RNAs in spermatogenesis," *Molecular and Cellular Endocrinology*, vol. 382, pp. 498–508, 2014.

[25] F. Bouhallier, N. Allioli, F. Lavial et al., "Role of miR-34c microRNA in the late steps of spermatogenesis," *RNA*, vol. 16, no. 4, pp. 720–731, 2010.

[26] Z. Niu, S. M. Goodyear, S. Rao et al., "MicroRNA-21 regulates the self-renewal of mouse spermatogonial stem cells," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 31, pp. 12740–12745, 2011.

[27] M. Kanatsu-Shinohara, J. Lee, K. Inoue et al., "Pluripotency of a single spermatogonial stem cell in mice," *Biology of Reproduction*, vol. 78, no. 4, pp. 681–687, 2008.

[28] 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.

[29] 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.

[30] J. M. Polo, E. Anderssen, R. M. Walsh et al., "A molecular roadmap of reprogramming somatic cells into iPS cells," *Cell*, vol. 151, no. 7, pp. 1617–1632, 2012.

[31] M. Kanatsu-Shinohara, N. Ogonuki, K. Inoue et al., "Long-term proliferation in culture and germline transmission of mouse male germline stem cells," *Biology of Reproduction*, vol. 69, no. 2, pp. 612–616, 2003.

[32] P. Tropel, D. Noël, N. Platet, P. Legrand, A. Benabid, and F. Berger, "Isolation and characterisation of mesenchymal stem cells from adult mouse bone marrow," *Experimental Cell Research*, vol. 295, no. 2, pp. 395–406, 2004.

[33] Y. F. Chang, J. S. Lee-Chang, S. Panneerdoss, J. A. MacLean II, and M. K. Rao, "Isolation of sertoli, leydig, and spermatogenic cells from the mouse testis," *BioTechniques*, vol. 51, no. 5, pp. 341–344, 2011.

[34] R. Li, C. Yu, Y. Li et al., "SOAP2: an improved ultrafast tool for short read alignment," *Bioinformatics*, vol. 25, no. 15, pp. 1966–1967, 2009.

[35] R. Song, G. W. Hennig, Q. Wu, C. Jose, H. Zheng, and W. Yan, "Male germ cells express abundant endogenous siRNAs," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 32, pp. 13159–13164, 2011.

[36] B. T. Phillips, K. Gassei, and K. E. Orwig, "Spermatogonial stem cell regulation and spermatogenesis," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 365, no. 1546, pp. 1663–1678, 2010.

[37] A. Peister, J. A. Mellad, B. L. Larson, B. M. Hall, L. F. Gibson, and D. J. Prockop, "Adult stem cells from bone marrow (MSCs) isolated from different strains of inbred mice vary in surface epitopes, rates of proliferation, and differentiation potential," *Blood*, vol. 103, no. 5, pp. 1662–1668, 2004.

[38] C. M. Kolf, E. Cho, and R. S. Tuan, "Mesenchymal stromal cells. Biology of adult mesenchymal stem cells: regulation of niche, self-renewal and differentiation," *Arthritis Research & Therapy*, vol. 9, no. 1, article 204, 2007.

[39] H. Gan, X. Lin, Z. Zhang et al., "piRNA profiling during specific stages of mouse spermatogenesis," *RNA*, vol. 17, no. 7, pp. 1191–1203, 2011.

[40] R. S. Pillai and S. Chuma, "piRNAs and their involvement in male germline development in mice," *Development Growth and Differentiation*, vol. 54, no. 1, pp. 78–92, 2012.

[41] R. Song, S. Ro, J. D. Michaels, C. Park, J. R. McCarrey, and W. Yan, "Many X-linked microRNAs escape meiotic sex chromosome inactivation," *Nature Genetics*, vol. 41, no. 4, pp. 488–493, 2009.

[42] 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.

[43] L. C. Laurent, J. Chen, I. Ulitsky et al., "Comprehensive microRNA profiling reveals a unique human embryonic stem cell signature dominated by a single seed sequence," *Stem Cells*, vol. 26, no. 6, pp. 1506–1516, 2008.

[44] 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. 4, pp. 610–621, 2008.

[45] B. Liao, X. Bao, L. Liu et al., "MicroRNA cluster 302–367 enhances somatic cell reprogramming by accelerating a mesenchymal-to-epithelial transition," *Journal of Biological Chemistry*, vol. 286, no. 19, pp. 17359–17364, 2011.

[46] D. Subramanyam, S. Lamouille, R. L. Judson et al., "Multiple targets of miR-302 and miR-372 promote reprogramming of human fibroblasts to induced pluripotent stem cells," *Nature Biotechnology*, vol. 29, no. 5, pp. 443–448, 2011.

[47] Z. Sun, Q. Wei, Y. Zhang, X. He, W. Ji, and B. Su, "MicroRNA profiling of rhesus macaque embryonic stem cells," *BMC Genomics*, vol. 12, article 276, 2011.

[48] G. X. Y. Zheng, A. Ravi, J. M. Calabrese et al., "A latent pro-survival function for the Mir-290-295 cluster in mouse embryonic stem cells," *PLoS Genetics*, vol. 7, no. 5, Article ID e1002054, 2011.

[49] N. C. Lau, A. G. Seto, J. Kim et al., "Characterization of the piRNA complex from rat testes," *Science*, vol. 313, no. 5785, pp. 363–367, 2006.

[50] Z. Yan, H. Y. Hu, X. Jiang et al., "Widespread expression of piRNA-like molecules in somatic tissues," *Nucleic Acids Research*, vol. 39, no. 15, pp. 6596–6607, 2011.

[51] C. Juliano, J. Wang, and H. Lin, "Uniting germline and stem cells: the function of piwi proteins and the pirna pathway in diverse organisms," *Annual Review of Genetics*, vol. 45, pp. 447–469, 2011.

[52] J. Yu and J. A. Thomson, "Pluripotent stem cell lines," *Genes and Development*, vol. 22, no. 15, pp. 1987–1997, 2008.

[53] K. Caires, J. Broady, and D. McLean, "Maintaining the male germline: regulation of spermatogonial stem cells," *Journal of Endocrinology*, vol. 205, no. 2, pp. 133–145, 2010.

[54] D. P. Bartel, "MicroRNAs: target recognition and regulatory functions," *Cell*, vol. 136, no. 2, pp. 215–233, 2009.

[55] G. M. Buchold, C. Coarfa, J. Kim, A. Milosavljevic, P. H. Gunaratne, and M. M. Matzuk, "Analysis of MicroRNA expression in the prepubertal testis," *PLoS ONE*, vol. 5, no. 12, Article ID e15317, 2010.

[56] W. Sun, Y. S. Julie Li, H. D. Huang, J. Y. Shyy, and S. Chien, "MicroRNA: a master regulator of cellular processes for bioengineering systems," *Annual Review of Biomedical Engineering*, vol. 12, pp. 1–27, 2010.

[57] J. Bao, D. Li, L. Wang et al., "MicroRNA-449 and MicroRNA-34b/c function redundantly in murine testes by targeting E2F transcription factor-retinoblastoma protein (E2F-pRb) pathway," *Journal of Biological Chemistry*, vol. 287, no. 26, pp. 21686–21698, 2012.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 986048, 6 pages
http://dx.doi.org/10.1155/2014/986048

Research Article

Advanced Heat Map and Clustering Analysis Using Heatmap3

Shilin Zhao, Yan Guo, Quanhu Sheng, and Yu Shyr

Center for Quantitative Sciences, Vanderbilt University, Nashville, TN 37232, USA

Correspondence should be addressed to Yu Shyr; yu.shyr@vanderbilt.edu

Received 6 June 2014; Accepted 2 July 2014; Published 16 July 2014

Academic Editor: Leng Han

Heat maps and clustering are used frequently in expression analysis studies for data visualization and quality control. Simple clustering and heat maps can be produced from the "heatmap" function in R. However, the "heatmap" function lacks certain functionalities and customizability, preventing it from generating advanced heat maps and dendrograms. To tackle the limitations of the "heatmap" function, we have developed an R package "heatmap3" which significantly improves the original "heatmap" function by adding several more powerful and convenient features. The "heatmap3" package allows users to produce highly customizable state of the art heat maps and dendrograms. The "heatmap3" package is developed based on the "heatmap" function in R, and it is completely compatible with it. The new features of "heatmap3" include highly customizable legends and side annotation, a wider range of color selections, new labeling features which allow users to define multiple layers of phenotype variables, and automatically conducted association tests based on the phenotypes provided. Additional features such as different agglomeration methods for estimating distance between two samples are also added for clustering.

1. Introduction

Gene expression analysis is one of the most popular analyses in the field of biomedical research. In the age of high-throughput genomics, microarray technology dominated the market of high-throughput gene expression profiling for over a decade until the introduction of RNA-seq technology. Regardless of which high-throughput gene expression profiling assay used, the heat map is one of the most popular methods of presenting the gene expression data. A heat map is a graphical representation of data where the individual values contained in a matrix are represented as colors. There are many variations of heat map such as web heat map and tree map. Here, we focus on the biology heat map, which is typically used to represent the level of expression of genes across a number of comparable samples. A gene expression heat map's visualization features can help a user to immediately make sense of the data by assigning different colors to each gene. Clusters of genes with similar or vastly different expression values are easily visible. The popularity of the heat map is clearly evidenced by the huge number of publications that have utilized it.

Cluster analysis is another popular method frequently used with gene expression study [1]. In our context, clustering refers to the task of grouping together a set of samples based on the similarity of their gene expression patterns. There are two major applications of cluster analysis. First, it is often used as a quality control measurement for identifying outliers. Second, it can be used to classify sample subtypes. The majority of the time in gene expression studies, gene expression is quantified from samples originating from multiple biological conditions. For example, most gene expression studies will consist of disease and control groups. Samples are selected based on their phenotype. In the ideal scenario, after performing the cluster, samples with a specific phenotype are in one cluster and samples without this phenotype are in another cluster. However, in the real world, many factors can affect the cluster results. For example, biological contamination can cause a sample to fail to cluster within the group. Also, the phenotype used to select the sample might not be the driving force in this sample's gene expression pattern. There may be other phenotypes that cause the sample's gene expression pattern to behave differently from other samples within the same group. Thus, cluster analysis is an ideal tool to detect outlier samples in gene expression studies [2]. Also, cluster analysis can be used to identify novel subtypes [3]. For example, the breast cancer study from The Cancer Genome Atlas (TCGA) project [4] used clustering techniques

to discover the subtype of samples based on their gene expression patterns. This is especially useful when subtypes of the samples are unknown. Also, the clustering technique can be applied to both sample and gene. When applied to both, the heat map can help visualizing potential novel pathways [5] and coexpression patterns [6].

The most popular tools to generate heat maps and clusters include the "heatmap" function in R and Cluster 3.0 [7]. However, these tools have some limitations. First, they can be slow and sometimes not able to finish for large expression matrices. Second, they are insufficient for producing advanced graphics. Third, they lack customizability. For example, in the breast cancer study from The Cancer Genome Atlas (TCGA) project [4] mentioned previously, the authors used heat map and cluster figures to present subtypes of the samples. The heat map used in that publication showed several additional bars to indicate phenotypes, and these phenotype bars are the result of meticulous work done by hand. A tool that can automatically display such phenotypes with the heat map is highly desirable. Driven by such motivation, we have produced "heatmap3," an advanced heat map and cluster analysis tool in R. Our "heatmap3" package significantly improves the original "heatmap" function's functionality by adding more powerful and convenient features including highly customizable legends, multiphenotype display bars including continuous phenotypes such as age, a wider range of color selection, a wider range of distance and agglomeration method selection, and automatic association tests of phenotype and cluster groups. Our "heatmap3" package allows users to generate heat maps and clusters and to make annotations easily. Users with basic skill in R can operate "heatmap3" without trouble.

2. Implementation

The "heatmap3" package is developed based on the "heatmap" function in R, and it is also backward compatible with it (i.e., if a code were written for the "heatmap" function, it will also run with the "heatmap3" package without problem). All the commands and parameters for "heatmap" can also be used in "heatmap3." We have implemented many new parameters in the "heatmap3" package in order to accommodate for the more powerful features. Detailed explanations and a manual of these parameters can be found at the hosting website of "heatmap3" (http://cran.r-project.org/web/packages/heatmap3/index.html).

2.1. Compute the Hierarchical Clustering between Rows and Columns. To assess the similarity of gene expression patterns between two samples, a distance or score needs to be computed. The original "heatmap" function used the Euclidean distance as the default distance method and complete linkage as the agglomeration method; it is not easy to change the default distance method within the original "heatmap" function. Our "heatmap3" package provides a wide selection of distance and agglomeration options, such as centered Pearson correlation, uncentered Pearson correlation, and average linkage. More importantly, "heatmap3" uses the clustering function in the "fastcluster" package when the expression matrix is large. This package efficiently implements the seven most widely used clustering schemes: single, complete, average, weighted, Ward, centroid, and median linkage. By using the "fastcluster" package, "heatmap3" is able to produce hierarchical clusters much faster and more efficiently than the original "heatmap" function.

2.2. Plot the Heat Map and Dendrogram. The "heatmap3" package sorts the rows and columns based on the hierarchical clustering result. The colors will then be assigned to the genes to represent the expression value. A balance option is provided here to ensure the median color will represent zero value. The heat map and dendrogram are plotted in the same fashion as the original "heatmap" function. However, more customization parameters are implemented. For example, the user now can choose to display or hide the dendrogram.

2.3. Plot the Color Bar, Annotation, and Legend. A color bar which represents the relationship between colors and values will be automatically generated at the top left side of the figure. The categorical phenotypes such as gender and race and the continuous phenotypes such as age and drug dose can be annotated in the column side of the heat map figure. This allows users to easily compare the annotation with the heat map results and make proper inference. Furthermore, "heatmap3" provides the function interfaces for generating the user's own annotations and legends. Users can use their own R functions to generate figures in the legend position and annotation position.

2.4. Cut and Statistically Test for Annotation in Different Groups. Our "heatmap3" package provides an automatic grouping method. A cutoff needs to be provided, and the dendrogram tree will be cut at the height of cutoff. The samples will be divided into several groups and labeled by different colors at the cutoff level. Then, statistical tests will be performed to see if the annotations are distributed equally in different groups. We used a chi-squared test for factor annotations and ANOVA for continuous annotations. These group results and P values will be returned to the user so that they can be used as criteria for selecting the genes that best separated the samples.

3. Results

To demonstrate the "heatmap3" package's efficiency and visualization power, we used RNA-seq gene expression results from the TCGA breast cancer (BRCA) dataset. The example dataset and its command can be downloaded from https://github.com/slzhao/heatmap3. The complete read count and clinical information can be seen in Tables S1 and S2 (see Supplementary Tables S1 and S2 available online at http://dx.doi.org/10.1155/2014/986048). To install the "heatmap3" package, type the following command in R:

```
install.packages("heatmap3")
```

First, we performed differential analysis by the "edgeR" [8] package to compare the gene expression between triple negative samples versus nontriple negative samples. The P values and fold changes for genes were taken as annotation

information (Table S3). We selected 500 genes with the largest standard deviations and randomly selected 30 samples to generate the heatmap. By selecting genes with large standard deviations, we effectively removed the nonexpressed genes across all samples, and the results still remained unbiased. We also included several important clinical variables for demonstration purposes. The selected phenotype variables were age, triple negative (TN) status, estrogen receptor (ER) status, progesterone receptor (PR) status, and human epidermal growth factor receptor 2 (HER2) status.

Using these data, a heat map with legend color bar, column side annotations, and row side annotations was generated (Figure 1). The legend color bar indicates the relation between scaled expression values and colors, and the colors were balanced to ensure the white color represented zero value. We provided two annotation methods: color bar and categorical bar. Color bar is ideal to represent multiple phenotypes that are mutually exclusive. For example, for phenotypes of disease and normal, a sample can only be disease or normal but not both. Categorical bar is ideal to represent multiple phenotypes that are not mutually exclusive. For example, a sample can be TN and ER negative simultaneously.

The annotation on the y-axis side demonstrates how the customized function can be used for annotation. Here, we used the "showAnn" function within the package as an example. The categorical phenotype annotations (ER, PR, HER2 and TN) were separated into two columns, and the samples were labeled by black squares. The numeric annotation (age) was demonstrated by a scatter plot, and the values were labeled at the right axis. The annotation on the row side indicated an example of annotation by color bar. The green to red and orange to white colors here represent the log2 fold changes and the negative log10 P values, respectively. We can easily find that the genes increased in tripe negative samples (red color in log2 fold change annotation) were clustered in the bottom of heat map, while the genes decreased in tripe negative samples (green color in log2 fold change annotation) were clustered in the top of the heat map.

Using a height cutoff of 0.85 for the dendrogram tree on the column side, clearly, the samples were divided into two groups and labeled by different colors. As expected, the triple negative samples were enriched in the right group and non-triple negative samples were enriched in the left group. For the ER, PR, and HER2 levels, we can find that most of the samples were HER2 negative, and the ER and PR negative samples were enriched in the right group. Based on the results from the heat map, we might able to infer that ER and PR positive appear more in patients and they may have more important roles in defining triple negative samples. To generate Figure 1 using example data, enter the following command in R:

```
# assume "counts" is the expression data, "colGene" contains the colors indicating fold changes and P value, and "clinic" contains the ER, PR, HER2, TN, and age information,
temp<-apply(counts,1,sd),
selectedGenes<-rev(order(temp))[1:500],
heatmap3(counts[selectedGenes,],labRow="",margin=c(7,0),RowSideColors=colGene[selectedGenes,],
ColSideCut=0.85,ColSideAnn=clinic,ColSideFun=function(x) showAnn(x),ColSideWidth=1.2,balanceColor=T).
```

Table 1: The statistical test result for categorical annotation in different groups.

	Cluster1	Cluster2	P value by chi-square test
ER			
Negative	2	11	0.003
Positive	13	4	
Positive Percent	0.87	0.27	
PR			
Negative	4	13	0.003
Positive	11	2	
Positive Percent	0.73	0.13	
HER2			
Negative	13	13	0.023
Positive	2	2	
Positive Percent	0.13	0.13	

Table 2: The statistical test result for age in different groups, ANNOVA P value: 0.429.

Age	Cluster1	Cluster2
Min.	41.00	46.00
1st Qu.	51.00	49.00
Median	61.00	55.00
Mean	60.00	57.13
3rd Qu.	64.25	62.50
Max.	89.00	80.00

Association tests between phenotype and cluster groups were performed automatically by "heatmap3" (Tables 1 and 2). The number of categorical phenotypes and quantiles of continuous phenotype variables in each cluster group are summarized and reported. Chi-square test for categorical variables and ANOVA for continuous variables are performed by "heatmap3." Based on the results, ER, PR, and HER2 were not equally distributed between the two clusters. On the other hand, age had no association with the two clusters ($P = 0.429$).

The "heatmap3" package also provides an option which allows the generation of multiple heat maps and dendrograms based on the threshold criteria selected by the user. Using the same dataset, we performed heat map and cluster analysis using all genes, the top 3000 genes, and the top 500 genes selected by standard deviation. Figure 2 shows the three dendrograms. All three dendrograms showed clearly two large clusters. Using TN status as the primary phenotype, each time a more stringent standard deviation cutoff was used, the clusters became clearer between TN and non-TN. This example illustrates the importance of selecting more statistically varied genes for subtyping purposes. We can

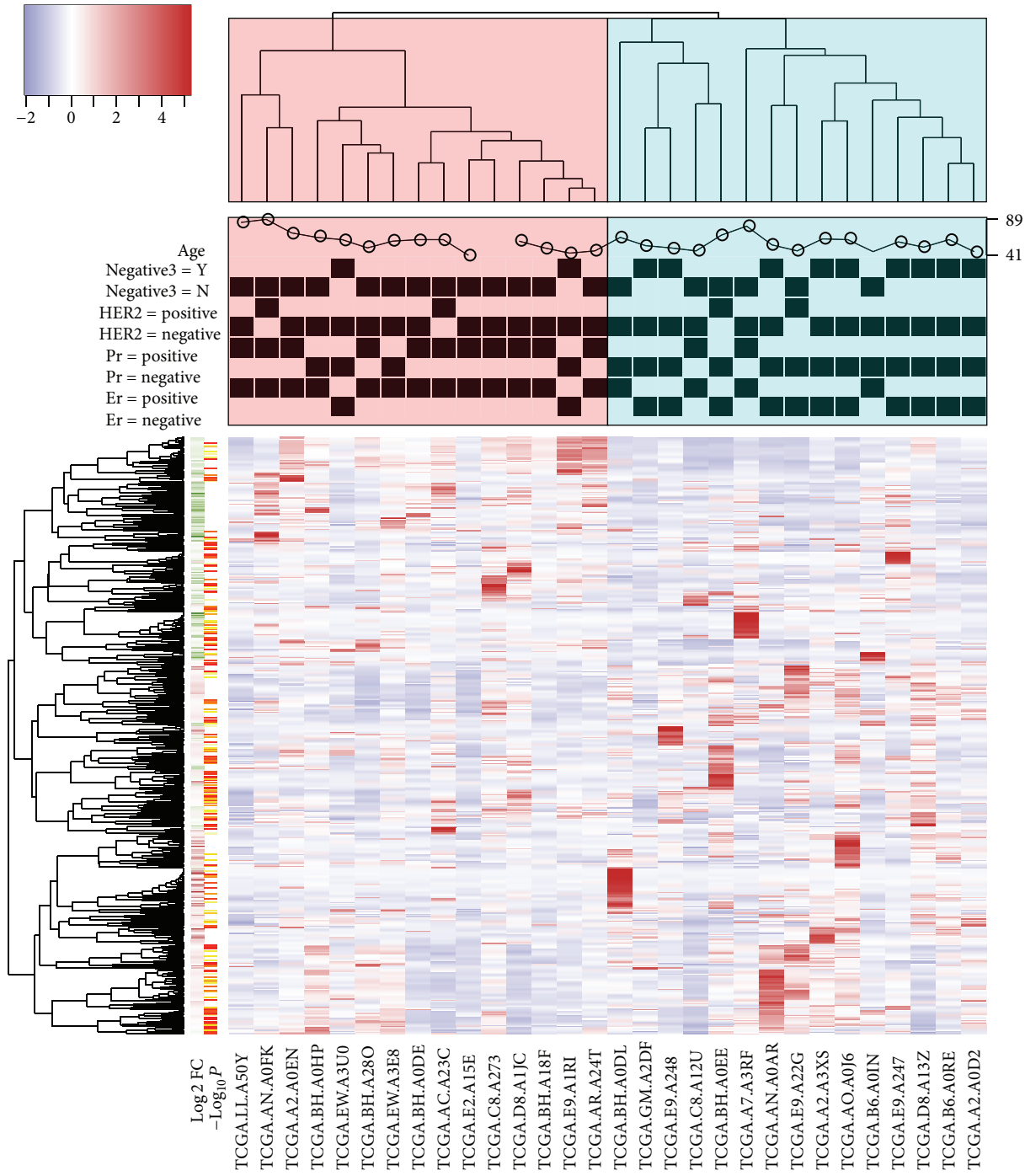


FIGURE 1: An example of "heatmap3" package. The heat map was generated based on 30 samples from TCGA BRCA dataset. The dendrogram of samples (top) was divided into two parts based on the correlation between samples' gene expression and then labeled, respectively. The categorical annotation bars (above heat map) demonstrate the annotation for age, TN, HER2, PR, and ER. The color bar on the left side demonstrates the log2 fold changes and negative log10 *P* values from comparison of triple negative patients versus nontriple negative patients.

conclude that the genes with the highest standard deviations can be used to separate the triple negative and nontriple negative samples. To generate Figure 2 using the example dataset, type the following commands in R:

assume "counts" is the expression data, "colGene" contains the colors indicating fold changes and *P* value, and "clinic" contains the ER, PR,HER2, TN, and age information,

heatmap3(counts,topN=c(500,3000,nrow(counts)), labRow="",margin=c(7,0),RowSideColors=colGene, ColSideCut=0.85,ColSideAnn=clinic,ColSideFun= **function**(x) showAnn(x),ColSideWidth=1.2,balance Color=T).

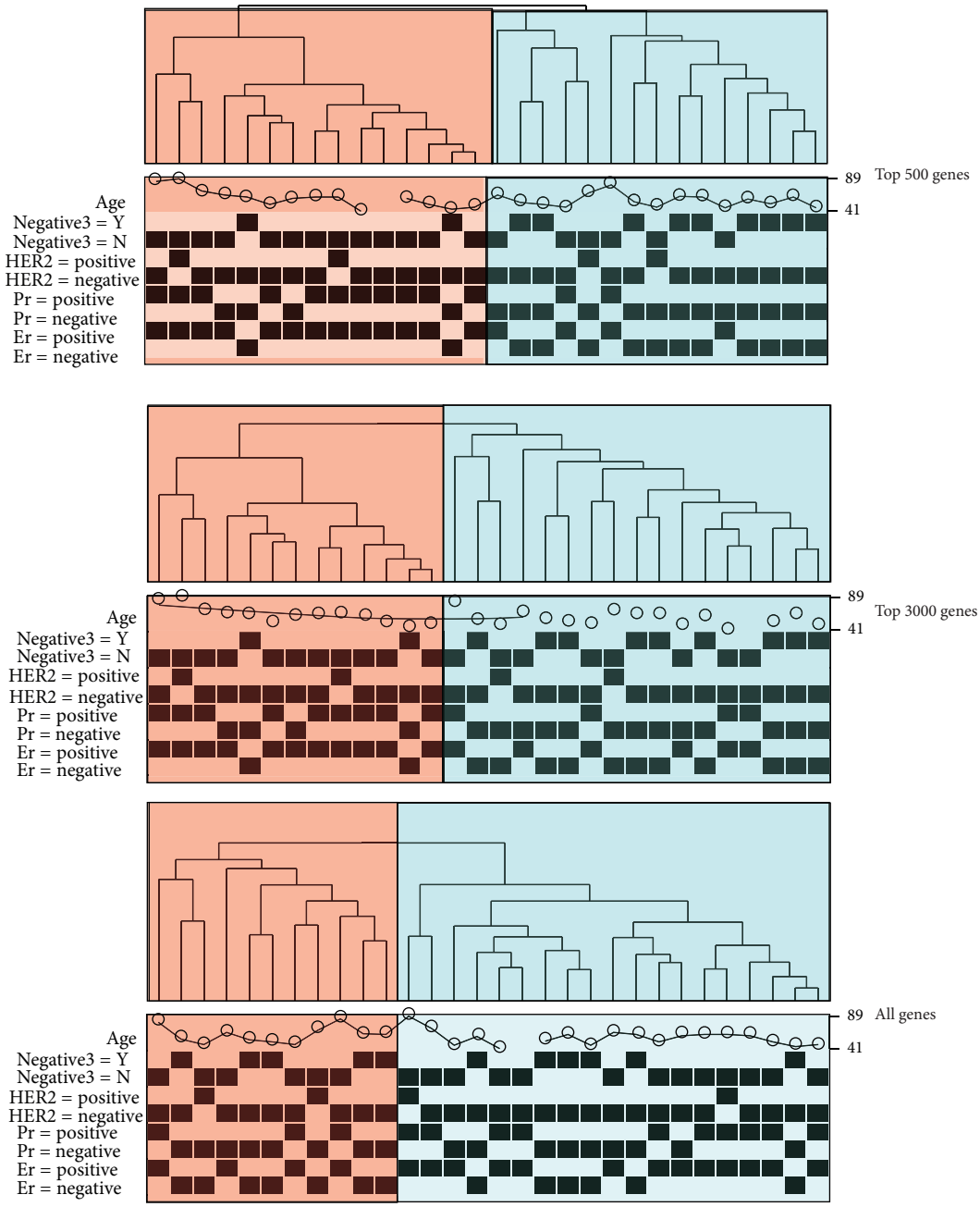


Figure 2: The dendrograms and clusters generated by top 500, top 3000, and all genes which were selected by standard deviation. The triple negative samples were more enriched in one group when genes with larger standard deviation were used. The results demonstrate that the "heatmap3" package can be helpful in selecting genes that best represent the phenotypes of samples.

4. Discussions

In this paper, we discussed the importance of heat map and clustering analysis as well as the limitations of existing heat map and clustering tools. To address these limitations, we implemented the "heatmap3" package in R and demonstrated its effectiveness using RNA-seq data from a breast cancer study in TCGA. The "heatmap3" package is designed with advanced options and is completely backward compatible with the original "heatmap" function in R. Users with limited R skill can generate sophisticated heat maps and dendrograms with ease. In summary, the "heatmap3" package fills the void of advanced graphical options in current heat map tools. It provides the much needed customizability for heat map and cluster analysis.

Conflict of Interests

The authors declare that they have no conflict of interests regarding the publication of this paper.

Authors' Contribution

Shilin Zhao and Yan Guo have equal contribution.

Acknowledgments

This study was supported by Grant CCSG (P30 CA068485). The authors would like to thank Margot Bjoring for her editorial support.

References

[1] Y. Guo, S. Zhao, F. Ye, Q. Sheng, and Y. Shyr, "MultiRankSeq: multiperspective approach for RNAseq differential expression analysis and quality control," *BioMed Research International*, vol. 2014, Article ID 248090, 8 pages, 2014.

[2] S. Yang, X. Guo, Y. Yang et al., "Detecting outlier microarray arrays by correlation and percentage of outliers spots," *Cancer Informatics*, vol. 2, pp. 351–360, 2006.

[3] A. Sadanandam, C. A. Lyssiotis, K. Homicsko et al., "A colorectal cancer classification system that associates cellular phenotype and responses to therapy," *Nature Medicine*, vol. 19, no. 5, pp. 619–625, 2013.

[4] Cancer Genome Atlas Network, "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, pp. 61–70, 2012.

[5] J. M. Lee and E. L. L. Sonnhammer, "Genomic gene clustering analysis of pathways in eukaryotes," *Genome Research*, vol. 13, no. 5, pp. 875–882, 2003.

[6] P. D'Haeseleer, S. Liang, and R. Somogyi, "Genetic network inference: from co-expression clustering to reverse engineering," *Bioinformatics*, vol. 16, no. 8, pp. 707–726, 2000.

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

[8] 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.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 694967, 11 pages
http://dx.doi.org/10.1155/2014/694967

Research Article

Metabolic Modeling of Common *Escherichia coli* Strains in Human Gut Microbiome

Yue-Dong Gao,[1] Yuqi Zhao,[2] and Jingfei Huang[2,3]

[1] *Kunming Biological Diversity Regional Center of Instruments, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China*
[2] *State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, 32 Eastern Jiaochang Road, Kunming, Yunnan 650223, China*
[3] *Kunming Institute of Zoology, Chinese University of Hongkong, Joint Research Center for Bio-Resources and Human Disease Mechanisms, Kunming 650223, China*

Correspondence should be addressed to Yuqi Zhao; zhaoyq@mail.kiz.ac.cn and Jingfei Huang; huangjf@mail.kiz.ac.cn

Received 21 April 2014; Revised 11 June 2014; Accepted 13 June 2014; Published 13 July 2014

Academic Editor: Zhixi Su

The recent high-throughput sequencing has enabled the composition of *Escherichia coli* strains in the human microbial community to be profiled en masse. However, there are two challenges to address: (1) exploring the genetic differences between *E. coli* strains in human gut and (2) dynamic responses of *E. coli* to diverse stress conditions. As a result, we investigated the *E. coli* strains in human gut microbiome using deep sequencing data and reconstructed genome-wide metabolic networks for the three most common *E. coli* strains, including *E. coli* HS, UTI89, and CFT073. The metabolic models show obvious strain-specific characteristics, both in network contents and in behaviors. We predicted optimal biomass production for three models on four different carbon sources (acetate, ethanol, glucose, and succinate) and found that these stress-associated genes were involved in host-microbial interactions and increased in human obesity. Besides, it shows that the growth rates are similar among the models, but the flux distributions are different, even in *E. coli* core reactions. The correlations between human diabetes-associated metabolic reactions in the *E. coli* models were also predicted. The study provides a systems perspective on *E. coli* strains in human gut microbiome and will be helpful in integrating diverse data sources in the following study.

1. Introduction

Escherichia coli (*E. coli*) is the most widely studied prokaryotic model organism and an important species in the fields of biotechnology and microbiology. *E. coli* constitutes about 0.1% of human gut flora [1], which benefits human beings by providing supplemental nutrition, by enhancing nutrient acquisition, and by preventing the establishment of pathogenic bacteria within the intestine [2]. The study of this bacterium is both of importance for applications, such as environmental testing and metabolic engineering [3], and of interest as a fundamental physical problem. For example, a recent study demonstrated an obvious increase in the number of *E. coli* in the stool, while diarrhea was apparent [4].

In the recent five years, the flood of deep sequencing data has set the latest wave of microbiome research apart from earlier studies, with the ability to enumerate all of the cells in a complex microbial community at once [5]. For instance, using deep sequencing, the Human Microbiome Project (HMP) was launched to characterize the microbial communities found at several different sites on the human body and to analyze the role of these microbes in human health and disease [6, 7]. This switch from the low-throughput technique, culture-based enumeration, to the high-throughput technology of deep sequencing offers several advantages, including high accuracy, culture-free sampling, and comprehensive information. However, there are still two challenges to address. First, due to the huge data size and high complexity of the different algorithms, it is difficult to determine the exact roles of the various species in human microbiome, let alone strains of the same species. The composition of *E. coli* strains is of value to human health;

for example, changes in the *E. coli* composition were observed associated with intestinal inflammatory disorders in human and mice [8, 9]. Second, most of the microbiota community structures obtained from sequencing were "static," while the human microbiomes are diverse and dynamic. The diet changes, individual differences, sampling sites, and physical conditions are responsible for the dynamic responses of human microbiome [10–12]. However, the comprehensive responses of microbiome to the dynamic microenvironments can hardly be obtained from one or several samples.

To solve these problems, considerable efforts have been made to develop metabolic networks of *E. coli* [3, 13, 14]. These in silico models have been successfully applied in many fields. For example, they were frequently used in prediction of steady-state or dynamic responses of cells to changes in ecosystems [3]. In addition, the metabolic models can be easily integrated with other data sources, such as DNA sequencing [15], expression profiles [16], proteomics [17], or metabolomics [18]. Goals of such data integration efforts are (1) to gain a better understanding of the observable phenotypes of the cell, (2) to predict potential functions of molecular signatures, and (3) to apply these in silico models for biological discovery and engineering applications. As a result, integration of relevant omics data with metabolic models as a representative species in the human gut microbiota elucidates the changes in the gut microbiota.

In this study, we performed in silico modeling of metabolic networks of *E. coli* strains in human gut microbiome. First, we determined *E. coli* strains in human gut microbiome using 148 fecal metagenomes. Next, we reconstructed genome-wide metabolic network of common *E. coli* strains in human gut. Then, the cellular phenotypes were predicted and validated using the genome variation of *E. coli* and diet changes. The findings of the study will help in developing new technologies and tools for computational analysis and exploring the relationship between disease and changes in the human microbiome.

2. Materials and Methods

2.1. Human Gut Metagenomes and Reference Genomes. High-quality short reads of 148 human gut samples were retrieved from Human Microbiome Project (HMP, http://www.hmpdacc.org/). The sequenced and well-annotated *E. coli* genomes (totally 61 genomes) deposited in GenBank were downloaded from NCBI database (http://www.ncbi.nlm.nih.gov/), to build a reference genome database. The reads were aligned against the *E. coli* reference genome using BLASTN (version 2.2.27+) with $E < 0.01$, minimal 99% identity cutoff and considering the reads that were aligned onto only a single position in the reference genome.

2.2. De Novo Assembly and Identification of Genes. The reads of human gut samples were assembled by Newbler (454/Roche GS Mapper/Assembler), following the protocol in HMP [19]. The assembled scaffolds were aligned against *E. coli* genomes using BLASTN with minimal 99% identity cutoff and best hit output.

2.3. Reconstruction of Strain-Specific Metabolic Network. The *E. coli* pan-genome (the union of the gene sets of all the strains of a species) metabolic network has been generated in a recent study [20]. The strain-specific metabolic model could be reconstructed based on the pan-genome metabolic network. We generated metabolic networks for the common *E. coli* strains in human gut microbiome based on the pan-genome metabolic network.

In the process, we derived the strain-specific metabolic models using two commonly used algorithms of top-down metabolic reconstructions, including GIMME [21] and iMAT [22]. These two algorithms are different: the GIMME is a linear programming procedure, while the iMAT is a mixed integer linear programming procedure.

2.4. Predictions of Cellular Phenotypes Using Metabolic Network. Fluxes through reactions in the metabolic models can be predicted using flux balance analysis (FBA) [23]. In the process, fluxes are constrained by steady-state mass balances, enzyme capacities, and reaction directionality, which yield a solution space of possible flux values. Besides, FBA uses an objective function to identify flux distributions that maximize (or minimize) the physiologically relevant predicted solution. Cellular growth rate (biomass production in another word) was used as an objective function for FBA analyses performed in this study. The same biomass equation, growth (GAM) and nongrowth (NGAM) associated ATP requirement values, and PO (number of ATP molecules produced per pair of electrons donated to the electron transport system) ratio were used for all the *E. coli* models and were the same as that in iAF1260 model [24]. When the metabolic models were used to simulate the change of carbon source (e.g., from glucose to succinate), we obtained the corresponding optimal growth rates and flux distributions for all the reactions. If the uptake/secretion flux for a reaction in the optimal flux solution was reduced or increased by over 10% ($\text{flux}.x > 1.1 \times \text{flux}.y$ or $\text{flux}.x < 0.9 \times \text{flux}.y$) between two conditions, we defined the reactions to be associated with the diet stress.

Uniform random sampling of the solution space for *E. coli* metabolic models in any environmental condition is a rapid and scalable way to characterize the structure of the allowed space of metabolic fluxes [25]. The set of flux distributions obtained from sampling can be interrogated further to answer a number of questions related to the metabolic network function. In the study, we studied how dependent two reactions within the *E. coli* network were on each other.

2.5. Flux Variability Analysis (FVA). Biological systems often contain redundancies that contribute to their robustness. FVA can be used to examine these redundancies by calculating the full range of numerical values for each reaction flux in a network [26]. In FVA, the process is carried out by optimizing for a particular objective, while still satisfying the given constraints set on biological systems. In the study, FVA was

applied to determine the ranges of fluxes that correspond to an optimal solution of the *E. coli* models determined through FBA. The maximum value of the objective function is first computed and this value is used with multiple optimizations to calculate the maximum and minimum flux values through each reaction.

3. Results

3.1. E. coli in Human Gut Microbiome. Deep metagenomic sequencing provides us the opportunity to explore the existence of a common set of *E. coli* species in human gut microbiome.

To obtain this goal, we built a nonredundant database of 61 sequenced and well-annotated *E. coli* genomes. After aligning the reads of each human gut microbial sample onto the reference database, we determined the proportion of the genomes covered by the reads (Methods). At a 99% identity threshold and 10-fold coverage (the genomes of *E. coli* strains are 5 M on average), we detected one in all gut samples, three in 80%, and seven in 60% of the 148 human gut samples (Table 1). We focused on the three common *E. coli* strains, including *E. coli* HS, UTI89, and CFT073. Other recent studies support our findings, including studies from human [27] and animal models [28].

Besides the genome-guided methods, the reads were used to perform de novo assembly, which can recover transcript fragments from regions missing in the genome assembly [29]. We first assembled metagenomes in 148 human gut microbiome samples using over 10 billion reads. Then, we mapped the 15 million gut scaffolds to the 293663 genes (target genes) of the 61 *E. coli* genomes in the human gut. At a 99% identity threshold, over 60% of the target genes of the seven *E. coli* in Table 1 had at least 80% of their length covered by a single scaffold, indicating that the genes of these *E. coli* strains were significantly enriched in the gut scaffolds (Fisher's exact test, $P < 10^{-10}$).

3.2. In Silico Metabolic Models of E. coli Strains. We generated genome-wide metabolic network of three common *E. coli* (*E. coli* HS, UTI89, and CFT073) from metabolic model of *E. coli* pan-genome using GIMME and iMAT algorithms.

The results indicate that the metabolic networks obtained with the two algorithms are identical (TEXT S1–S3 available online at http://dx.doi.org/10.1155/2014/694967). We then explored the differences in network properties among the three models. It shows that these models are different in network structure (Figure 1(a), Table S1). For example, compared with *E. coli* CFT073 and *E. coli* UTI89, *E. coli* HS model has 41 specific metabolic reactions catalyzed by 36 genes (Figure 1(b)). These reactions are associated with alternate carbon metabolism, murein recycling, nitrogen metabolism, and inner membrane transport. Most of the reactions tend to form a subnetwork rather than are scattered in an apparently random manner in the metabolic network. We also observed 32 different metabolites not included in all the three models (Table 2). Only three of the metabolites (including allantoate, tRNA-Ala, and tRNA-Phe) can be detected in the human metabolic model Recon2 [30], suggesting that most of these different metabolites are not involved in direct interactions of gut microbiome host. However, some of these metabolites are of importance to strain-specific characteristics and closely related to human-microbe interactions. For example, GDP-L-fucose plays important roles in microbial infection and numerous ontogenic events [31].

The genome-wide metabolic networks for *E. coli* CFT073 and UTI89 have recently been reconstructed based on the comparative genomics analysis [20]. We compared our models (TEXT S1–S3) with the previous ones and found that our models included more metabolic genes because the deep sequencing has been proven to lead to the identification of large populations of novel as well as missing transcripts that might reflect Hydra-specific evolutionary events [32].

3.3. Optimal Flux Distributions for E. coli Strains. In the previous studies, one of the most fundamental genome-scale phenotypic calculations is the simulation of cellular growth using flux balance analysis (FBA) [25]. As a result, we defined biomass composition of the cell as the biomass objective function and performed FBA on the model in order to maximize the objective function. It shows that the optimal biomass flux for the three models are pretty close (optimal flux = 0.7287 for CFTO73, while optimal flux = 0.7367 for HS and UTI89). However, the optimal flux distributions are of different in the networks. Figure 2 shows the optimal flux distribution map of core metabolic network in three *E. coli* strains. It shows that the fluxes of ACS (acetyl-CoA synthetase), PTAr (phosphotransacetylase), and ACKr (acetate kinase) in CFTO73 model are obviously different from that in the other two models.

We then estimated the effect of reducing flux through metabolic reactions on biomass production of three models. Two reactions ACOAD6F (acyl-CoA dehydrogenase, tetradecanoyl-CoA) and PGK (phosphoglycerate kinase) were taken as examples here (Figure S1). It shows that the growth rate is sustained near the optimal value over a range of values for PGK in all three models, indicating the same network robustness with respect to flux changes in the reaction (Figure S1A). However, the effects of reducing flux through ACOAD6F on growth are different between *E. coli* CFTO73 and the other two models (Figure S1B). Besides, the growth rate is sharply reduced after reaching the optimal value in HS and UTI89 models.

3.4. Dynamic Responses of Metabolic Networks to Changes in Carbon Sources. Although a few human gut microbiome projects have been launched, the interrelationships between our diets and the structure and operations of our gut microbial communities are poorly understood. Here, we predicted the human gut *E. coli*'s response to diet using metabolic modeling.

We simulated the optimal growth rates for three models on carbon source as acetate, ethanol, glucose, and succinate, respectively (uptake rate sets all changed to 9 mmol gDW^{-1} h^{-1}). The average growth rates of three metabolic models corresponding to four diet conditions are

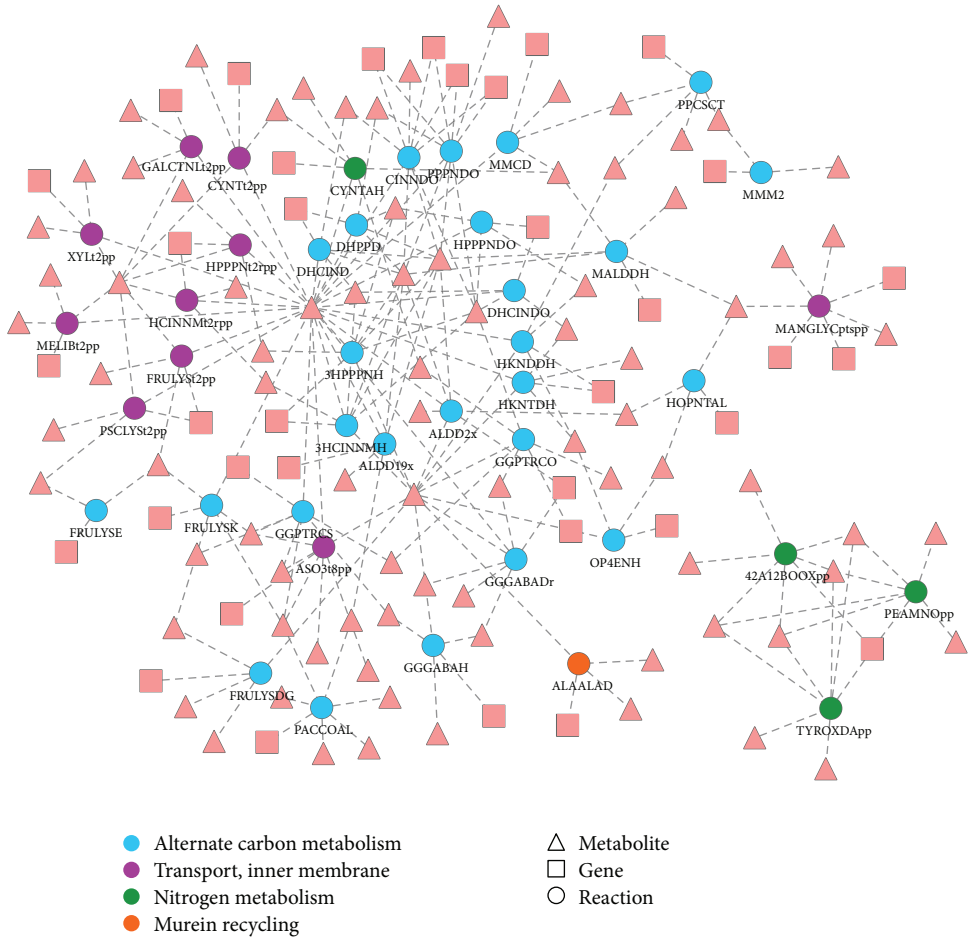
Table 1: Common *Escherichia coli* strains in human gut.

Escherichia coli strains	Samples count[a]	Genome size	Gene counts	Protein count	Genes by *de novo* assembly
E. coli HS	148	4.6 M	4629	4377	3606
E. coli UTI89	134	5.0 M	5127	5017	3435
E. coli CFT073	125	5.2 M	5579	5369	3406
E. coli KO11FL	115	4.9 M	4756	4533	3512
E. coli NA114	94	5.0 M	4975	4873	3381
E. coli 536	90	4.9 M	4779	4619	3488
E. coli O127:H6 str. E2348/69	90	5.0 M	4890	4552	3179

[a]There are 148 individual samples in the analysis.

Strains	Genes	Reactions	Metabolites	Reversible reactions
*E.coli*_CFT073	1149	2226	1621	838
*E.coli*_HS	1223	2330	1646	841
*E.coli*_UTI89	1193	2314	1632	845

(a)



(b)

Figure 1: Comparisons of metabolic networks of three *E. coli* strains. (a) Basic parameters of metabolic models. (b) Strain-specific reactions in *E. coli* HS model.

Table 2: Different metabolites in *E. coli* strains.

Metabolites	Descriptions	Formulas	Charges
4h2opntn	4-Hydroxy-2-oxopentanoate	C5H7O4	−1
acglc-D	6-Acetyl-D-glucose	C8H14O7	0
acmalt	Acetyl-maltose	C14H24O12	0
alatrna	L-Alanyl-tRNA(Ala)	C3H6NOR	1
all6p	D-Allose 6-phosphate	C6H11O9P	−2
alltt	Allantoate	C4H7N4O4	−1
allul6p	Allulose 6-phosphate	C6H11O9P	−2
cechddd	cis-3-(3-Carboxyethyl)-3,5-cyclohexadiene-1,2-diol	C9H11O4	−1
cenchddd	cis-3-(3-Carboxyethenyl)-3,5-cyclohexadiene-1,2-diol	C9H9O4	−1
cinnm	trans-Cinnamate	C9H7O2	−1
dhcinnm	2,3-Dihydroxicinnamic acid	C9H7O4	−1
dhpppn	3-(2,3-Dihydroxyphenyl)propanoate	C9H9O4	−1
dtdp4d6dm	dTDP-4-dehydro-6-deoxy-L-mannose	C16H22N2O15P2	−2
dtdprmn	dTDP-L-Rhamnose	C16H24N2O15P2	−2
frulysp	Fructoselysine phosphate	C12H24N2O10P	−1
gdpddman	GDP-4-Dehydro-6-deoxy-D-mannose	C16H21N5O15P2	−2
gdpfuc	GDP-L-Fucose	C16H23N5O15P2	−2
gdpofuc	GDP-4-oxo-L-Fucose	C16H21N5O15P2	−2
gg4abut	Gamma-glutamyl-gamma aminobutyric acid	C9H15O5N2	−1
ggbutal	Gamma-glutamyl-gamma-butyraldehyde	C9H16O4N2	0
ggptrc	Gamma-glutamyl-putrescine	C9H20O3N3	1
hkndd	2-Hydroxy-6-oxonona-2,4-diene-1,9-dioate	C9H8O6	−2
hkntd	2-Hydroxy-6-ketononatrienedioate	C9H6O6	−2
malt6p	Maltose 6′-phosphate	C12H21O14P	−2
man6pglyc	2(alpha-D-Mannosyl-6-phosphate)-D-glycerate	C9H14O12P	−3
op4en	2-Oxopent-4-enoate	C5H5O3	−1
pac	Phenylacetic acid	C8H7O2	−1
phaccoa	Phenylacetyl-CoA	C29H38N7O17P3S	−4
phetrna	L-Phenylalanyl-tRNA(Phe)	C9H10NOR	1
trnaala	tRNA(Ala)	R	0
trnaphe	tRNA(Phe)	R	0
urdglyc	(-)-Ureidoglycolate	C3H5N2O4	−1

shown in Figure 3(a). We can see that the growth rates for three models are similar in different conditions. Besides, it demonstrates substantially decreased anaerobic growth as compared with aerobic ($18\,\text{mmol}\,\text{gDW}^{-1}\,\text{h}^{-1}$) growth with the same glucose uptake rate, which was supported by recent studies that *E. coli* requires aerobic respiration to compete successfully in the mouse intestine [8, 9]. For *E. coli* strains in human gut, carbon sources are diverse, but glucose is most suitable for their growth.

These responses of *E. coli* to the diet changes involve many metabolic genes and pathways. We explored the perturbations in the metabolic networks and found 10 genes (including *ADH5*, *ALDH5A1*, *DLD*, *FECH*, *GCLC*, *GPT*, *GSR*, *KARS*, *MPST*, and *TST*) closely associated with the diet stress (Figure 3(b)). The glycolysis, gluconeogenesis, and glycerophospholipid metabolism were enriched in the metabolic reactions catalyzed by these genes ($P < 10^{-3}$ using Fisher's exact test). Besides, we found that these enzymes were evolutionarily conserved from *E. coli* to human and were involved in the interactions between human and *E. coli* [14, 33]. Especially, nine out of these 10 genes (except *GPT*, glutamic-pyruvate transaminase) were found to be increased in human obesity [34].

3.5. Analyzing Flux Correlations in Diabetes-Associated Pathways in E. coli Using Sampling. Assessment and characterization of gut microbiota (*E. coli* acts as an integral component) has become a major research area in human type 2 diabetes, the most prevalent endocrine disease worldwide. A recent metagenomic research identified and validated over 400 type-2-diabetes-associated markers in *E. coli*, including over 100 metabolic genes [35]. In the study, we performed uniform random sampling for three models under glucose-limiting aerobic growth conditions to explore the relationships between the diabetes-associated pathways.

We detected 158 metabolic reactions in *E. coli* models that were associated with human type 2 diabetes (Table S2). It shows that these reactions participate in many subsystems,

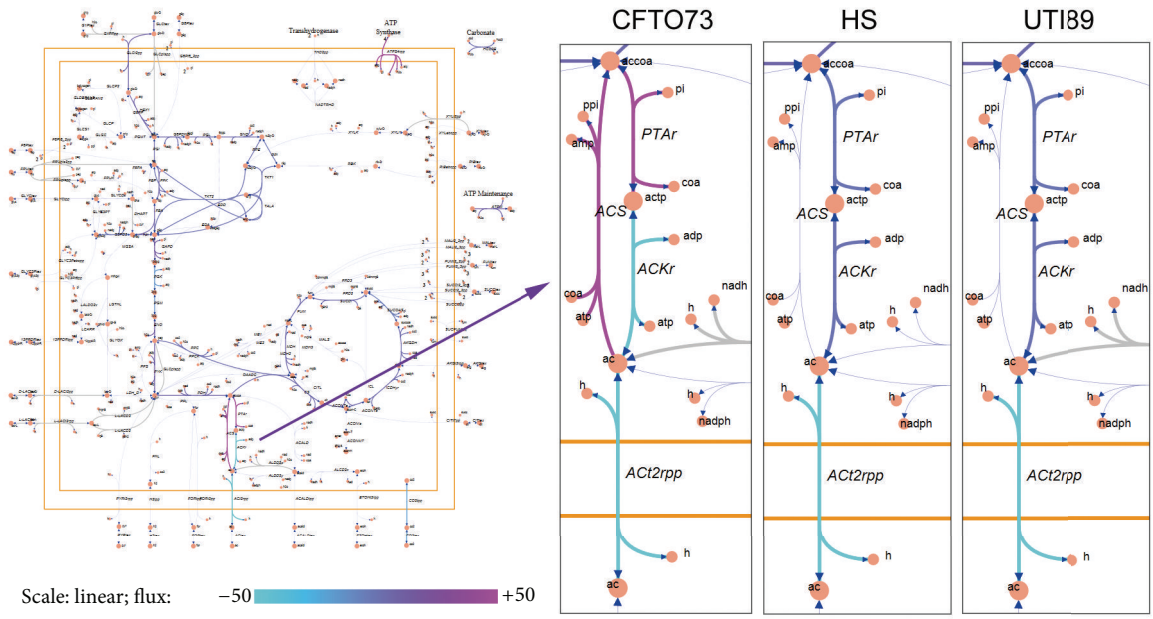


Figure 2: Flux balance analysis of metabolic models. The figure shows the core metabolic map (a) in *E. coli* and the reactions with different fluxes (b) among three *E. coli* models. ACS: acetyl-CoA synthetase; PTAr: phosphotransacetylase; ACKr: acetate kinase.

of which over 30% are associated with lipid metabolism and cofactor/prosthetic group biosynthesis. Correlation ships between some metabolic reactions can be observed in Figure 4. For example, PGL (6-phosphogluconolactonase) and GND (phosphogluconate dehydrogenase) fluxes are positively correlated in the *E. coli* HS model, whereas PGL shows negative correlation with RPI (ribose-5-phosphate isomerase) fluxes. The correlation ships between these diabetes-associated reaction fluxes are the same in other two models.

3.6. Flux Variability Analysis (FVA) of E. coli Models. FBA returns a single flux distribution that corresponds to maximal biomass production under given growth conditions. However, alternate optimal solutions may exist, which correspond to maximal growth. As a result, we performed FVA for the three *E. coli* models under glucose-limited aerobic growth conditions (glucose and oxygen were changed to 10 and 18 mmol gDW^{-1} h^{-1}, resp.).

It shows that the minimum and maximum fluxes for the reactions in *E. coli* models are different. Figure 5 illustrates FVA result for the seven reactions in pyrimidine biosynthesis. All the reactions have different flux range in three networks, especially carbamate kinase and dihydroorotic acid dehydrogenase.

4. Discussion and Conclusion

In our study, we determined the common *E. coli* strains in human gut microbial communities based on HMP datasets. We applied two widely used algorithms (GIMME and iMAT) to reconstruct genome-wide metabolic models for three common *E. coli* strains (*E. coli* HS, UTI89, and CFT073) and compared the network characteristics of these models. These models were then used to predict the cellular phenotypes and dynamic responses to the diverse gut microenvironment. The models were also applied in exploring the relationships between *E. coli* and human diabetes. The results will be helpful in exploring the dynamic responses of gut microbiome to the environmental perturbations.

The *E. coli* strains have been proven to be significantly different among individuals, although the species is abundant in human gut [36]. Although it is well accepted that the composition of *E. coli* strains in human gut flora is associated with health status, the exact molecular mechanism is still unclear. We detected the common *E. coli* strains in human gut and systematically compared their functions through in silico modeling, which have two advantages over the traditional methods. First, the sequencing data allows for a much more accurate determination of microbiome composition. The advent of next-generation sequencing (NGS) enabled several high-profile collaborative projects including the HMP Consortium (http://www.hmpdacc.org/project_catalog.html) and MetaHIT Consortium (http://www.sanger.ac.uk/resources/downloads/bacteria/metahit/), which have released a wide range of data on the human microbiome. Using these datasets, we applied different methods (genome-guided mapping and de novo assembly) to determine the common *E. coli* strains, making the following study of interconnectivity between gut microbiota, diet, and cell molecular responses available. Second, the metabolic modeling can allow us to see how a biological system might respond [37]. This will guide the wet lab experiments and avoid most of the mistakes in the process. In fact, developing computational methods

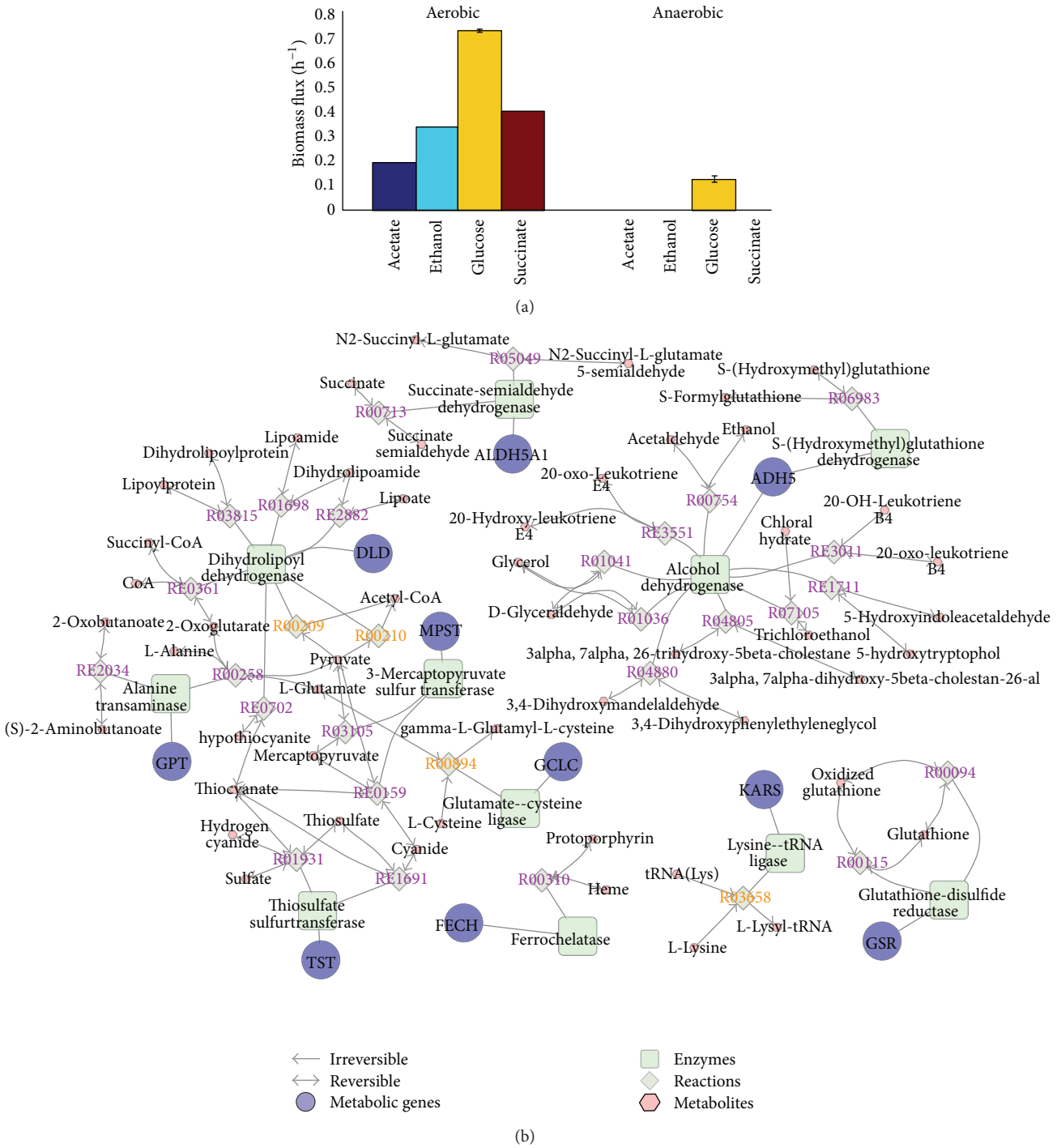


Figure 3: Optimal growth rates for *E. coli* strains on different carbon sources and the associated gene-protein reactions. (a) Optimal growth rates for *E. coli* strains on nutrition sources in human gut. The length of each bar represents the average optimal growth rates for three models on the same carbon source. (b) The diet stress-associated metabolic network in gut *E. coli*.

capable of predicting metabolic flux by integrating these data sources with a metabolic network is a major challenge of systems biology [18]. For example, the predicted behaviors of diabetes-associated reactions in *E. coli* (Table S2) can be integrated with experimental validations to detect the causal genes in human diabetes.

The *E. coli* is regarded as the prototypical pluripotent pathogens capable of causing a wide variety of illnesses in a broad array of species, including pyelonephritis, diarrhea, dysentery, and the hemolytic-uremic syndrome [38]. In particlar, human gut *E. coli* and its relationship to complex diseases, such as cancer [39] and diabetes [40], has attracted

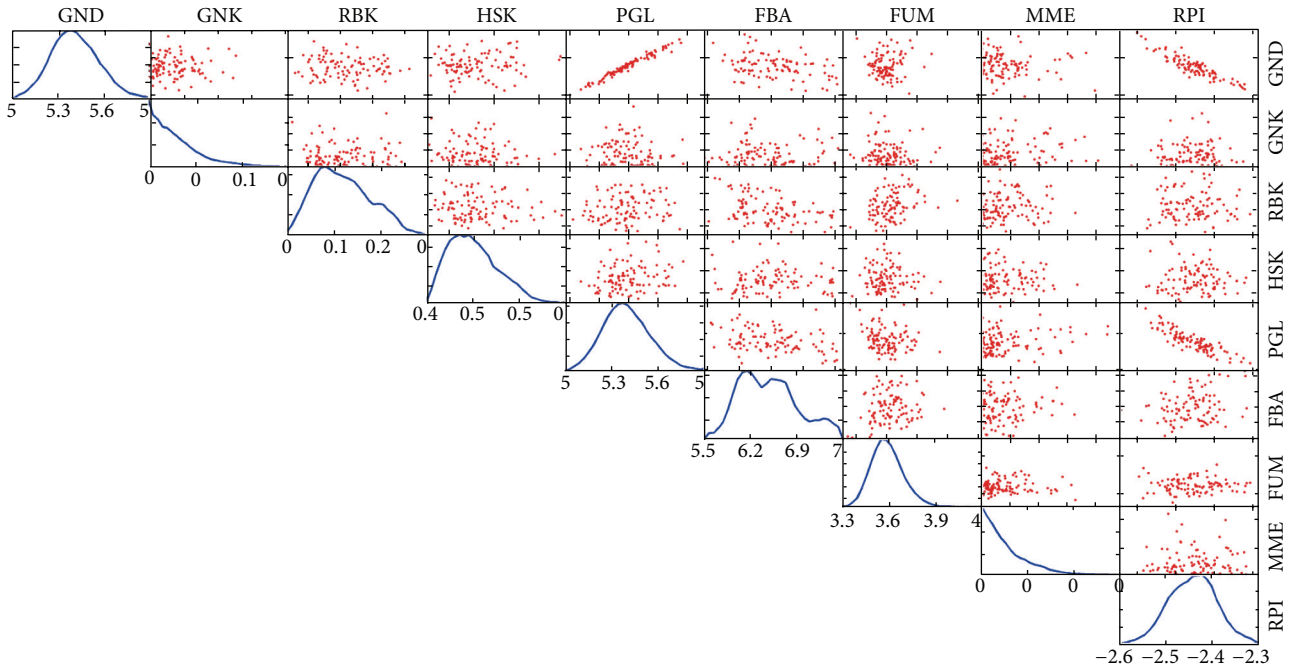
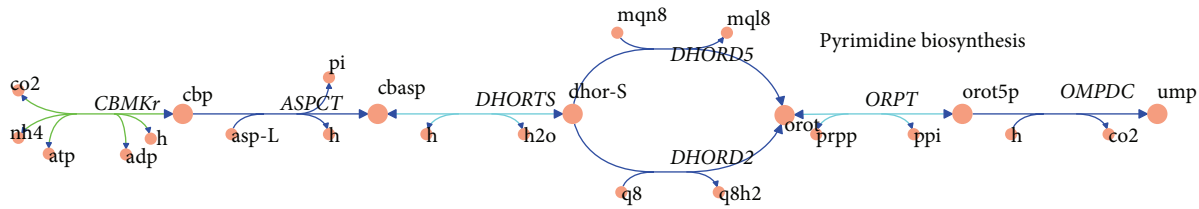


Figure 4: Flux sampling of *E. coli* HS model. Flux distribution histograms (diagonal) and pairwise scatterplots (off-diagonal) for diabetes-associated metabolic reactions in *E. coli* HS model. The *x*-axis of the histograms indicates the magnitude of the flux through the particular reaction. The scatterplots on the off-diagonal elements show the relationship between fluxes through two reactions. GND: phosphogluconate dehydrogenase; CAT: catalase; GNK: gluconokinase; RBK: ribokinase; HSK: homoserine kinase; TMK: thiamine kinase; PGL: 6-phosphogluconolactonase; FBA: fructose-bisphosphate aldolase; FUM: fumarase; MME: methylmalonyl-CoA epimerase; RPI: ribose-5-phosphate isomerase.

increasing interest in the last few years. A question then arises: "How is it possible for this Jekyll and Hyde species to both coexist peacefully with its host and cause devastating illness?" [38]. The answer mainly lies in the existence of different strains of *E. coli* with variable pathogenic potential [41]. However, we can hardly draw a complete picture of how the *E. coli* strains respond physiologically to the complex gut microenvironment. Our study can provide valuable information based on the systematic comparisons of different *E. coli* strains. It shows that although the optimal growth rates are similar for three *E. coli* strains, the optimal flux distributions are different for three models, even in *E. coli* core reactions. The detected different reactions, such as ACS (acetyl-CoA synthetase) and PTAr (phosphotransacetylase) were approved to be involved in the virulence of *E. coli* and be associated with human complex diseases [35, 42]. The results can be integrated with other data sets, such as human clinical trials and virulence profiles, which will help establish the extent of commonality between food-source and human gut *E. coli* [43] and estimate the contribution of strain-specific reactions or genes to infections in humans.

We found that the *E. coli* responded distinctly to different gut diets and the stress-associated genes were closely associated with obesity. With the high prevalence of diet-induced health concerns, such as diabetes and obesity, there remains a need for approaches that treat the causal factors. Among these factors, gut microbiome is drawing more attention [35, 44] for it is suitable as disease markers and drug targets. For example, Qin et al. carried out a metagenome-wide association study which indicated that patients with type 2 diabetes have only moderate intestinal dysbiosis but that butyrate-producing bacteria are less abundant and opportunistic pathogens are more abundant in these individuals than in healthy controls [35]. The underlying mechanisms of interactions between gut microbiome and human health are complicated; however the stress-associated pathways (such as the detected the gluconeogenesis, and glycerophospholipid metabolism) may play important roles in the disease development. The diet changes first induced changes of involved metabolic genes (such as *ADH5*, alcohol dehydrogenase 5), which trigger the downstream signaling pathways. These signaling pathways mainly associated with immune responses and development [44, 45]. It is commonly accepted that the gut microbiota interacts with the immune system, providing signals to promote the maturation of immune cells and the normal development of immune functions [46]. The dynamic interactions between all components of the microbiota and host tissue over time will be crucial for building predictive models for diagnosis and treatment of diseases linked to imbalances in our microbiota.

In summary, the findings here represent a significantly expanded and comprehensive reconstruction of the *E. coli* metabolic network in human gut. This work will enable a wider spectrum of studies focused on microbe-host interactions and serve as a means of integrating other omics sets in systems biology.



(a)

	CFTO73		HS		UTI89	
Reactions	minFlux	maxFlux	minFlux	maxFlux	minFlux	maxFlux
CBMKr	−4.15	2.20	−4.26	2.26	−7.24	3.20
ASPCT	0.25	1.96	0.26	2.03	0.22	1.08
DHORTS	−1.98	−0.25	−2.03	−0.26	−1.08	−0.22
DHORD2	0	1.98	0	2.03	0	1.08
DHORD5	0	1.98	0	2.03	0	1.08
ORPT	−1.96	−0.25	−2.03	−0.26	−1.08	−0.22
OMPDC	0.25	1.98	0.26	2.03	0.22	1.08

Bidirectional/reversible
Unidirectional/reversible
reverse
Unidirectional/irreversible

(b)

Figure 5: FVA of *E. coli* models. Shown is a map of metabolic reactions in pyrimidine biosynthesis pathway of *E. coli* models. Using FVA, the minimum (min) and maximum (max) allowable flux values for each reaction were determined. The values shown in the table correspond to the min and max allowable fluxes for each reaction shown in the map. The results were further characterized by the direction of predicted flux (bidirectional or unidirectional) computed using FVA. The full names of the metabolic reactions are included in TEXT S1–S3.

Conflict of Interests

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

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant no. 31123005) and the Instruments Function Deployment Foundation of CAS (Grants nos. yg2010044 and yg2011057).

References

[1] P. B. Eckburg, E. M. Bik, C. N. Bernstein et al., "Microbiology: diversity of the human intestinal microbial flora," *Science*, vol. 308, no. 5728, pp. 1635–1638, 2005.

[2] G. Reid, J. Howard, and B. S. Gan, "Can bacterial interference prevent infection?" *Trends in Microbiology*, vol. 9, no. 9, pp. 424–428, 2001.

[3] D. McCloskey, B. Ø. Palsson, and A. M. Feist, "Basic and applied uses of genome-scale metabolic network reconstructions of *Escherichia coli*," *Molecular Systems Biology*, vol. 9, no. 1, article 661, 2013.

[4] S. Nakamura, T. Nakaya, and T. Iida, "Metagenomic analysis of bacterial infections by means of high-throughput DNA sequencing," *Experimental Biology and Medicine*, vol. 236, no. 8, pp. 968–971, 2011.

[5] M. S. Donia and M. A. Fischbach, "Dyeing to learn more about the gut microbiota," *Cell Host and Microbe*, vol. 13, no. 2, pp. 119–120, 2013.

[6] L. Rup, "The human microbiome project," *Indian Journal of Microbiology*, vol. 52, no. 3, p. 315, 2012.

[7] P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, and J. I. Gordon, "The human microbiome project," *Nature*, vol. 449, no. 7164, pp. 804–810, 2007.

[8] S. A. Jones, F. Z. Chowdhury, A. J. Fabich et al., "Respiration of Escherichia coli in the mouse intestine," *Infection and Immunity*, vol. 75, no. 10, pp. 4891–4899, 2007.

[9] S. E. Winter, M. G. Winter, M. N. Xavier et al., "Host-derived nitrate boosts growth of E. coli in the inflamed gut," *Science*, vol. 339, no. 6120, pp. 708–711, 2013.

[10] I. Cho and M. J. Blaser, "The human microbiome: at the interface of health and disease," *Nature Reviews Genetics*, vol. 13, no. 4, pp. 260–270, 2012.

[11] C. Huttenhower, D. Gevers, R. Knight et al., "Structure, function and diversity of the healthy human microbiome," *Nature*, vol. 486, no. 7402, pp. 207–214, 2012.

[12] C. F. Maurice, H. J. Haiser, and P. J. Turnbaugh, "Xenobiotics shape the physiology and gene expression of the active human gut microbiome," *Cell*, vol. 152, no. 1-2, pp. 39–50, 2013.

[13] A. Heinken, S. Sahoo, R. M. Fleming, and I. Thiele, "Systems-level characterization of a host-microbe metabolic symbiosis in the mammalian gut," *Gut Microbes*, vol. 4, no. 1, pp. 28–40, 2013.

[14] S. Shoaie, F. Karlsson, A. Mardinoglu, I. Nookaew, S. Bordel, and J. Nielsen, "Understanding the interactions between bacteria in the human gut through metabolic modeling," *Scientific Reports*, vol. 3, article 2532, 2013.

[15] C. Lozupone, K. Faust, J. Raes et al., "Identifying genomic and metabolic features that can underlie early successional and opportunistic lifestyles of human gut symbionts," *Genome Research*, vol. 22, no. 10, pp. 1974–1984, 2012.

[16] Y. Q. Zhao and J. F. Huang, "Reconstruction and analysis of human heart-specific metabolic network based on transcriptome and proteome data," *Biochemical and Biophysical Research Communications*, vol. 415, no. 3, pp. 450–454, 2011.

[17] J. Zhao, C. Geng, L. Tao et al., "Reconstruction and analysis of human liver-specific metabolic network based on CNHLPP data," *Journal of Proteome Research*, vol. 9, no. 4, pp. 1648–1658, 2010.

[18] K. Yizhak, T. Benyamini, W. Liebermeister, E. Ruppin, and T. Shlomi, "Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model," *Bioinformatics*, vol. 26, no. 12, pp. i255–i260, 2010.

[19] B. A. Methe, K. E. Nelson, M. Pop et al., "A framework for human microbiome research," *Nature*, vol. 486, no. 7402, pp. 215–221, 2012.

[20] D. J. Baumler, R. G. Peplinski, J. L. Reed, J. D. Glasner, and N. T. Perna, "The evolution of metabolic networks of E. coli," *BMC Systems Biology*, vol. 5, article 182, 2011.

[21] S. A. Becker and B. O. Palsson, "Context-specific metabolic networks are consistent with experiments," *PLoS Computational Biology*, vol. 4, no. 5, Article ID e1000082, 2008.

[22] H. Zur, E. Ruppin, and T. Shlomi, "iMAT: an integrative metabolic analysis tool," *Bioinformatics*, vol. 26, no. 24, pp. 3140–3142, 2010.

[23] J. D. Orth, I. Thiele, and B. O. Palsson, "What is flux balance analysis?" *Nature Biotechnology*, vol. 28, no. 3, pp. 245–248, 2010.

[24] A. M. Feist, C. S. Henry, J. L. Reed et al., "A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information," *Molecular Systems Biology*, vol. 3, article 121, 2007.

[25] S. A. Becker, A. M. Feist, M. L. Mo, G. Hannum, B. Ø. Palsson, and M. J. Herrgard, "Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox," *Nature Protocols*, vol. 2, no. 3, pp. 727–738, 2007.

[26] R. Mahadevan and C. H. Schilling, "The effects of alternate optimal solutions in constraint-based genome-scale metabolic models," *Metabolic Engineering*, vol. 5, no. 4, pp. 264–276, 2003.

[27] S. L. Chen, M. Wu, J. P. Henderson et al., "Genomic diversity and fitness of *E. coli* strains recovered from the intestinal and urinary tracts of women with recurrent urinary tract infection," *Science Translational Medicine*, vol. 5, no. 184, 2013.

[28] J. S. Ayres, N. J. Trinidad, and R. E. Vance, "Lethal inflammasome activation by a multidrug-resistant pathobiont upon antibiotic disruption of the microbiota," *Nature Medicine*, vol. 18, no. 5, pp. 799–806, 2012.

[29] P. Jain, N. M. Krishnan, and B. Panda, "Augmenting transcriptome assembly by combining de novo and genome-guided tools," *PeerJ*, vol. 1, article e133, 2013.

[30] I. Thiele, N. Swainston, R. M. T. Fleming et al., "A community-driven global reconstruction of human metabolism," *Nature Biotechnology*, vol. 31, no. 5, pp. 419–425, 2013.

[31] D. J. Becker and J. B. Lowe, "Fucose: biosynthesis and biological function in mammals," *Glycobiology*, vol. 13, no. 7, pp. 41R–53R, 2003.

[32] Y. Wenger and B. Galliot, "RNAseq versus genome-predicted transcriptomes: a large population of novel transcripts identified in an Illumina-454 Hydra transcriptome," *BMC Genomics*, vol. 14, no. 1, article 204, 2013.

[33] Z. Qi and M. R. O'Brian, "Interaction between the bacterial iron response regulator and ferrochelatase mediates genetic control of heme biosynthesis," *Molecular Cell*, vol. 9, no. 1, pp. 155–162, 2002.

[34] S. Greenblum, P. J. Turnbaugh, and E. Borenstein, "Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 2, pp. 594–599, 2012.

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

[36] M. Martinez-Medina, X. Aldeguer, M. Lopez-Siles et al., "Molecular diversity of *Escherichia coli* in the human gut: new ecological evidence supporting the role of adherent-invasive *E. coli* (AIEC) in Crohn's disease," *Inflammatory Bowel Diseases*, vol. 15, no. 6, pp. 872–882, 2009.

[37] J. R. Karr, J. C. Sanghvi, D. N. MacKlin et al., "A whole-cell computational model predicts phenotype from genotype," *Cell*, vol. 150, no. 2, pp. 389–401, 2012.

[38] M. Donnenberg, *Escherichia coli: Pathotypes and Principles of Pathogenesis*, Academic Press, New York, NY, USA, 2013.

[39] H. Tlaskalová-Hogenová, R. Tpánková, H. Kozáková et al., "The role of gut microbiota (commensal bacteria) and the mucosal barrier in the pathogenesis of inflammatory and autoimmune diseases and cancer: Contribution of germ-free and gnotobiotic animal models of human diseases," *Cellular and Molecular Immunology*, vol. 8, no. 2, pp. 110–120, 2011.

[40] N. Larsen, F. K. Vogensen, F. W. J. van den Berg et al., "Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults," *PLoS ONE*, vol. 5, no. 2, Article ID e9085, 2010.

[41] R. M. Robins-Browne, "Traditional enteropathogenic *Escherichia coli* of infantile diarrhea," *Reviews of Infectious Diseases*, vol. 9, no. 1, pp. 28–53, 1987.

[42] K. Guan and Y. Xiong, "Regulation of intermediary metabolism by protein acetylation," *Trends in Biochemical Sciences*, vol. 36, no. 2, pp. 108–116, 2011.

[43] J. R. Johnson, M. A. Kuskowski, K. Smith, T. T. O'Bryan, and S. Tatini, "Antimicrobial-resistant and extraintestinal pathogenic *Escherichia coli* in retail foods," *Journal of Infectious Diseases*, vol. 191, no. 7, pp. 1040–1049, 2005.

[44] J. C. Clemente, L. K. Ursell, L. W. Parfrey, and R. Knight, "The impact of the gut microbiota on human health: an integrative view," *Cell*, vol. 148, no. 6, pp. 1258–1270, 2012.

[45] P. D. Gluckman, K. A. Lillycrop, M. H. Vickers et al., "Metabolic plasticity during mammalian development is directionally dependent on early nutritional status," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 31, pp. 12796–12800, 2007.

[46] J. Chow, S. M. Lee, Y. Shen, A. Khosravi, and S. K. Mazmanian, "Host-bacterial symbiosis in health and disease," *Advances in Immunology*, vol. 107, pp. 243–274, 2010.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 931825, 10 pages
http://dx.doi.org/10.1155/2014/931825

Research Article

Integrated Analysis Identifies Interaction Patterns between Small Molecules and Pathways

Yan Li,[1] Weiguo Li,[1] Xin Chen,[2] Hong Jiang,[1] Jiatong Sun,[1] Huan Chen,[1] and Sali Lv[1]

[1] *Department of Bioinformatics, School of Basic Medical Sciences, Nanjing Medical University, Nanjing 210029, China*
[2] *Faculty of Health Sciences, University of Macau, Macau*

Correspondence should be addressed to Sali Lv; lvsali@njmu.edu.cn

Received 1 March 2014; Revised 13 May 2014; Accepted 22 May 2014; Published 13 July 2014

Academic Editor: Siyuan Zheng

Previous studies have indicated that the downstream proteins in a key pathway can be potential drug targets and that the pathway can play an important role in the action of drugs. So pathways could be considered as targets of small molecules. A link map between small molecules and pathways was constructed using gene expression profile, pathways, and gene expression of cancer cell line intervened by small molecules and then we analysed the topological characteristics of the link map. Three link patterns were identified based on different drug discovery implications for breast, liver, and lung cancer. Furthermore, molecules that significantly targeted the same pathways tended to treat the same diseases. These results can provide a valuable reference for identifying drug candidates and targets in molecularly targeted therapy.

1. Introduction

Recently, with the development of molecular biology techniques, molecularly targeted therapy has been applied in clinical practice [1, 2]. In cancer research, molecularly targeted therapy aims to identify the agent for a known therapeutic target. The agent can modify the expression or activity of the target during the growth and progression of the cancer [3]. Unfortunately, the set of drug candidates must be determined by rigorous and repeated experiments [4], a process that is beset with difficulties and is usually time consuming.

Small molecules act by simultaneously participating in multiple biological processes and triggering a variety of changes that lead to diverse reactions. A phenotype is always caused by a series of complex molecular reactions. A pathway embodies complex interactions between small molecules and macromolecules, and most pathways are interrelated [5, 6]. Thus, it is of great biological importance to detect the links between small molecules and their target pathways and to perceive the intervention effects of small molecules on disease through these pathways [7]. In addition, the concept of biochemical pathways aids in understanding the mechanisms of cancer [8]. Considering a pathway as a functional unit facilitates the unravelling of the mode of action for small molecules [9]. A number of studies have reported drug targeted pathway that was the effective therapeutic approach in treating cancer [10–15]. For example, a hedgehog pathway inhibitor, vismodegib, has recently been approved by the US FDA for the treatment of skin cancer, while several drug candidates for the Wnt pathway are entering clinical trials [12]. Azole drugs, which are commonly used in infection treatment, play a part in azole therapy by targeting the sterol biosynthetic pathway [11]. The findings of Chian and his colleagues demonstrated that luteolin inhibits the NRF2 pathway in vivo and can serve as an adjuvant in the chemotherapy of NSCLC [10]. Collins and Workman reported that there were several kinds of potential drug targets: oncogene products downstream, proteins in a key pathway and oncogenic support processes [16]. Disease-related pathways that are affected by the intervention of small molecules are more likely to include target genes [17]. Therefore, employing computational methods to explore the links between small molecules and pathways provides a new perspective for molecularly targeted therapy. With the ongoing research into genome, proteome, and

transcriptome, various databases for small molecules and pathways emerge one after another, such as Connectivity Map [18, 19], DrugBank [20, 21], CTD [22], KEGG [23], and the NCBI PubChem [24]. These databases provide abundant data resources for high-throughput analysis, and this availability allows the creation of a computational method to construct the links between small molecules and their target pathways, which can provide complementary and supporting evidence to experimental studies.

Based upon the above considerations, we have proposed a novel method to detect the links between small molecules and pathways. First, differentially expressed genes related to diseases were identified and enriched into KEGG pathways. Next, molecules that target each pathway were identified by Connectivity Map. We further constructed a link map between the molecules and their target pathways and analyzed the topological features of the link map. Moreover, by applying this method to a chosen set of data, we identified three link patterns. We also found that if molecules significantly targeted the same pathways, then they tended to treat the same disease. Besides, we provide potential candidate for drug experiment by mining and predicting medicinal small molecules and their target pathway in in silico method. This application may provide valuable information to molecularly targeted therapy from a pathway-based perspective.

2. Materials and Methods

2.1. Data Sources. The microarray data were downloaded from Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/, accession number GSE5364). In this study, we used three datasets including both tumour tissues and adjacent normal tissues (tissue type/tumour/normal: breast/183/13, liver/9/8, and lung/18/12). In addition, another dataset was added into each type of cancer to establish link map as well as to predict target relationship. The data included both tumour tissues and adjacent normal tissues (tissue type/tumour/normal: breast (GSE15852)/43/43, liver (GSE9166)/15/18, and lung (GSE7670)/27/27).

Connectivity Map (http://www.connectivitymap.org/cmap/) consisted of more than 7,000 gene expression profiles treated with 1,309 small molecules. These expression profiles represented about 6,000 instances, each of which comprised a treatment and vehicle pair. By comparing the expression pattern similarity of the input genes and the genes perturbed in Connectivity Map instances, a list of molecules related to the input genes would be identified.

2.2. Screening of Differentially Expressed Genes. For each dataset, probes corresponding to more than one gene were discarded. Log_2 transformation of the expression value was performed for each probe, and the data was normalized by the quantile normalization method. Up- and downregulated probes were determined according to their fold difference [25]. All probes were then mapped to Entrez gene IDs using mean values. Differentially expressed genes were screened by significance analysis of microarrays (SAM) [26] with FDR = 0.001.

2.3. Detection of Significant Links between Small Molecules and Pathways. The differentially expressed genes from GSE5364 were annotated into KEGG pathways using WebGestalt [27] (http://bioinfo.vanderbilt.edu/webgestalt/). When it came to data used to predict target relationship, we manipulated DAVID to annotate pathways and carried out corresponding enrichment analysis. Then, statistically enriched pathways, which are potentially relevant to diseases, were obtained using hypergeometric test ($P \leq 0.05$). Differentially expressed genes were partitioned into enriched KEGG pathways. For each pathway, up- and downregulated probes corresponding to differential genes were input into Connectivity Map. The small molecules that were significantly related to a certain pathway were identified according to P value from the permuted results given by Connectivity Map. In this way, small molecules for all the enriched pathways could be found. Thus, we could construct an adjacency matrix between the small molecules and the pathways. The elements of the matrix were set to one when $P \leq 0.01$; otherwise, they would be set to zero. Suppose the adjacency matrix was A. When $A_{ij} = 1$, then that small molecule i significantly targets pathway j. A bipartite M-P network of molecules and pathways can be constructed based on the adjacency matrix. The nodes of the M-P network were small molecules and pathways, respectively. A link was placed between a molecule and a pathway if the molecule significantly targets the pathway. The M-P network was also called the M-P link map. Figure 1 shows the process of constructing the M-P link map.

2.4. Detect the Links Where a Single Molecule Robustly Targets a Single Pathway from the M-P Link Map. Two types of unipartite links were derived from the M-P link map: molecule-molecule links and pathway-pathway links. Such a relationship between two molecules will exist if the two molecules significantly target the same pathways, and such a relationship between two pathways will exist if the two pathways were significantly targeted by the same molecules. The two unipartite links were inferred by the cumulative hypergeometric distribution [28]. The details were as follows.

We suppose that two different molecules target n_1 and n_2 pathways. The size of the intersection and the union of their target pathways were share and n, respectively. Then, the significance level of the two molecules targeting the same pathways was calculated as follows:

$$P = 1 - \sum_{i=0}^{\text{share}-1} p_i, \tag{1}$$

where

$$p_i = \frac{\binom{n_2}{i}\binom{n-n_2}{n_1-i}}{\binom{n}{n_1}}, \quad i = 0, 1, 2, \ldots \text{share} - 1. \tag{2}$$

Similarly, the significance level of two pathways targeted by the same molecules could be calculated. We further extracted the links where a single molecule robustly targets a single pathway from the original M-P link map using molecule-molecule links and pathway-pathway links. Robust links between molecules and pathways were defined

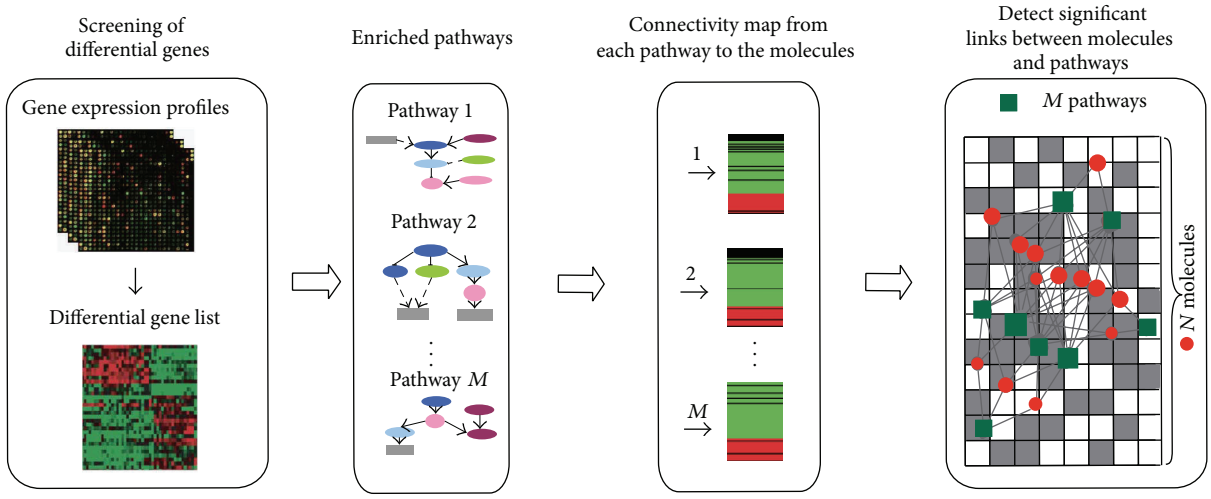


Figure 1: Flow chart for the construction of a link map between molecules and their target pathways. First, the gene expression profiles for each cancer were obtained from both tumours and adjacent normal tissues and differentially expressed genes were screened. Second, enriched pathways were obtained by enriching differentially expressed genes into KEGG pathways. Third, differential genes in each enriched pathway were input into Connectivity Map to identify the small molecules related to each pathway. Finally, link map between small molecules and pathways (the M-P link map) was constructed based on the significance level of their association. In the background grid, the corresponding cell is coloured grey if small molecules significantly link to the pathway in the M-P link map, while the cell was white if such links did not exist. The front network was a visualisation of the background grid in which each pathway node was represented by a green rectangle, while each molecule node was represented by a red circle. The node size increased with the number of first-order neighbours in the link map.

as follows: small molecule A and some other molecules significantly target pathway B; at the same time, pathway B and some other pathways were significantly targeted by molecule A.

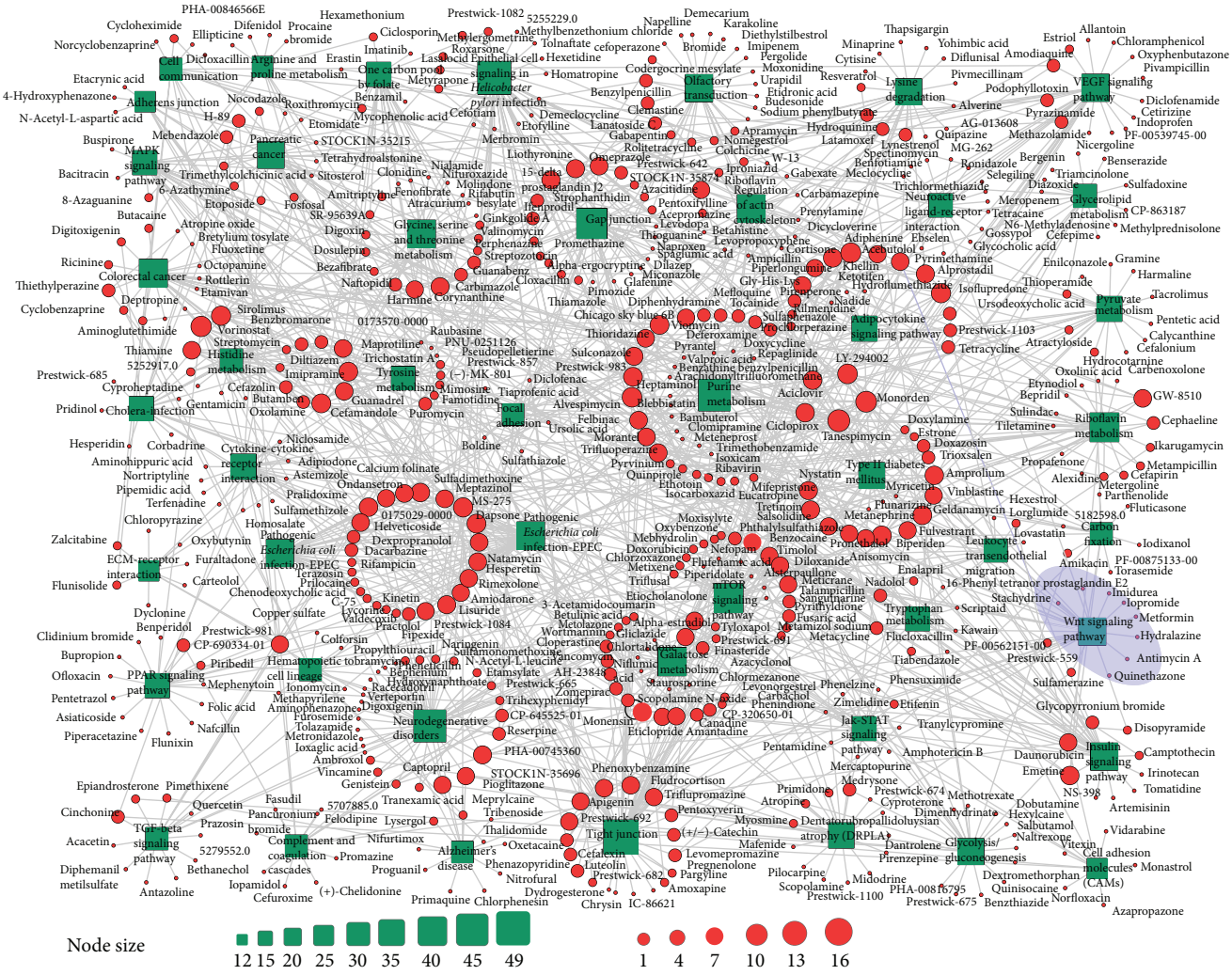
3. Results

3.1. The Topological Characteristics of the M-P Link Map. We proposed a method to construct a link map between small molecules and pathways (M-P link map), as described in Section 2. The method was applied to analyze breast cancer, liver cancer, and lung cancer, as described in the Materials and Methods. M-P link maps for these three cancers from GSE5364 were constructed. The M-P link map for breast cancer and its degree distribution are shown in Figure 2. The heatmap for M-P link map of breast cancer is shown in Figure 4(a). The M-P link maps and degree distributions for liver and lung cancers are shown in Figures S1 and S2, respectively (see Figures S1 and S2 in Supplementary Material available online at http://dx.doi.org/10.1155/2014/931825). Their corresponding heatmaps are shown in Figures S3(a) and S4(a).

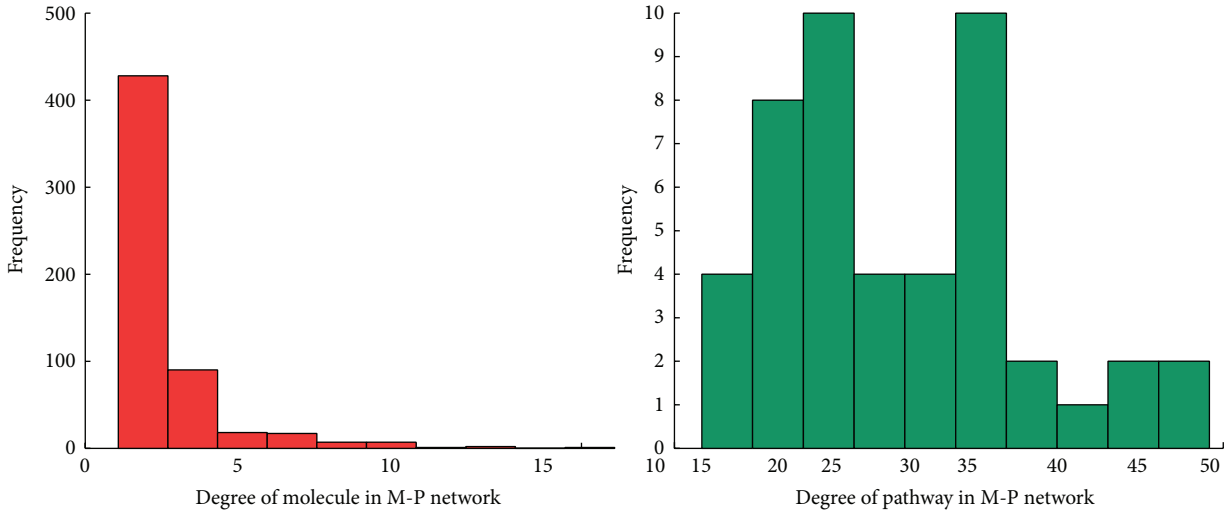
For the M-P link map in Figure 2, the degree of molecules followed a power law distribution. Relatively few molecules were of high degree, targeting multiple pathways. Instead, the majority of small molecules were of low degrees because they target few pathways. Of the 571 small molecules in the M-P link map for breast cancer, eight (1.4%) molecules had degrees larger than ten, while 300 (52.5%) molecules have degree one, meaning that each molecule in the latter category links to only one pathway. Of the 263 small molecules in the M-P link map for liver cancer, only one (0.38%) molecule had the highest degree eight, and 183 (69.6%) molecules had degree one. Of the 276 small molecules in the M-P link map for lung cancer, only one (0.36%) molecule had degree eight, and 196 (71.0%) molecules had degree one.

The degree of the pathways did not show a similar distribution to the degree distribution for small molecules. There were no significant differences in the number of pathways over various degrees. The average degree of the pathways was higher than that of the small molecules. Additionally, the degrees of the pathways were all larger than one. This result indicated that one pathway might not be specifically targeted by one small molecule and was consistent with the biological fact that one pathway might be targeted by distinct small molecules through different genes in the pathway. Figure 3 depicts an instance of gap junction, which turned out to be target pathways of amiodarone, L-DOPA, and lisuride. By directly interfering ADRB1, DRD1, DRD2, and HTR2 of the pathway, amiodarone, L-DOPA, and lisuride established target relationships with gap junction pathway. Here, gap junction partially manifests the target protein-related section rather than the whole pathway.

3.2. The Link Patterns in the M-P Link Map. Based on the M-P link map, there are several link patterns that help reveal the intervention features of molecules on their target pathways. The first pattern is that a single molecule targets multiple pathways (sM-mP). Molecules from this pattern can intervene in multiple pathways. Most tumours involve the regulation of multiple genes and processes. There are usually many potential targets in these tumours [29]. All cell signal transductions cannot be blocked by inhibiting a



(a)



(b)

Figure 2: Visualisation of the M-P link map for breast cancer and its degree distribution. (a) The red circles and green rectangles correspond to the small molecules and pathways, respectively. Node size is proportional to the degree of the node. (b) The degree distribution of small molecules and pathways.

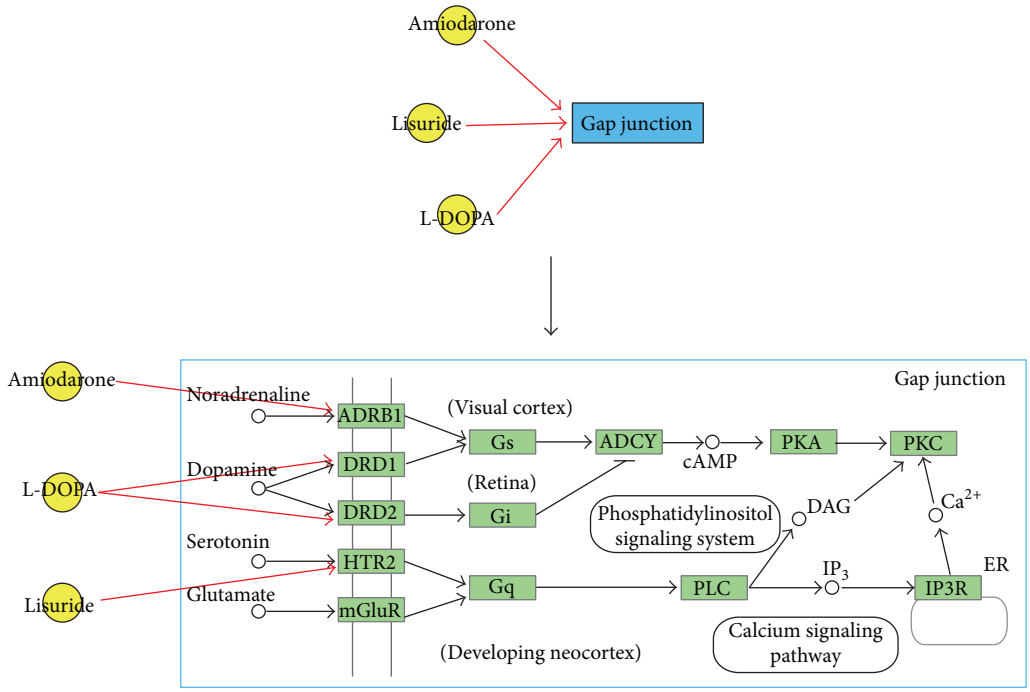


FIGURE 3: The case of three small molecules amiodarone, L-DOPA, and lisuride targeted gap junction. The yellow points represent small molecules, while the blue box represents gap junction pathway. The green boxes represent the genes in gap junction.

single receptor or target. Therefore, molecules with the sM-mP pattern may be considered as antineoplastic candidates, but they may cause many side effects. For breast cancer from GSE5364, we found that molecules with the sM-mP pattern, which had a high degree, could be used for cancer treatment. Table 1 provides detailed descriptions for eight small molecules with degrees not less than ten. Among these molecules, LY-294002 is a potent PI3K inhibitor. PI3K is related to human tumourigenesis, including breast cancer [30]. Vorinostat has also been shown to correlate with breast neoplasm. Tanespimycin, monorden, alvespimycin, and monensin are directly related to the occurrence of cancer. For example, vorinostat targets 12 pathways, which are marked with a green box at the left in Figure 4(a). The descriptions of high degree small molecules of the other two sets of data can be found in Tables S1 and S2.

The second link pattern is where a single molecule targets a single pathway robustly (sM-sP), which can be extracted from the original M-P link map using molecule-molecule links and pathway-pathway links (as described in Section 2). Disease-specific links may be identified in sM-sP links. In the case of breast cancer from GSE5364, the link between LY294002 and the VEGF signaling pathway belongs to the sM-sP pattern. Figure 4(b) shows the possible mechanism through which LY294002 intervenes in the VEGF signalling pathway. LY-294002 is a potent PI3K inhibitor. PI3K, a target gene of LY294002, is involved in the VEGF signaling pathway [31]. The VEGF signaling pathway is closely related to angiogenesis, and PI3K is a key regulator of this biological process. PI3K is related to human tumourigenesis, including breast cancer, lung cancer, melanoma, and lymphoma. LY-294002 may first target PI3K, which is located at the upstream of the VEGF signaling pathway, thereby leading to the differential expression of downstream genes. Another example of an sM-sP link is between isoniazid and the glutathione metabolism pathway in the link map for liver cancer from GSE5364 (Figure S3(b)). Isoniazid is an antibacterial agent used primarily as a tuberculostatic. It remains a preferred choice for the treatment of tuberculosis. ABAT, a target of isoniazid, is involved in glutamate metabolism. According to KEGG pathway, glutamate metabolism is part of glutathione metabolism. A third example of an sM-sP link is between alsterpaullone and the insulin signaling pathway in the link map for lung cancer from GSE5364 (Figure S4(b)). Alsterpaullone is usually involved in protein kinase activity [32]. GSK3B, which is one of its target genes, can inhibit tumourigenesis and tumour diffusion. The GSK3B enzyme also plays crucial roles in various tumours, including breast, colon, kidney, and stomach cancers. It has been examined as a potential target in cancer therapy [33]. If the GSK3B enzyme is inactivated, the risk of occurrence of liver cancer will increase [34].

The third link pattern is where multiple molecules target a single pathway (mM-sP). In this pattern, one pathway is targeted by many molecules that do not link to any other pathways. Molecules in this pattern have few side effects. Therefore, these small molecules may be candidate adjuvant drugs for anticancer agents. Additionally, parts of the molecules tend to have similar efficacy. For the first instance, there are a total of 31 small molecules that target the Wnt

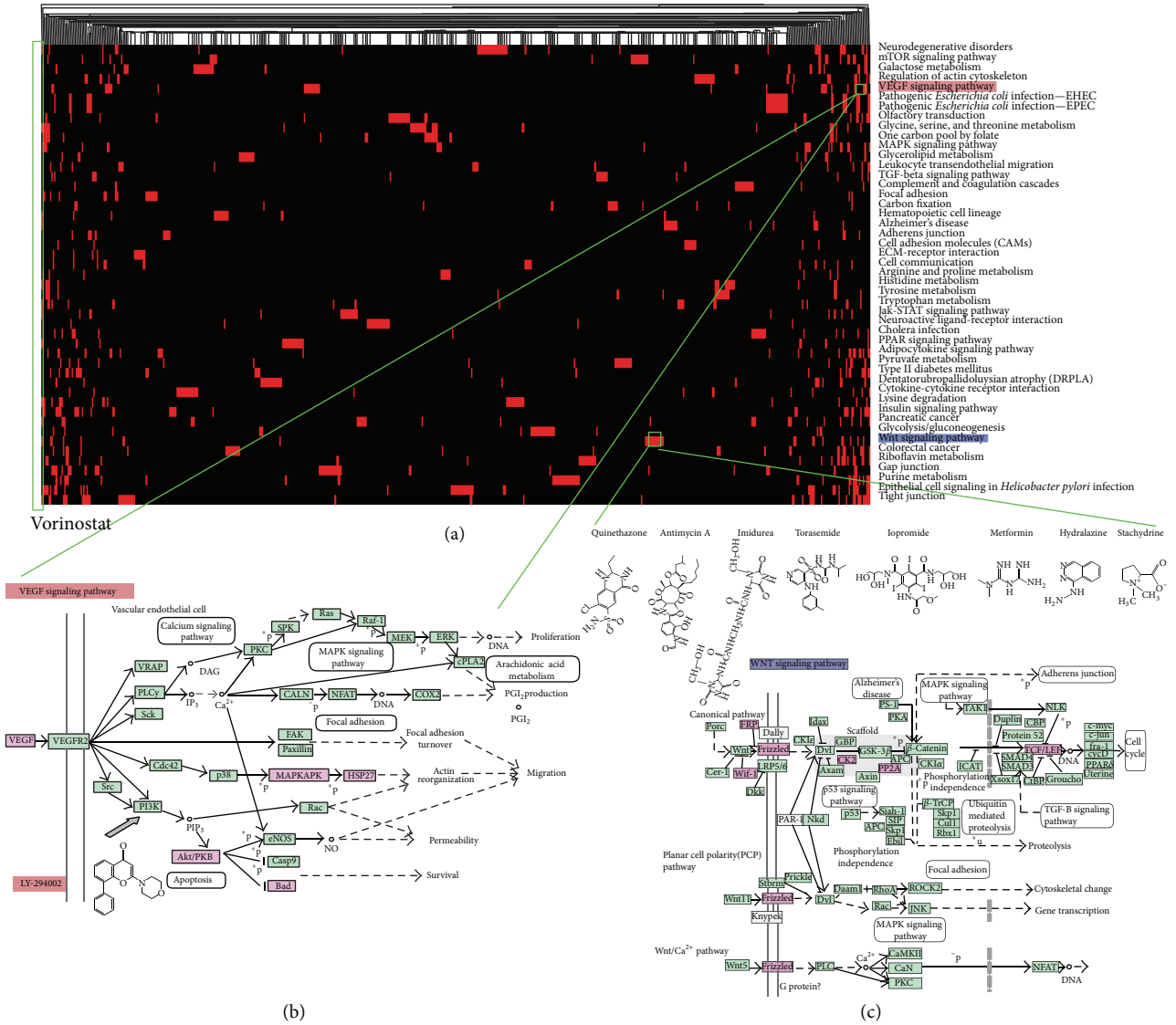


Figure 4: Hierarchical clustering in the M-P link map for breast cancer from GSE5364. (a) Hierarchical clustering between 571 small molecules and 47 metabolic pathways. The corresponding cells are coloured red, where small molecules link to the pathways in the M-P link map. The labels for the corresponding pathways are shown on the right of the figure. (b) Zoomed-in plot of an sM-sP link between LY-294002 and the VEGF signaling pathway. The gene indicated by the arrow is the drug target of LY-294002. Differentially expressed genes in this pathway are coloured pink, while other genes in green are human disease-related genes. (c) Zoomed-in plot of mM-sP links between eight small molecules and the Wnt signaling pathway. These eight small molecules target only the Wnt signaling pathway. The differentially expressed genes in this pathway are coloured pink, while other genes in green are human disease-related genes.

Table 1: The characteristics of molecules with the sM-mP pattern in breast cancer.

Small molecule	Degree in M-P link map	Description of the small molecule
Tanespimycin	16	Antineoplastic antibiotic, HSP90 inhibitor
Monorden	13	HSP90 inhibitor, DNA topoisomerase VI inhibitor
Vorinostat	12	Treats cutaneous T cell lymphoma and breast neoplasm
Adiphenine	11	Antispasmodic agent
LY-294002	10	PI3 kinase inhibitor
Alvespimycin	10	Antineoplastic antibiotic, HSP90 inhibitor
Monensin	10	Blocks intracellular protein transport and exhibits antibiotic and antimalarial efficacy
Biperiden	10	Unknown

signaling pathway in breast cancer from GSE5364. There are eight molecules with degree one, which are shaded light slate blue in Figure 2(a). These eight molecules are stachydrine, torasemide, quinethazone, hydralazine, imidurea, antimycin, iopromide, and metformin. The first four are used to treat hypertension and cardiovascular disease. Details of their intervention with genes in the Wnt signaling pathway are shown in Figure 4(c). We may identify target genes for these molecules from differential genes in this pathway. The second example is a set of 15 small molecules targeting only the focal adhesion in liver cancer from GSE5364 (Figure S3(c)). Orciprenaline, budesonide, and hydrocortisone are used to treat asthma or asthma-related diseases. Trihexyphenidyl, tiletamine, and thiocolchicoside have sedative and muscle relaxant effects. The case is shaded light slate blue in Figure S1(a). Another instance is the separate subnetwork shaded by light slate blue in the M-P link map for lung cancer from GSE5364 (Figure S4(c)). The central node is the cell communication; this node is connected to 16 small molecules with degree one, of which 14 have relatively clear efficiency. Six molecules are mainly used as either antibacterial or anti-inflammatory agents, and four molecules are used for the treatment of psychoses. For the molecules in this link pattern, we can explore drug substitution or efficacy prediction for small molecules. Small molecules affecting the same gene ontology (GO) modules could be considered for drug substitution or efficacy prediction [35].

3.3. Potential Cocurative Effects of Molecules That Significantly Target Same Pathways. For some diseases, single-agent therapy may not be sufficiently effective [36]. To increase the effectiveness of drugs, combination therapy was used to enhance their efficacy by synergistic effect. Links between small molecules are implied in the M-P link map. We can also construct molecule-molecule links if two small molecules are both used to treat the same disease. We call these small molecule pairs cocurative molecule pairs. Small molecules were input into CTD to query diseases related to them. We calculate the overlap between cocurative molecule pairs and molecule pairs targeting the same pathways. The proportions of overlap in breast cancer, liver cancer, and lung cancer from GSE5364 are 41.3%, 41.2%, and 22.7%, respectively. These overlaps are statistically significant by Fisher's exact test with $P < 0.001$ [28]. The results showed that the small molecules that significantly targeted the same pathways tended to treat the same diseases. This conclusion has been mentioned in a recent published work [37]. Such small molecule pairs may be used as alternative medications or as a drug combination. As our knowledge of small molecules accumulates, the proportion of known pairs will increase. At the same time, the efficacy of molecules can be predicted more easily if the efficacies of corresponding molecules which target the same pathways are known. However, when administered together, antagonism between the cocurative drugs should be considered to reduce adverse reactions.

3.4. Common Small Molecules and Pathways Shared by Three Cancers. From the M-P link map, we found that there were 71 small molecules shared by three cancers from GSE5364. All molecules with degrees of five or more from the three cancers were included in this group. Of these 71 molecules, 58 were recorded in existing databases. Nine of these 58 molecules can be directly used as antitumour agents or directly impact tumour-related proteins. They are trichostatin A, geldanamycin, azacitidine, puromycin, streptozotocin, vorinostat, genistein, camptothecin, and sulfadimethoxine. The five common pathways are ECM-receptor interaction, focal adhesion, insulin signalling, cell communication, and Type II diabetes mellitus. The first three pathways are related to cell differentiation, proliferation, and apoptosis. Pathways appearing in only one cancer seldom show these characteristics. In liver cancer, for example, the majority of the pathways are involved in biological processes related to the metabolic functions of the liver. For instance, glycerophospholipid metabolism and glycerolipid metabolism are part of lipid metabolism, while fructose and mannose metabolism and pyruvate metabolism belong to carbohydrate metabolism, and glutathione metabolism metabolises abnormal amino acids. In particular, lipid metabolism and pyruvate metabolism are closely related to liver cancer [38–40].

3.5. Analysis and Validation of Link Relationship. To validate the link relationships between small molecules and pathways, we developed the reliability analysis. We used three sets of data from GEO, that is, GSE15852, GSE9166, and GSE7670, respectively, to construct M-P link. As a consequence, we obtained M-P links corresponding to six sets of data. Actually, there were two sets of data in each type of cancer. Firstly, for the links between 571 small molecules and 47 pathways from our method for breast cancer dataset, we extracted known targeted genes for each small molecule using DrugBank database and identified whether these targeted genes were members of 47 pathways. If so, the predicted target relationship can be validated. In the results, the relationships of 15 small molecules targeting 25 pathways have been identified. As a validated example, genistein was identified by CTD database, which was an antineoplastic and antitumor agent by targeting leukocyte transendothelial migration [41]. Furthermore, we added another breast cancer dataset to confirm our analysis and then identified the relations of 10 small molecules targeting 13 pathways. Cheng et al. found that quercetin induced tumor-selective apoptosis through downregulation of valine, leucine, and isoleucine degradation [42]. The overlap between two breast cancer datasets was 402 small molecules and showed high consistency. We also furthered the validation analysis by comparing the relationship of small molecules targeting pathways to DrugBank and CTD databases for liver and lung cancers and known targeted therapy results were available in Table S3 [43–45]. The validation results were summarised in Figure 5.

4. Discussion and Conclusions

In this study, we constructed a link map between small molecules and pathways with Connectivity Map database and analysed the topological properties of this link map.

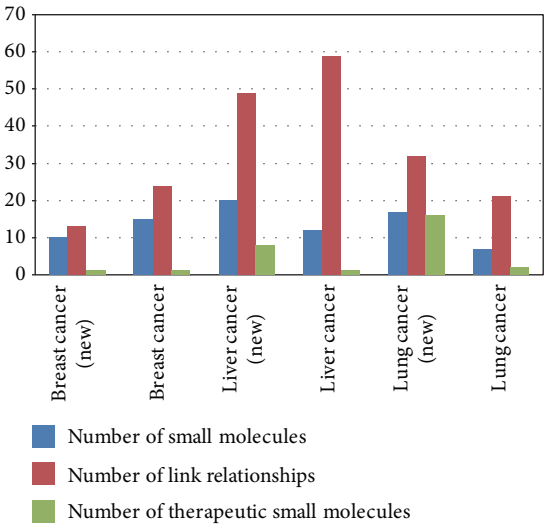


Figure 5: A chart indicates the validated six datasets of three cancers. The blue bars represent the validated small molecules, while the red bars turn out to be the validated M-P link relationships. Besides, the green bars are the small molecules for anticancer drugs proved.

Pathways were regarded as drug targets, and we studied small molecules' effects on disease from the perspective of these pathways, which provides valuable information for molecularly targeted therapy.

To detect pathways that were targeted by small molecules, disease-related differentially expressed genes are first enriched into pathways. In this way, differentially expressed genes were divided into various subsets based on KEGG pathways they participate in. Thus, it is possible to detect potential candidate small molecules from the perspective of the pathways that include differentially expressed genes. Although there is low reproducibility of differentially expressed genes between diverse data sources for the same kind of diseases, the expression correlation of those genes is high [46]. This result indicates that separate experiments can identify different but functionally correlated sets of differential genes. In the proposed method, we selected one dataset for each cancer on the same platform. Nevertheless, it is still possible to identify significant links between small molecules and pathways. By comparing the expression pattern similarity of the differentially expressed genes enriched in a pathway and the genes perturbed in Connectivity Map instances, molecules that intervene in the pathway were identified. Then, we constructed the link map between small molecules and all pathways. From the target pathway of the molecules, we can further screen their target genes.

The proposed approach established the mapping between disease-related pathways and small molecules. After analysing three cancer datasets from GSE5364, we were able to describe specific characteristics of the link maps of small molecules with high degrees in the sM-mP pattern that were always used for cancer treatment. Most tumours involve the regulation of multiple genes and processes, and there are usually many potential targets for treating tumours [29]. Thus, small molecules with high degrees may be considered as antineoplastic candidates. For small molecules in the mM-sP pattern, some of them are of similar efficacy. This similarity facilitates the prediction of drug efficacy and the identification of new efficacies for existing drugs. In the established link map, the number of genes shared by both pathways indeed plays part in impacting the number of their cotargeted small molecules. In spite of this, it is not the determining factor. By application of KS-like test of Connectivity Map to evaluate the expression deregulation caused by small molecules, the function of genes that remarkably interfered with small molecules became increasingly obvious. Even though the two pathways might share the identical genes, their function can vary from pathway to pathway, which would influence the number of corresponding small molecules that had been mined.

Although there are common molecules and pathways for the three cancers, there is no identifiable overlap among the M-P links in our datasets. This result is not surprising. In the process of pathway enrichment analysis for the three datasets, diverse differentially expressed genes may be enriched into different subpathways in the same pathway. When the genes are input into Connectivity Map to detect significant molecules, different genes in the same pathways may link to distinct molecules. Thus, in different datasets, the same pathway may be targeted by different molecules. Thus, there may not be any overlap among the M-P links for the three cancers.

Meanwhile, in order to validate the stability of the results, we added another set of data in each type of cancer to construct link maps. By comparing to the original data, we found a good repeatability. The number of repeat small molecules between two sets of breast cancer data is 402; when it comes to lung cancer and liver cancer, the numbers are 218 and 276, respectively. Axon guidance, which proved to be small molecules of tumor treatment in the previous analysis, was found in both sets of lung cancer data. In breast cancer data, we had found the meticrane targeting PPAR signaling pathway that had been reported before. Meanwhile, we found the known clomipramine targeting metabolism of xenobiotics by cytochrome P450 pathway in liver cancer.

In the future, we can expand the candidate target pathways affected by small molecules through the integration of multiple cancer datasets. As the information on pathways and small molecules accumulates, more comprehensive results will be obtained. In our future work, we will expand the types of cancer and provide creative thinking for molecularly targeted therapy from the perspective of systematic biology. We will focus our creative thinking on further researching mechanisms of drug functions of validated small molecules and make a contribution to the development of clinical anticancer drugs. Besides, we are taking other types of targets of small molecules, namely, miRNA, to specify the target relationships between small molecules and pathways. We will explore the molecule-impacted target pathway from different perspectives.

Conflict of Interests

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

Authors' Contribution

Yan Li, Weiguo Li, and Xin Chen contributed equally to this work.

Acknowledgments

This paper was supported in part by the National Natural Science Foundation of China (Grant no. 31100948) and the Natural Science Foundation of Jiangsu Province (Grant no. BK20131385).

References

[1] A. M. Gonzalez-Angulo, G. N. Hortobágyi, and F. J. Esteva, "Adjuvant therapy with trastuzumab for HER-2/neu-positive breast cancer," *Oncologist*, vol. 11, no. 8, pp. 857–867, 2006.

[2] J. Tabernero, "The role of VEGF and EGFR inhibition: implications for combining anti-VEGF and anti-EGFR Agents," *Molecular Cancer Research*, vol. 5, no. 3, pp. 203–220, 2007.

[3] T. J. Hobday and E. A. Perez, "Molecularly targeted therapies for breast cancer," *Cancer Control*, vol. 12, no. 2, pp. 73–81, 2005.

[4] F. G. Kuruvilla, A. F. Shamji, S. M. Sternson, P. J. Hergenrother, and S. L. Schreiber, "Dissecting glucose signalling with diversity-oriented synthesis and small-molecule microarrays," *Nature*, vol. 416, no. 6881, pp. 653–657, 2002.

[5] R. A. Pache, A. Zanzoni, J. Naval, J. M. Mas, and P. Aloy, "Towards a molecular characterisation of pathological pathways," *The FEBS Letters*, vol. 582, no. 8, pp. 1259–1265, 2008.

[6] C. F. Thorn, M. Whirl-Carrillo, T. E. Klein, and R. B. Altman, "Pathway-based approaches to pharmacogenomics," *Current Pharmacogenomics*, vol. 5, no. 1, pp. 79–86, 2007.

[7] D. C. Altieri, "Survivin, cancer networks and pathway-directed drug discovery," *Nature Reviews Cancer*, vol. 8, no. 1, pp. 61–70, 2008.

[8] M. K. Hellerstein, "A critique of the molecular target-based drug discovery paradigm based on principles of metabolic control: advantages of pathway-based discovery," *Metabolic Engineering*, vol. 10, no. 1, pp. 1–9, 2008.

[9] D. R. Rhodes, S. Kalyana-Sundaram, S. A. Tomlins et al., "Molecular concepts analysis links tumors, pathways, mechanisms, and drugs," *Neoplasia*, vol. 9, no. 5, pp. 443–454, 2007.

[10] S. Chian, R. Thapa, Z. Chi, X. J. Wang, and X. Tang, "Luteolin inhibits the Nrf2 signaling pathway and tumor growth in vivo," *Biochemical and Biophysical Research Communications*, vol. 447, no. 4, pp. 602–608, 2014.

[11] C. Gallo-Ebert, M. Donigan, I. L. Stroke et al., "Novel antifungal drug discovery based on targeting pathways regulating the fungus-conserved Upc2 transcription factor," *Antimicrobial Agents and Chemotherapy*, vol. 58, no. 1, pp. 258–266, 2014.

[12] S. M. An, Q. P. Ding, and L. S. Li, "Stem cell signaling as a target for novel drug discovery: recent progress in the WNT and Hedgehog pathways," *Acta Pharmacologica Sinica*, vol. 34, no. 6, pp. 777–783, 2013.

[13] S. Huber, S. Valente, P. Chaimbault, and H. Schohn, "Evaluation of △2-pioglitazone, an analogue of pioglitazone, on colon cancer cell survival: evidence of drug treatment association with autophagy and activation of the Nrf2/Keap1 pathway," *International Journal of Oncology*, vol. 45, no. 1, pp. 426–438, 2014.

[14] R. C. Arend, A. I. Londono-Joshi, R. S. Samant et al., "Inhibition of Wnt/beta-catenin pathway by niclosamide: a therapeutic target for ovarian cancer," *Gynecologic Oncology*, vol. 134, no. 1, pp. 112–120, 2014.

[15] C. Garcia-Echeverria and W. R. Sellers, "Drug discovery approaches targeting the PI3K/Akt pathway in cancer," *Oncogene*, vol. 27, no. 41, pp. 5511–5526, 2008.

[16] I. Collins and P. Workman, "New approaches to molecular cancer therapeutics," *Nature Chemical Biology*, vol. 2, no. 12, pp. 689–700, 2006.

[17] A. W. E. Chan and M. P. Weir, "Using chemistry to target treatments," *Chemical Innovation*, vol. 31, no. 12, pp. 12–17, 2001.

[18] J. Lamb, E. D. Crawford, D. Peck et al., "The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease," *Science*, vol. 313, no. 5795, pp. 1929–1935, 2006.

[19] J. Lamb, "The Connectivity Map: a new tool for biomedical research," *Nature Reviews Cancer*, vol. 7, no. 1, pp. 54–60, 2007.

[20] D. S. Wishart, C. Knox, A. C. Guo et al., "DrugBank: a comprehensive resource for in silico drug discovery and exploration," *Nucleic Acids Research*, vol. 34, pp. D668–D672, 2006.

[21] D. S. Wishart, C. Knox, A. C. Guo et al., "DrugBank: a knowledgebase for drugs, drug actions and drug targets," *Nucleic Acids Research*, vol. 36, pp. D901–D906, 2008.

[22] A. P. Davis, C. G. Murphy, C. A. Saraceni-Richards, M. C. Rosenstein, T. C. Wiegers, and C. J. Mattingly, "Comparative toxicogenomics database: a knowledgebase and discovery tool for chemical-gene-disease networks," *Nucleic Acids Research*, vol. 37, no. 1, pp. D786–D792, 2009.

[23] J. Wixon and D. Kell, "The Kyoto encyclopedia of genes and genomes—KEGG," *Yeast*, vol. 17, no. 1, pp. 48–55, 2000.

[24] E. W. Sayers, "Database resources of the National Center for Biotechnology Information," *Nucleic Acids Research*, vol. 37, supplement 1, pp. D5–D15, 2009.

[25] A. J. Butte, J. Ye, H. U. Häring, M. Stumvoll, M. F. White, and I. S. Kohane, "Determining significant fold differences in gene expression analysis," *Pacific Symposium on Biocomputing*, pp. 6–17, 2001.

[26] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 9, pp. 5116–5121, 2001.

[27] B. Zhang, S. Kirov, and J. Snoddy, "WebGestalt: an integrated system for exploring gene sets in various biological contexts," *Nucleic Acids Research*, vol. 33, no. 2, pp. W741–W748, 2005.

[28] I. Rivals, L. Personnaz, L. Taing, and M. Potier, "Enrichment or depletion of a GO category within a class of genes: which test?" *Bioinformatics*, vol. 23, no. 4, pp. 401–407, 2007.

[29] M. R. Green, "Targeting targeted therapy," *The New England Journal of Medicine*, vol. 350, no. 21, pp. 2191–2193, 2004.

[30] F. Chang, J. T. Lee, P. M. Navolanic et al., "Involvement of PI3K/Akt pathway in cell cycle progression, apoptosis, and neoplastic transformation: a target for cancer chemotherapy," *Leukemia*, vol. 17, no. 3, pp. 590–603, 2003.

[31] C. Blancher, J. W. Moore, N. Robertson, and A. L. Harris, "Effects of ras and von Hippel-Lindau (VHL) gene mutations on hypoxia-inducible factor (HIF)-1α, HIF-2α, and vascular endothelial growth factor expression and their regulation by the phosphatidylinositol 3′-kinase/Akt signaling pathway," *Cancer Research*, vol. 61, no. 19, pp. 7349–7355, 2001.

[32] T. Lahusen, A. de Siervi, C. Kunick, and A. M. Senderowicz, "Alsterpaullone, a novel cyclin-dependent kinase inhibitor, induces apoptosis by activation of caspase-9 due to perturbation in mitochondrial membrane potential," *Molecular Carcinogenesis*, vol. 36, no. 4, pp. 183–194, 2003.

[33] L. Meijer, M. Flajolet, and P. Greengard, "Pharmacological inhibitors of glycogen synthase kinase 3," *Trends in Pharmacological Sciences*, vol. 25, no. 9, pp. 471–480, 2004.

[34] E. ter Haar, J. T. Coll, D. A. Austen, H. M. Hsiao, L. Swenson, and J. Jain, "Structure of GSK3beta reveals a primed phosphorylation mechanism," *Nature Structural Biology*, vol. 8, no. 7, pp. 593–596, 2001.

[35] Y. Li, P. Hao, S. Zheng et al., "Gene expression module-based chemical function similarity search," *Nucleic Acids Research*, vol. 36, no. 20, article e137, 2008.

[36] C. Bokemeyer, C. Kollmannsberger, S. Stenning et al., "Metastatic seminoma treated with either single agent carboplatin or cisplatin-based combination chemotherapy: a pooled analysis of two randomised trials," *The British Journal of Cancer*, vol. 91, no. 4, pp. 683–687, 2004.

[37] Y. Liu, B. Hu, C. Fu, and X. Chen, "DCDB: drug combination database," *Bioinformatics*, vol. 26, no. 4, pp. 587–588, 2010.

[38] J. Jiang, P. Nilsson-Ehle, and N. Xu, "Influence of liver cancer on lipid and lipoprotein metabolism," *Lipids in Health and Disease*, vol. 5, article 4, 2006.

[39] J. T. Jiang, C. Wu, N. Xu, and X. Zhang, "Mechanisms and significance of lipoprotein(a) in hepatocellular carcinoma," *Hepatobiliary and Pancreatic Diseases International*, vol. 8, no. 1, pp. 25–28, 2009.

[40] X. Liang, A. R. Chavez, N. E. Schapiro et al., "Ethyl pyruvate administration inhibits hepatic tumor growth," *Journal of Leukocyte Biology*, vol. 86, no. 3, pp. 599–607, 2009.

[41] H. S. Seo, D. G. DeNardo, Y. Jacquot et al., "Stimulatory effect of genistein and apigenin on the growth of breast cancer cells correlates with their ability to activate ER alpha," *Breast Cancer Research and Treatment*, vol. 99, no. 2, pp. 121–134, 2006.

[42] S. Cheng, N. Gao, Z. Zhang et al., "Quercetin induces tumor-selective apoptosis through downregulation of Mcl-1 and activation of bax," *Clinical Cancer Research*, vol. 16, no. 23, pp. 5679–5691, 2010.

[43] S. T. Philips, Z. L. Hildenbrand, K. I. Oravecz-Wilson, S. B. Foley, V. E. Mgbemena, and T. S. Ross, "Toward a therapeutic reduction of imatinib refractory myeloproliferative neoplasm-initiating cells," *Oncogene*, 2013.

[44] M. A. Scheper, N. G. Nikitakis, R. Chaisuparat, S. Montaner, and J. J. Sauk, "Sulindac induces apoptosis and inhibits tumor growth in vivo in head and neck squamous cell carcinoma," *Neoplasia*, vol. 9, no. 3, pp. 192–199, 2007.

[45] M. R. Ramsey, L. He, N. Forster, B. Ory, and L. W. Ellisen, "Physical association of HDAC1 and HDAC2 with p63 mediates transcriptional repression and tumor maintenance in squamous cell carcinoma," *Cancer Research*, vol. 71, no. 13, pp. 4373–4379, 2011.

[46] M. Zhang, L. Zhang, J. Zou et al., "Evaluating reproducibility of differential expression discoveries in microarray studies by considering correlated molecular changes," *Bioinformatics*, vol. 25, no. 13, pp. 1662–1668, 2009.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 758672, 13 pages
http://dx.doi.org/10.1155/2014/758672

Research Article

Effect of Duplicate Genes on Mouse Genetic Robustness: An Update

Zhixi Su,[1] Junqiang Wang,[1] and Xun Gu[1,2]

[1] *State Key Laboratory of Genetic Engineering and MOE Key Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University, Shanghai 200433, China*
[2] *Department of Genetics, Development & Cell Biology, Iowa State University, Ames, IA 50010, USA*

Correspondence should be addressed to Zhixi Su; zxsu@fudan.edu.cn and Xun Gu; xungufudan@gmail.com

Received 15 May 2014; Revised 15 June 2014; Accepted 16 June 2014; Published 10 July 2014

Academic Editor: Leng Han

In contrast to *S. cerevisiae* and *C. elegans*, analyses based on the current knockout (KO) mouse phenotypes led to the conclusion that duplicate genes had almost no role in mouse genetic robustness. It has been suggested that the bias of mouse KO database toward ancient duplicates may possibly cause this knockout duplicate puzzle, that is, a very similar proportion of essential genes (P_E) between duplicate genes and singletons. In this paper, we conducted an extensive and careful analysis for the mouse KO phenotype data and corroborated a strong effect of duplicate genes on mouse genetics robustness. Moreover, the effect of duplicate genes on mouse genetic robustness is duplication-age dependent, which holds after ruling out the potential confounding effect from coding-sequence conservation, protein-protein connectivity, functional bias, or the bias of duplicates generated by whole genome duplication (WGD). Our findings suggest that two factors, the sampling bias toward ancient duplicates and very ancient duplicates with a proportion of essential genes higher than that of singletons, have caused the mouse knockout duplicate puzzle; meanwhile, the effect of genetic buffering may be correlated with sequence conservation as well as protein-protein interactivity.

1. Introduction

Functional compensation of duplicate (paralogous) genes has been thought to play an important role in genetic robustness [1–7]. Indeed, existence of a close paralog in the same genome could result in null mutations of the gene with little effect on the organismal fitness (nonessential gene), as observed in both yeast and nematode [1–4]. However, the role and magnitude of the duplicate genes contributing to genetic robustness in mammals remain controversial [8–13]. Two studies on mouse knockout phenotypes [9, 10] observed that the proportion of essential genes (P_E) is similar between duplicate genes and singletons in mouse, sharply contrasted to those well-known findings that removing a duplicate gene usually generates less deleterious phenotypes than removing a single-copy gene [1–4]. On the other hand, Hsiao and Vitkup [8] suggested an important role in robustness against deleterious mutations of duplicate genes in human [8]. We call this controversy the knockout duplicate puzzle in mammals. Since knockout mice have been widely used as animal models for human diseases, resolving this issue may have a significant impact on biomedical sciences.

In summary, there are three alternative hypotheses proposed.

(i) The Duplicability Hypothesis. By combining the protein-protein interaction data into the analysis, Liang and Li [9] found that mouse duplicate genes tend to have much higher protein connectivity than those for singletons. Since high connectivity means high functional centrality in the gene network, they proposed that mouse duplicates probably are more important than singletons and that this factor could compromise the contribution of duplicate compensation. In other words, functionally important genes may have more chance to be duplicated. It remains unexplained why more important mouse genes tend to be duplicated, while yeast genes may have the opposite trend [14].

(ii) The No-Role Hypothesis. In contrast, Liao and Zhang [10] argued that the compensational role of duplicates in mouse genetic robustness is negligible. After examining a number of

genomic factors, they discussed several possibilities that may result in similar proportion of essential genes between singletons and duplicates. It implies that most recently duplicated mouse genes, for example, 26 rodent-specific prolactin-like proteins [15], may have lost functional compensations to each other. This prediction seems to be counterintuitive and does not receive much experimental evidence for supporting.

(iii) Age-Distribution Hypothesis. Su and Gu [11] have noticed the effect of sampling bias: recently duplicated genes, for example, after the mammalian radiation, are severely underrepresented in the current mouse KO database. Because most of the mouse gene knockouts were generated by individual laboratories for finding knockout phenotypes, recently duplicated genes may have been purposely avoided to minimize the experimental cost due to negative-phenotype results. In other words, the age distribution of duplicates in the data sample is upwardly biased, resulting in underestimation of the overall duplicate effect on the genetic robustness.

(iv) The Functional Importance Hypothesis. Makino et al. (2009) reported that there is a strong sampling bias towards the duplicated genes generated by whole genome duplication (WGD) in current mouse KO phenotype dataset [12].

Since most of the mouse WGD duplicates are ancient duplicate genes, their conclusion that the mouse knockout duplicate puzzle may be caused by sampling bias of WGD duplicate genes is consistent with age-distribution hypothesis. Previous studies [16–18] have shown that mammalian duplicate genes can be characterized as two waves (Wave-I for young duplicates and Wave-II for those duplicated around the origin of vertebrates) and the ancient component (prior to the split of vertebrates and *Drosophila*). We [11] observed that the mouse (Wave-I) young duplicates were indeed severely underrepresented, and, for duplicates in the knockout experiments, their characteristic age (duplication time) could be as ancient as that of Wave-II (early vertebrates) or even more ancient. Obviously, very ancient duplicates certainly have little effect on the genetic robustness. However, due to the space limit, in the short communication we only had a brief discussion about the other two hypotheses. In this paper, we conducted an extensive and careful analysis for the updated mouse gene deletion phenotype data to evaluate the relative merit between the duplicability hypothesis, the no-role hypothesis, and the age-distribution of duplicates hypothesis.

In this paper, we use an updated mouse KO dataset to carry out an extensive analysis. To facilitate the study, we proposed an empirical evolutionary model of gene essentiality—the A&B model (Age of duplication and genetic Buffering)—to explain knockout duplicate puzzle. Our results suggest that duplication age and genetic buffering determine the essentiality of mouse duplicates.

2. Results

2.1. Similar P_E between Singletons and Duplicates Caused by Strong Bias in Mouse KO Genes toward Ancient Duplicates. Of the 4123 mouse genes with available phenotypic data, 1921 were identified as essential genes. Meanwhile, we identified 2479 duplicate genes and 464 singleton genes and calculated proportions of essential genes (P_E), respectively. Consistent with previous studies [9–12], the updated mouse KO dataset shows no statistical difference of P_E between singletons and duplicates (44.8% versus 46.3%; $P = 0.56$). That is, proportions of essential genes in mouse singletons and duplicates are similar, in contrast to the well-known observations in other model organisms [1, 3]. Based on a more broad definition of gene essentiality (Materials and Methods), that is, genes with premature death or induced morbidity phenotype were considered as essential genes, we found the same pattern (data not shown).

Though it is highly suspected that recently duplicated genes may have been underrepresented in the mouse KO database, detection of such bias at the genome level has been shown to be nontrivial [9–11], and Su and Gu [11] proposed a practically feasible solution: estimate the age of duplication event from the assumption of molecular clock. Since time estimation is well known to be error prone and based on a number of assumptions [19, 20], we have to develop a robust analytical pipeline to minimize the potential errors (see Materials and Methods). As shown in Figure 1, the histogram of mouse duplication events, short for the genome set, has recaptured the unique evolutionary feature of vertebrate gene families [16]. That is, it shows a pattern characterized by two waves (I, II) and an ancient (III) component [21, 22].

In the same manner, we estimated the duplication times between 2260 mouse knockout genes and their closest paralogs and found that the age distribution of duplicate pairs differs significantly between the genome set and the knockout set ($P < 10^{-16}$, χ^2-test). The histograms in Figure 1 clearly show that mouse KO experiments have been designed to avoid recently duplicated genes, for example, only 1.4% for those duplicated within 100 mya (around or after the mammalian radiation) in the KO set, compared to 19.6% in the mouse genome set. Consequently, the ages of duplicate genes in the mouse knockout dataset are typically around 500 to 700 mya (in early vertebrates), with a long-tail toward even more ancient ones (>1000 mya). In other words, the sampling bias toward ancient duplicates in the currently available mouse KO target genes has been nontrivial. These ancient duplicates may have undergone substantial functional divergence so that they have lost the capacity of functional compensation. In contrast, recent gene duplications, those duplicated around the mammalian radiation or in the rodent lineage, are expected to have significant contributions to the gene robustness in the current mouse genome. While these young duplicates were considerably underrepresented in the mouse knockout dataset, the observed proportion of essential duplicate genes is upwardly biased close to the value of singletons.

2.2. The Duplication-Age and Buffering Model (Age-Buffering Model) of Gene Essentiality. Since initially duplicated genes were completely compensated, the loss process of duplicate compensation is apparently time dependent, during which the outcome can be influenced by many gene-specific factors. To have a complete understanding of gene essentiality in

duplicates and singletons, an evolutionary model is needed. We formulate a simple *A&B* model as follows, short for *A*ge of duplication and genetic *B*uffering. Without genetic buffering, we assume that the probability of a duplicate remains nonessential, that is, functionally compensated by another duplicate copy in the same genome, and decayed exponentially with the time t (the age of gene duplication), that is, $e^{-\lambda t}$, where λ is the loss rate of duplicate compensation by mutations. Next, let g be the probability that a gene is genetically buffered. Together, the *A&B* model demonstrates that a gene to be essential depends on two mechanisms: the effect of genetic buffering (g) and the age-dependent effect of duplication compensation ($e^{-\lambda t}$). Obviously, the probability of a duplicate gene being essential is the probability for both mechanisms failure, that is,

$$P_E = (1 - g)\left(1 - e^{-\lambda t}\right). \quad (1)$$

Under this model, the negligible role hypothesis [10] actually claimed a very high loss rate (λ) of functional compensation such that $P_E \approx P_E^*$ in the current mouse genome. On the other hand, the duplicability model [9] assumes that the effect of genetic buffering (g) of duplicates is lower than that of singletons denoted by g^*, that is, $g < g^*$, such that $(1 - g)(1 - e^{-\lambda t}) \approx 1 - g^*$ holds. In fact, (1) suggests that three parameters, t (duplication age), g (genetic buffering), and λ (loss rate of functional compensation), together determine the gene essentiality of mouse duplicates. Particularly, we have two claims: (i) the proportion of essential genes in mouse duplicates (P_E) is age dependent on gene duplications; (ii) gene essentiality correlates to sequence conservation or protein connectivity in either duplicates or singletons largely because these two factors affect the efficiency of genetic buffering (g), rather than the functional compensation between duplicates. Our preliminary analysis [11] has shown the first claim. In the following we provide a detailed analysis to address some technical issues and doubts.

Our models suggested that, for sufficient time, P_E approaches to a level that is roughly equal to P_E of singleton. However, it does not mean that all these ancient duplicates are subject to the genetic buffering. A likely situation is that genetic buffering and duplication coevolve. In other words, the reason why some duplicates can remain dispensable for a long time is because they were integrated into existing or novel genetic buffering mechanisms.

Chen et al. (2010) found that in *Drosophila* new genes could become essential rapidly after the gene duplications [23]. This mechanism is also likely to exist in mammals. To take this factor into account, we modify (1) as follows:

$$P_E = (1 - g)\left[1 - (1 - \rho)\, e^{-2\lambda t}\right], \quad (2)$$

where the parameter $\rho > 0$ indicates the process of rapid essentiality in the early stage after gene duplication. Because the number of mouse KO genes is small for very young duplicates, a further investigation requires when the data are available.

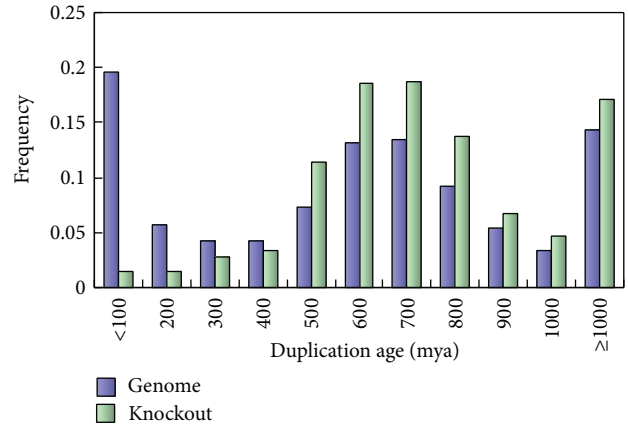


Figure 1: Duplication age distribution of mouse genome set (blue bars) and knockout gene set (green). The x-axis indicates the duplication age (t) between a duplicated gene and its closest paralog. The y-axis indicates the frequency of the duplicates in each duplication age category.

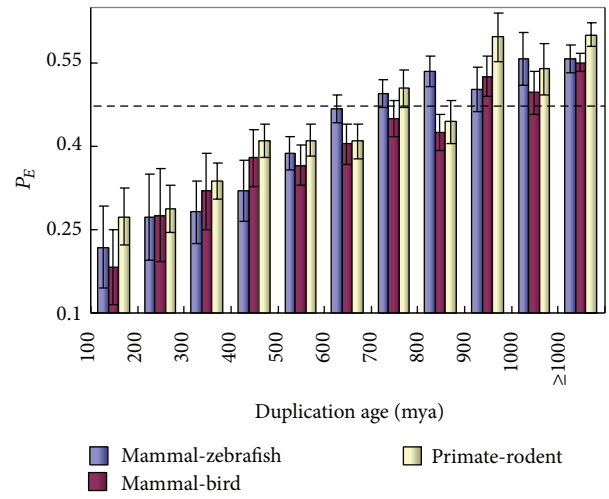


Figure 2: Relationship between P_E in duplicate genes and the duplication age. Error bars show one standard error. The dashed line indicates the P_E level of single-copy genes.

2.3. Proportions of Essential Genes (P_E) in Mouse Duplicates Are Age Dependent. A simple solution to correct this knockout sampling bias is to calculate P_E under a given age bin. We implemented several approaches to minimize the noise effect in time estimation. First, we used three time calibration points to date mouse duplication events: the mammal-zebrafish split (430 mya), the mammal-bird split (310 mya), and the primate-rodent split (80 mya), respectively, and calculated P_E for every age bin of 100 million years. As shown in Figure 2, in all cases we observed that P_E increases from a low value in young duplicates with the increasing of duplication ages; this P_E-age (t) correlation is statistically significant ($P < 10^{-4}$, χ^2-test). To be concise, in the following of this paper, we mainly present the results based on the mammal-zebrafish split time calibration. Noticeably, we found that P_E

in ancient duplicates, say, >700 mya, is unexpectedly higher than that of singletons; $P_E = 0.542 \pm 0.016$, $P < 0.001$. Hence, there are two reasons for why the overall P_E in duplicates has no difference from that of singletons: the sampling bias toward ancient duplicates and very ancient duplicates with a higher P_E than that of singletons. In addition, we conducted simulations to examine the effect of violation of molecular clock (constant evolutionary rate) on the estimation of P_E. Our results showed that the age dependency of P_E can be weakened or even vanished by the violation of molecular clock. In other words, our conclusion of P_E-age correlation seems to be conserved (not shown). Finally, we inferred the phylogenetic locations of mouse KO duplication events in three intervals: after the mammal-zebrafish split, after the mammal-bird split, and after the primate-rodent split. In each interval we calculated P_E, which is compatible to the proportion of essential genes, with respect to the three major speciation events in vertebrates: P_E is ~23% for those duplicated after the mammalian radiation, ~31% for those duplicated after the bird-mammal split, and close to ~39% for those duplicated after the teleost-tetrapod split. Although a decreasing P_E in younger duplicates is biologically intuitive, it is subject to the statistical uncertainty due to small sample size. Nevertheless, under a more broad age category, such as before the split of land animals and fishes versus the more ancient duplicates, the difference is statistically significant ($P < 0.01$).

In a separate study, we developed a simple bias-correcting procedure to obtain a bias-corrected P_E and test whether it is significantly lower than in singletons. We predicted that $P_E = 41.7\%$ for all duplicate genes, which are impressive compared to $P_E = 46.3\%$ observed in sample duplicates and $P_E = 47\%$ in sample singletons [11]. However, in this study, when we used a more stringent criterion to define single-copy genes, we found that there is no statistical significant difference between the predicted P_E and P_E of single-copy genes (41.7% versus 44.8%, $P = 0.21$). We want to emphasize that, even after taking this bias into consideration, the difference between P_E for singletons and P_E for duplicates at the genome level is still small. This may be because the contribution of functional compensation by young duplicates cancels the contribution of higher intrinsic importance of ancient duplicate, which is consistent with the duplicability hypothesis [9].

2.4. Age Dependence of P_E in Mouse Duplicates and Sequence Conservation. Though a simple interpretation for the P_E-t correlation is that the capability of duplicate compensation decays with the evolutionary time since the duplication [11], some other alternatives cannot be ruled out, which were based on the correlation of gene essentiality with, for instance, sequence conservation or protein connectivity [9, 10, 24]. We have addressed these issues carefully.

To measure the sequence conservation, we used the conventional ratio of the number of nonsynonymous substitutions per site (d_N) to the number of synonymous substitutions per site (d_S), which was estimated from the mouse gene and its human ortholog (see Materials and Methods). A low d_N/d_S ratio indicates high sequence conservation of the gene. Consistent with previous studies [10, 25], we showed that essential mouse genes tend to be more conserved: P_E decreases with the increase of d_N/d_S for both duplicates (Spearman rank $\rho = -0.23$, $P < 10^{-15}$) and singletons ($\rho = -0.18$, $P < 10^{-15}$; see Figure 3(a) for binned results). After calculating the mean d_N/d_S ratio for each age bin of mouse duplicates, we unexpectedly found that sequence conservation is actually positively correlated with the duplication age (t) (Figure 3(b), $P < 10^{-10}$). This unexpected inverse age-d_N/d_S relationship raises the possibility that the observed P_E-t (age) correlation could be confounded by theP_E-d_N/d_S correlation conjugated with the age-d_N/d_S correlation.

We first claim that the P_E-d_N/d_S correlation is the consequence of the inverse relationship between the genetic buffering (g) and the sequence conservation (d_N/d_S). Hence, the inverse age-d_N/d_S relationship in mouse duplicates suggests less effect of genetic buffering in ancient duplicates than that in recent duplicates, implying that the genetic buffering of duplicates g could be age dependent. One possible evolutionary mechanism for the age-g inverse relationship could be the neofunctionalization in the late stage after the gene duplication so the preexisting (ancestral) genetic buffering systems did not work for the newly acquired functions.

Suppose that the effects of genetic buffering (g) are similar between singletons and duplicates, as long as they have a similar d_N/d_S ratio; we designed a simple procedure as follows to take the effect of sequence conservation into account. That is, for each age bin (t) of duplicates, the buffering effect ($1 - g$) was estimated from the P_E of the singleton mouse KO genes, corrected by the linear regression with the d_N/d_S ratio, and denoted by $P_E^*(t)$ (Figure 3(a)). To be clear, we used $P_{E\text{-dup}}(t)$ for the age-bin (t) of mouse duplicates. Figure 3(c) plotted both $P_{E\text{-dup}}(t)$ and $P_E^*(t)$ against age bins of duplicates. As expected, $P_E^*(t)$ increases with the duplication age t, but much slower than $P_{E\text{-dup}}(t)$, indicating that the P_E-d_N/d_S correlation can only explain a small portion of the P_E-age correlation in duplicates. According to (1), the relative essentiality in duplicates, $P_{E\text{-dup}}(t)/P_E^*(t)$, is given by

$$\frac{P_{E\text{-dup}}(t)}{P_E^*(t)} = 1 - e^{-\lambda t}, \tag{3}$$

which measures the pure duplication effect on gene essentiality and does not depend on the sequence conservation. Indeed, we found a significantly positive correlation between the ratio $P_{E\text{-dup}}(t)/P_E^*(t)$ and the duplication age ($P < 0.001$; Figure 3(d)). We repeated our analysis using d_N/d_S ratio of mouse-rat orthologous gene pairs and obtained a virtually same result (Figure S1; see Figure S1 in Supplementary Material available online at http://dx.doi.org/10.1155/2014/758672). We therefore conclude that the proportion of essential genes (P_E) of mouse duplicates is age dependent, even after correcting the potential confounding effect from the essentiality-conservation dependence.

2.5. Age Dependence of P_E in Mouse Duplicates and Protein Connectivity. The proportion of essential genes is positively correlated with protein connectivity in mouse [9]. In our

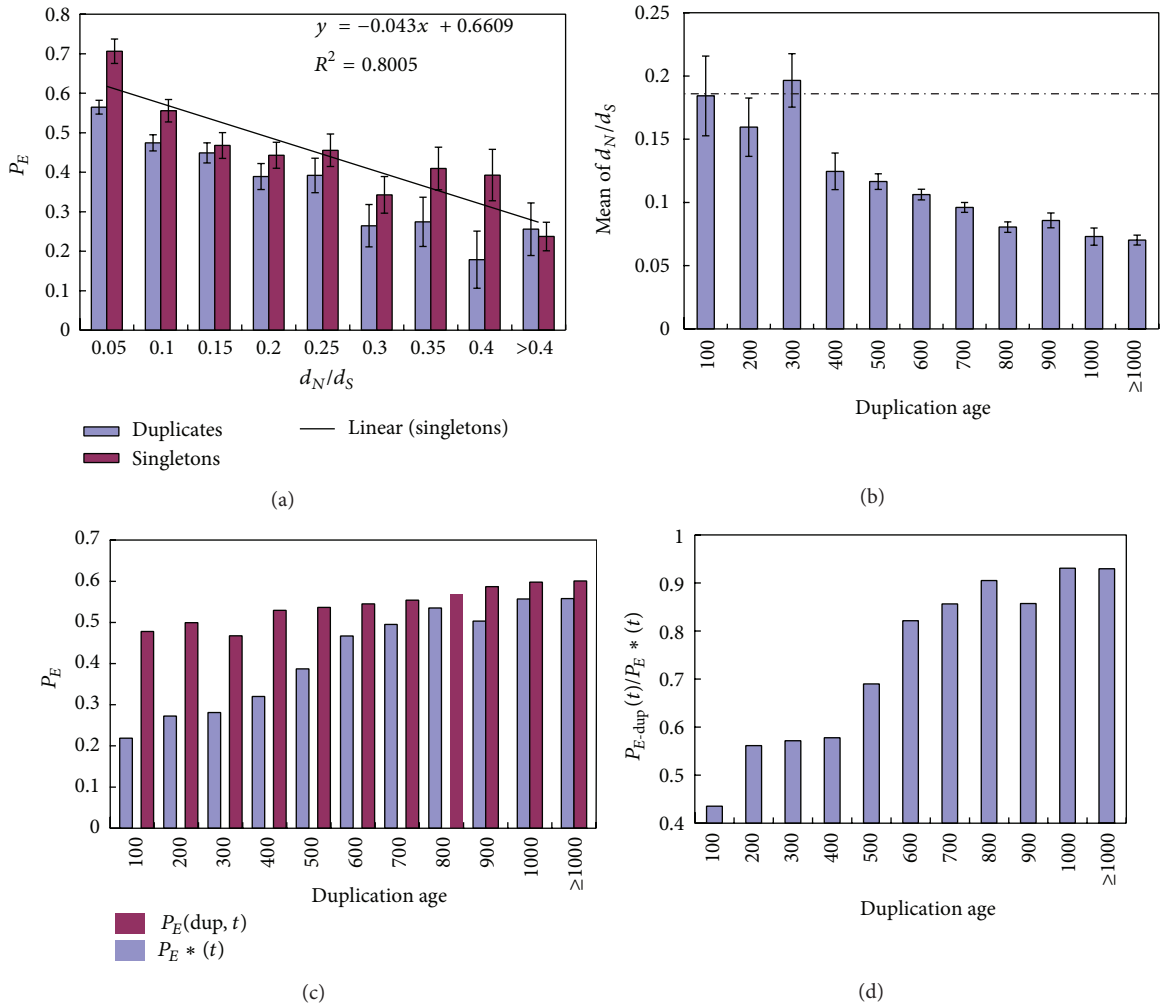


FIGURE 3: The effect of sequence conservation on the relationship between P_E and duplication age. (a) Relationship between P_E in duplicate genes (blue) or singletons (purple) and the evolutionary conservation of the gene, measured by the ratio of the nonsynonymous (d_N) to synonymous (d_S) nucleotide distances between the target gene and its human ortholog. Linear regression line and regression equation between d_N/d_S ratio and P_E in knockout single-copy genes are presented on the panel. (b) Mean d_N/d_S ratio for each age bin of duplicates. Dashed line denotes the mean d_N/d_S ratio of singleton mouse knockout genes. (c) P_E in each age bin of duplicates—P_E (dup, t)—and that of singletons with the same d_N/d_S ratio—$P_E^*(t)$. $P_E^*(t)$ is calculated based on the mean d_N/d_S ratio for duplicates in each age bin (panel b) and the linear regression equation (panel a). (d) Ratio of P_E (dup, t) and $P_E^*(t)$ in each age bin of duplicates. Error bars show one standard error.

updated mouse KO dataset, we compiled 211 singleton mouse KO targeted genes with available protein connectivity data, as well as 845 mouse KO duplicates [26]. Consistent with [9], we confirmed a weak but significant positive correlation between protein connectivity and P_E in both duplicates (Spearman rank $\rho = 0.11$, $P = 0.001$) and singletons ($\rho = 0.11$, $P = 0.003$; see Figure 4(a) for binned results). Similar to the effect of sequence conservation, the A&B model interprets this finding as genes with high connectivity may have low genetic buffering. Due to the small sample size, we further group the 845 genes into seven age groups. We then calculated the mean of protein interaction number for duplicated genes in each age bin and found no correlation of the mean protein connectivity with the duplication age (t) (Spearman rank $\rho = 0.04$, $P = 0.19$, Figure 4(b)).

We thus hypothesize that P_E-connectivity and P_E-age correlations reflect two independent underlying mechanisms. To further test this hypothesis, we divided duplicate genes with interaction data into two groups, those with high connectivity (larger than the median interaction, i.e., >2 interactions) and those with low connectivity (otherwise). The proportion of essential genes in the high-connectivity group is apparently higher than that in the low-connectivity group ($P < 0.001$). But, as shown in Figure 4(c), the inverse relationship between P_E and the age of duplicates holds in both gene groups. We thus conclude that age dependence of the proportion of essential genes (P_E) in duplicates is unlikely to be confounded by the effect of protein connectivity.

2.6. Age Dependence of P_E Is Irrespective of Sampling Bias toward Essential Genes, Developmental Genes, or WGD Duplicates. It was proposed that individual researchers might tend to report a gene with a discernible phenotype in the KO

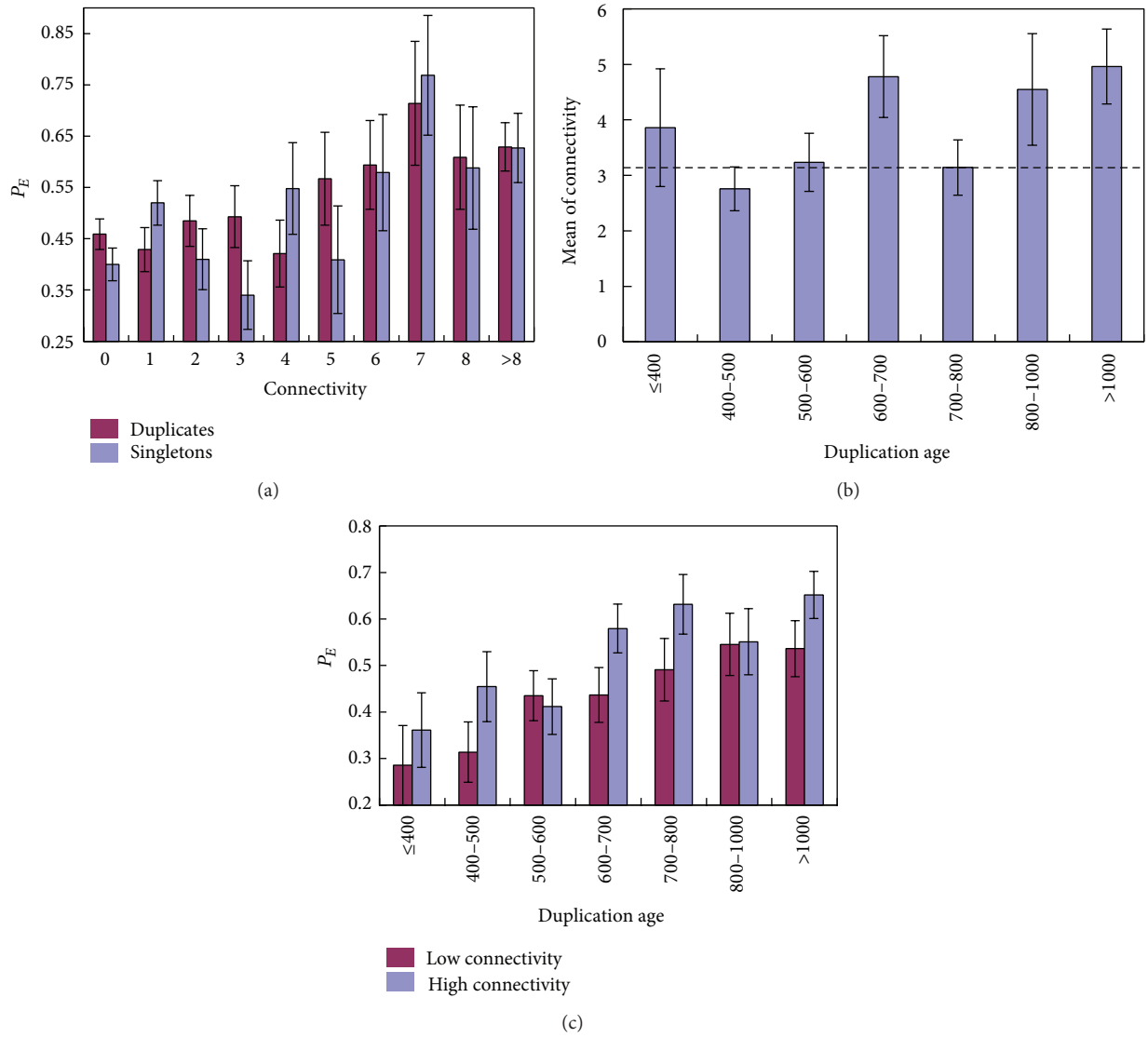


FIGURE 4: The effect of protein connectivity on the relationship between P_E and duplication age. (a) Relationship between P_E in duplicate genes (blue) or singletons (purple) and the protein connectivity of the gene. (b) Mean interaction number for each age bin of duplicates. Dashed line denotes the mean interaction number of singleton mouse knockout genes. (c) Relationship between P_E in duplicate genes and the duplication age for high connectivity genes and low connectivity genes. Error bars show one standard error.

experiments [10, 12]. Therefore, reports of gene knockouts with stronger phenotype (essential genes) are likely to be dramatically overrepresented in the KO dataset. A previous study found that the developmental genes and duplicated genes generated by WGD tend to be more essential than the nondevelopmental genes and small-scale duplication (SSD) duplicated genes, respectively. Besides, the current mouse KO dataset is biased toward developmental genes and WGD duplicates. Therefore, it is suspected that the ancient duplicates bias of KO duplicates and P_E-t correlation might be only a byproduct of the above factors. Here, we tested whether the bias of ancient duplicates of KO dataset is a side effect of the biased sampling of WGD genes or developmental genes and whether age dependency of P_E still holds after controlling the influences of the above factors.

If the sampling bias towards the ancient duplicates is just caused by the preferential report of the essential genes by individual mouse KO experiments, no age distribution difference would be expected between KO nonessential duplicates and the whole genome set. We then compared the age distribution of nonessential KO duplicates with the whole genome set. As shown in Figure 5, even after removing all essential genes, the KO duplicates still show strong age bias toward ancient duplicated genes. Therefore, we conclude that the age bias of KO genes is not an artifact of sampling bias of essential genes.

To test the influence of the sampling bias of developmental genes, we subdivided all the mouse genes with at least one GO item as developmental genes and nondevelopmental genes, based on the approaches of [12]. In the KO dataset,

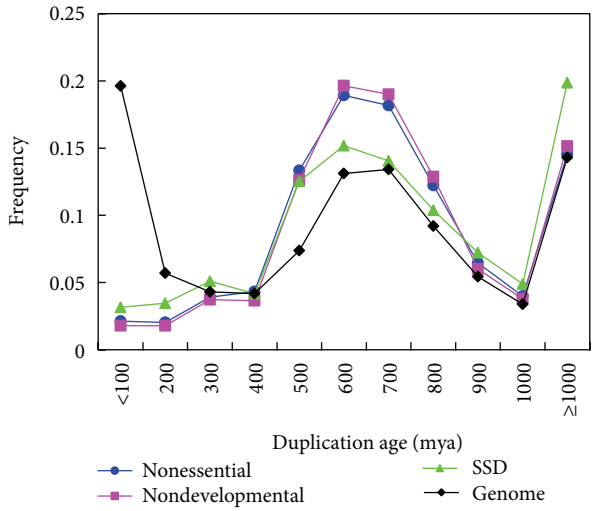


FIGURE 5: Duplication age distribution of mouse genome set (black), nonessential duplicates, nondevelopmental duplicates, and SSD duplicates.

we found that the P_E of developmental genes is significantly higher than the P_E of nondevelopmental genes (66.1% versus 34.8%, $P < 2.2e-16$, χ^2 test). For all of the duplicate genes with at least one GO item in KO dataset, we found 36.8% of them belonging to developmental genes, which is significantly larger than the proportion of developmental genes in whole genome set (13.4%, $P < 2.2e-16$). The similar bias also has been found in single-copy genes (28.9% versus 8.3%, $P < 2.2e-16$). These findings indicate that developmental genes were enriched in the mouse KO dataset irrespective of single-copy genes or duplicated genes, which is consistent with previous study [12]. If the sampling bias of KO duplicates toward the ancient duplicated genes is only caused by the bias of developmental genes, it is expected that the age distribution of KO nondevelopmental duplicates will be similar to that of whole genome set. However, for the nondevelopmental duplicates, we found that the age distribution of duplicates differs significantly between the genome set and KO set. That is, recently duplicated nondevelopmental genes have been underrepresented in the mouse nondevelopmental KO dataset (Figure 5). Since developmental genes are more essential than other genes, it is reasonable to suspect that the positive P_E-t correlation might be simply because of the trend that ancient duplicates have more developmental genes. To address this issue, we calculated the P_E-t correlation for developmental and nondevelopmental genes, respectively. We found that the P_E-t correlation is statistically significant, in both developmental genes ($\rho = 0.1$, $P = 0.002$, Spearman rank test) and nondevelopmental genes ($\rho = 0.2$, $P < 1e-5$).

The sampling bias of WGD duplicates also may confound our analysis. More and more evidences indicated that there may have been two rounds of WGD that occurred during the early stage of vertebrate evolution (500–700 mya), and duplicate developmental genes created by WGD were preferentially retained in vertebrate genome [12, 27]. We tested if we rule out the influence of WGD duplicates the A&B model still holds. Following the methods of [12], we obtained a list of human duplicated genes created by WGD inferred by [28]. We then inferred the mouse duplicated genes generated by whole genome duplication through one-to-one orthology relationships with the human genes. We identified 1237 mouse WGD duplicated genes and 1242 SSD duplicated genes with phenotype data. We found that the P_E of WGD duplicates is 51.1%, which is larger than the P_E of singletons (44.7%, $P = 0.02$). We then estimated the duplication age between all SSD duplicated KO genes and their closest paralogs and found that the age distribution of SSD duplicates still differs significantly between the genome set and SSD KO set ($P < 1e-16$, χ^2 test). Figure 5 clearly shows that, even after ruling out the WGD genes, the KO duplicates dataset is still biased toward ancient duplicates. We further calculated the P_E for each bin of age (100 mya) and observed that P_E-t correlation holds for SSD KO genes ($\rho = 0.21$, $P < 1e-11$).

2.7. What Determines Duplicate Compensation: Evolutionary Time (Age) or Sequence Conservation? The protein sequence divergence between duplicate genes, or the evolutionary distance (d), was widely used as a proxy measure of the age of duplicates. In our study we used the Poisson-corrected method to estimate the protein sequence distance (d). Figure 6(a) shows no correlation between P_E and d, as claimed in [10]. A straightforward explanation is that the sequence distance between duplicates (d) is determined by $d = 2vt$, where v is the evolutionary rate of the protein sequence and t is the age of duplicates. As shown in Figure 3(b), an ancient duplicate gene (a large t) tends to be conserved (low v as measured by low d_N/d_S ratio) so that the P_E-d independence could be the result of canceled P_E-t andP_E-d_N/d_S correlations.

Our conclusion that the P_E-d relationship is not fundamental differs from Liao and Zhang [10]. Assuming that it is the protein sequence similarity, not the age of gene duplication, which determines the likelihood of compensation between duplicates, the authors of [10] argued that the lack of correlation between P_E and d may indicate the negligible role of duplicate genes in the mouse genetic robustness. Here, we conduct a simple case-study to show that it may not be the case. We divided 135 mouse KO duplicate pairs with $d < 0.2$ (corresponding to 82% sequence identity between KO duplicates and their paralogs) into the "young" group (age <310 mya, after the bird-mammal split) or the "old" group (≥310 mya). Strikingly, we found $P_E = 0.39$ for the young group and $P_E = 0.58$ for the old group ($\chi^2 = 4.56$, $P = 0.03$) (Figure 6(b)). Moreover, we calculated the mean sequence conservation (the d_N/d_S ratio) in both groups: $d_N/d_S = 0.12$ for young duplicates and 0.02 for ancient duplicates. Does this mean that different P_E in young and old groups is caused by the difference in sequence conservation? From theP_E-d_N/d_S regression in singletons (Figure 3(a)), we predict that, if there is no functional compensation between duplicates, the young group should have the $P_E = 0.56$ versus the old group $P_E = 0.64$ (Figure 6(b)), which is contradictory to our observation. We therefore conclude that, for these duplicate pairs with >82% protein sequence identity, recent duplicate pairs are functionally more compensated than ancient pairs.

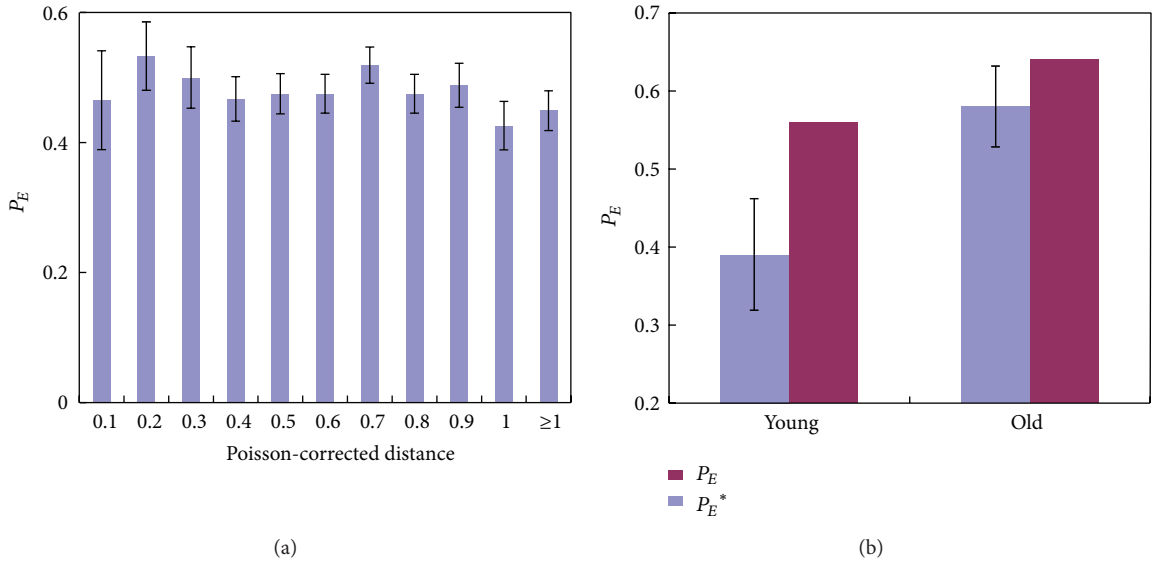


FIGURE 6: Relationship between P_E and protein sequence divergence. (a) P_E in duplicate genes is not correlated with the Poisson-corrected distance (d) between the target gene and its closest paralog in the genome. Error bars show one standard error. (b) P_E and P_E^* of mouse knockout duplicate pairs with sequence divergence $d < 0.2$. "Young" group represents the knockout genes with duplication age <310 mya, and "old" group represents the knockout genes with duplication age ≥310 mya. P_E^* is calculated based on the mean d_N/d_S ratio for each group and the linear regression equation of Figure 3(a).

The A&B model we proposed suggests that the age of gene duplication plays an important role in functional compensation between duplicates, while the sequence conservation indicates the likelihood of a duplicate gene actually genetically buffered by other (nonhomologous) genes, as supported by recent double deletions of yeast duplicate pairs [29, 30]. Noticing that, in many cases, the sequence similarity and functional similarity between paralogs may not be strongly correlated [31], we tentatively propose the transient hypothesis for the observed P_E-age correlation. That is, because only a few nucleotide substitutions are responsible for the compensation loss between duplicates, the time interval for maintaining the effective compensation between duplicates mainly depends on the "waiting time" for these substitutions to occur.

3. Discussion

In this study, we formulated an evolutionary model (A&B model) to address the knockout duplicate puzzle in mouse. That is, a duplicate gene to be essential depends on two mechanisms: the effect of genetic buffering (g) and the age-dependent effect of duplication compensation. We convincingly showed that the role of duplicates in mouse genetic robustness is nontrivial, similar to other simple model organisms [1–4]. There are substantial segmental or tandem gene duplications in the mouse genome around the mammalian radiation or even during the rodent lineage. These recently duplicated genes are expected to play major roles in the mouse gene robustness [11]. In spite of the fact that they were considerably underrepresented in the current mouse KO database, after the careful analysis that ruled out the potential confounding effect from sequence conservation, protein connectivity, functional bias, or bias of WGD duplicates, we reached the conclusion that differs sharply from the previous statement [10] of negligent duplicate effect on mouse genetic robustness. It is interesting to find that P_E seems to increase with organismal complexity. That is, though a greater fraction of genes in complex organisms may have been essential to ensure viability and fertility than that in simple organisms, for example, under laboratory conditions, P_E is ~7% in *Escherichia coli* [32], 17% in yeast [8, 33], and >46% in mouse, the age-dependent effect of duplicates on gene robustness remains similar from simple to complicated organisms. Of course, a more complete mouse KO database is crucial for further investigation.

Although there is no big difference between mouse and yeast in the role of duplicate genes in genetic robustness, mouse genetic robustness indeed reveals some unique features deserving further investigations: (i) why the P_E of mouse WGD duplicates is larger than the P_E of average single-copy mouse genes, but, in yeast, it is much smaller than its counterpart; (ii) why the P_E of yeast singletons is much larger than the P_E of duplicates, but the difference is not very evident in mouse even after controlling the sampling bias; (iii) why protein connectivity is high in mouse duplicated genes, in contrast to the case in the yeast [9, 14]. Though one may speculate that each problem may have several possible explanations, we propose a unified evolutionary model that can interpret these observations, which is the quite different age distribution of duplicated genes between mouse and budding yeast resulting from different evolutionary origins.

In the yeast *Saccharomyces cerevisiae*, the most recent WGD event occurred relatively recently (in the last ~100

million years) [34]. The majority of the yeast duplicated genes are quite young. For example, we found that only 13.1% of the yeast duplicates were generated 500 mya. In contrast, 58.9% of the mouse duplicates were created 500 mya (unpublished data). As shown in Figure 1, a significant portion of duplicate genes in vertebrates, including fishes, birds, and mammals, were generated by large-scale genome-wide duplications in the early stage of vertebrates [26, 35–39]. Though there still remains some controversies on how many rounds of WGDs had occurred during the evolution of early vertebrates, a general agreement has been reached that these duplication events may result in concomitant increase of developmental genes involving signal-transduction and transcription regulation that may be relevant to the expansion of cell types in the origin of vertebrates. For instance, we found a significant increase of paralogous genes in GPCRs (G-protein coupled receptors) and GPCR-pathway related protein families during the early stage of vertebrates. Transcription factors and protein kinases also show the same pattern [40]. These signaling-related molecules apparently tend to have more numbers of protein-protein interactions; many of them actually act as hubs in the process of signaling. If the evolutionary process of transition from invertebrate to vertebrate required the increase of tissue-specific signaling pathways, signaling-related duplicate genes may be favorably preserved in the genome. This hypothesis explains why protein connectivity in mammals is high in duplicate genes.

Another intriguing observation is the specific features of ancient duplicates. We found that ancient duplicates tend to be more conserved, and the ancient duplicate gene tends to be more essential than an average single-copy gene. First thought for why ancient duplicates are more conserved is puzzling, because it is generally believed that duplicated genes may have experienced a relaxed evolution due to the functional redundancy. Hence, an interpretation based on positive selection could be that the follow-up neofunctionalization may impose stronger functional constraints on these ancient duplication genes. Though it stands as an interesting hypothesis, we offer a much simpler explanation. For those ancient duplicate genes originated over 500 mya, only highly conserved duplicate pairs can be detected by the standard homologous search. In other words, sequence similarity between ancient duplicate genes with relatively low sequence conservation may be too low to be detected. Our simple calculation has shown that it may occur very likely. Suppose that the evolutionary rate of a gene is typically 3×10^{-9} per change/year. Since the ancient duplication event (500 mya), the sequence identity between duplicate copies, under the simplest Poisson model, is estimated to be $\exp[-2 \times 3 \times 10^{-9} \times 500 \times 10^{6}] = e^{-3} \approx 0.0498$! Note that the cutoff for sequence similarity in homologous search is usually around 0.25. An interesting explanation for why ancient mouse duplicates even have a higher degree of gene essentiality than the average of singletons invokes acquisition of new functions that facilitates the loss process of functional compensation between duplicates. However, our analysis (Figure 3(c)) shows that a nonadaptive alternative may be more likely; that is, ancestral genes for those duplicated in early or prior to vertebrates may have stronger sequence conservation. In this case, using the overall proportion of essential genes in singletons as a reference may be misleading.

Since functional compensation of duplicated genes has been found to play an important role in genetic robustness in various species, from simple eukaryote yeast to complicated mammal mouse, it is highly expected that the similar scenario holds in human. However, owing to the impossibility of getting the large-scale human gene KO phenotypic data, it is not possible to systematically verify this expectation. Recently, several studies showed evidences that disrupt duplicate genes have less phenotype effect in human genome, indicating a possible contribution of duplicate genes to the human genetic robustness. For example, two separate studies found that the human specific nonprocessed pseudogenes or long-established lost genes are overrepresented in genes belonging to large gene families, such as olfactory receptor or zinc finger protein family [41, 42]. These results might indicate that loss of duplicate genes could be compensated by their close paralogous genes. Similarly, through a large-scale experimental survey of nonsense SNPs in the human genome, Yngvadottir et al. (2009) discovered 99 genes with homozygous nonsense SNPs in healthy human population. These genes could be considered as nonessential genes [43]. They found that 51% of nonessential genes have at least one paralog, whereas in comparison only 35% of all human genes are reported to have a paralog ($P < 0.05$). So, it is possible that their function is "backed up" by duplicated paralogs in the human genome. Moreover, Hsiao and Vitkup (2008) found that genes with close homologs are significantly less likely to harbor known disease mutations compared to genes with remote homologs [8]. In addition, close duplicates affect the phenotypic consequences of deleterious mutations by making a decrease in life expectancy less likely. If all the gene samples of above studies represent the entire genome, the results would mean that the effect of duplicate genes on genetic robustness holds in human genome.

In our study, the duplication age was estimated between the mouse KO gene and its closest paralog. Many mouse KO genes have more than one paralog, consisting of a large gene family. In such cases the pattern of functional compensation is complex, which cannot be revealed because most members have no KO phenotype information. Our approach is based on the premise that the closest paralog is the major determinant of functional compensation. Of course our treatment could be biased, and the future study should be gene-family based. The bottleneck still is the lack of sufficient KO genes. We indeed conducted a preliminary survey of the distribution of KO genes in a family but the dataset is too small to be useful at the current stage. Another technical issue is about the age of singleton. While we use the common procedure to determine singletons, the age of gene does affect P_E in both duplicate and singleton genes. One may see Chen et al. (2012) for details [44].

The mouse KO database provides a valuable resource to study the genomic features of vertebrate evolution from gene essentiality [9, 10, 45] to pleiotropy [46]. Since mouse tissue-specific developmental genes were largely duplicated

in the early stage of vertebrates (~500 mya), while mammalian character-related genes were duplicated recently, the contribution of duplicates to genetic robustness may be more associated to mammalian-specific phenotypes. On the other hand, duplication events in the early stage of vertebrates were tightly associated with the expansion of signaling pathways for the evolution of vertebrate-specific multicellularity [16]. This may explain why gene duplicability and protein interactions are positively correlated [9], as signaling-related proteins tend to have high number of protein interactions. The effect of gene duplications on genetic robustness depends on the distribution of young duplicate genes in the current genome. Therefore, its impact varies among species, mainly because each species has its unique age distribution of gene duplications. For instance, due to recent polyploidizations, duplicate genes may dominate the genetic robustness in plant genomes [47]. It will be interesting to see whether the conclusions made in mouse hold in general when more invertebrate null mutation phenotypic data become available for such analyses.

4. Materials and Methods

4.1. Genomic Data. Protein sequences of mouse (NCBIM36), human (NCBI36), chicken (WASHUC2), and zebrafish (Zfish6) genes were extracted from Ensembl (release 59). If a gene had more than one alternative-splicing form, the longest isoform was used. Since several processed pseudogenes inserted into the genome very recently could be erroneously annotated as functional genes in Ensembl [48], we identified the single-exon genes with protein sequence identity ≥98% to multiple-exon genes as processed pseudogenes. The identified processed pseudogenes were excluded in the following analysis. The transcript and exon data of mouse genes were also obtained from Ensembl. For each alternatively spliced gene, the exon number was defined as the largest exon number of its all transcript isoforms.

Mouse phenotype and genotype association file (MGI_PhenoGenoMP.rpt) was downloaded from Mouse Genome Informatics (ftp://ftp.informatics.jax.org) (release 08/23/2010) [49]. This file contains specific mammalian phenotype (MP) ontology terms annotated to genotypes. Mammalian phenotype browser (http://www.informatics.jax.org/searches/MP_form.shtml) was used to match MP terms and phenotype details. Here, an essential gene was defined as a gene whose KO phenotype is annotated as lethality (including embryonic, prenatal, and postnatal lethality) or infertility [9]. We excluded all the phenotypic annotations due to multiple gene KO experiments and only used those of null mutation homozygotes by target deletion or gene-trap technologies. Totally, 4123 genes with phenotypic information were extracted from this file. We then classified these genes into 1921 essential genes and 2202 nonessential genes. Some different criteria were used to examine the effect of the definition of "essential genes" that we used above. For example, we followed the methods of [10, 45] to define essential genes. We found that though P_E varies under different criteria for essential genes, it does not change our major results qualitatively (data not shown).

Homology information of mouse-human genes (mouse-rat) was obtained from Ensembl BioMart (release 59). The number of synonymous substitutions per synonymous site (d_S) and the number of nonsynonymous substitutions per nonsynonymous site (d_N) between mouse and human orthologs were estimated by the maximum likelihood method using PAML [50] and were retrieved from Ensembl EnsMart. For mouse genes have many human (rat) orthologs, the pair with the smallest d_N/d_S ratio was used for further analysis.

We calculated the protein connectivity (k) based on the protein-protein interaction data of one-to-one human orthologous genes (including both yeast two-hybrid and literature-curated interactions, but excluding self-binding interactions) [26]. Because of the absence of the large-scale mouse protein-protein interaction experiment and the function similarity between human-mouse orthologous genes, here we use the protein connectivity of corresponding human orthologs to approximately represent that of the mouse KO genes.

4.2. Identification of Duplicate Genes and Singletons. We used a method similar to that of Gu et al. [51] to identify duplicate genes and single-copy genes. Because we want to detect the differences in P_E between real duplicates and singletons, we use stringent criteria to define duplicate genes and singletons. Briefly, every protein was used as the query to search against all other proteins by using Blastp ($E = 1e - 10$) [52]. Two proteins are scored as forming a link if (1) the alignable region between them is >80% of the longer protein and (2) the identity (I) between them is $I \geq 30\%$ if the alignable region is longer than 150aa and $I \geq 0.01n + 4.8L^{-0.32[1+\exp(-L/1000)]}$ for all other protein pairs, in which $n = 6$ and L is the alignable length between the two proteins. We deleted proteins if they formed a hit due to the presence of a repetitive element of the same family. The Blastp non-self best hit of a duplicate gene was defined as its closest paralog. A singleton gene is defined as a protein that does not hit any other proteins in the Blastp search with $E = 1e - 10$; this loose similarity search criterion was used to make sure that a singleton is indeed a singleton. Our results were essentially unchanged when we chose an even looser criterion, such as $E = 1e - 5$.

4.3. Dating Duplication Time of Mouse Duplicate Genes. We developed an analytical pipeline to estimate the duplication times (ages) of mouse duplicate genes on a large scale, using the split-time between the mouse and zebrafish (430 million years ago, mya) as a calibration. First, we shall define *Inparalogs clusters of mouse and zebrafish*; that is, those paralogs duplicated after the mouse-zebrafish split, in either mouse or zebrafish lineage. One may see Figure 7 for illustration. Apparently, there are two modes for each duplicate pair: duplicated after the mouse-zebrafish split (Figure 7(a)) or before mouse-zebrafish split (Figure 7(b)).

We used the Inparanoid program (Version 2.0) to infer Inparalogs clusters of mouse and zebrafish [53]. Mouse and zebrafish genes in the same cluster are then identified as orthologs. A multiple alignment including the mouse duplicate genes, their closest paralogs, and their Inparalogs clusters

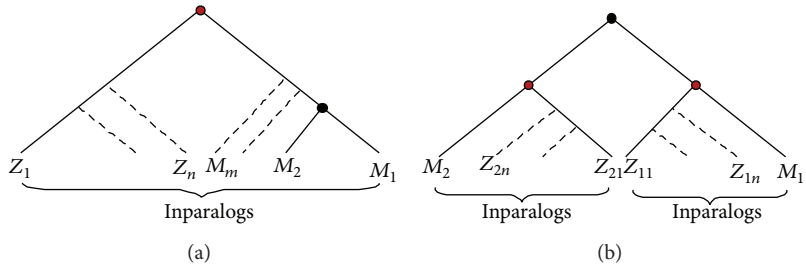


Figure 7: Illustration of the evolutionary relationship between mouse and zebrafish genes. Mouse duplicate events may occur after mouse-zebrafish split (a) or before it (b). Red node represents the speciation event and black node represents the duplication events. Genes under a red node represent a mouse-zebrafish inparalog cluster.

(orthologs) was obtained by Tcoffee [54]. For those clusters containing more than 10 mouse or zebrafish Inparalogs, to reduce the complexity of calculation, besides mouse duplicate pair, 10 mouse or zebrafish Inparalogs were randomly selected for further alignment. Poisson-corrected distances between duplicates (d_m) or orthologs were calculated after all alignment gaps were eliminated.

In each case (a) or (b) (Figure 7), we calculated the distance between the mouse knockout duplicate and its closest paralog and the averaged distance between mouse and zebrafish orthologs, which can be easily converted to the geological time (million years ago) under the assumption of molecular clock [16]. By this method, the duplicate time between each of 9503 mouse genes and its closest paralog was estimated (whole genome set). Among them, 2260 genes were KO target genes (knockout set).

Abbreviations

P_E: Proportion of essential genes
Mya: Million years ago
GPCRs: G-protein coupled receptors
WGD: Whole genome duplication
SSD: Small-scale duplication.

Conflict of Interests

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

Acknowledgments

This work was in part supported by the National Science Foundation of China (31272299) and the China State Key Basic Research Program (2012CB910101) and grants from Fudan University. Zhixi Su was supported by the Shanghai Pujiang Program (13PJD005). The authors are grateful to Wen-Hsiung Li, Han Liang, and Jianzhi Zhang for critical comments on early version of the paper, and Gangbiao Liu for his assistance.

References

[1] G. C. Conant and A. Wagner, "Duplicate genes and robustness to transient gene knock-downs in Caenorhabditis elegans," *Proceedings of the Royal Society B: Biological Sciences*, vol. 271, no. 1534, pp. 89–96, 2004.

[2] E. J. Dean, J. C. Davis, R. W. Davis, and D. A. Petrov, "Pervasive and persistent redundancy among duplicated genes in yeast," *PLoS Genetics*, vol. 4, no. 7, Article ID e1000113, 2008.

[3] Z. Gu, L. M. Steinmetz, X. Gu, C. Scharfe, R. W. Davis, and W. Li, "Role of duplicate genes in genetic robustness against null mutations," *Nature*, vol. 421, no. 6918, pp. 63–66, 2003.

[4] Y. Guan, M. J. Dunham, and O. G. Troyanskaya, "Functional analysis of gene duplications in *Saccharomyces cerevisiae*," *Genetics*, vol. 175, no. 2, pp. 933–943, 2007.

[5] X. Gu, "Evolution of duplicate genes versus genetic robustness against null mutations," *Trends in Genetics*, vol. 19, no. 7, pp. 354–356, 2003.

[6] R. S. Kamath, A. G. Fraser, Y. Dong et al., "Systematic functional analysis of the Caenorhabditis elegans genome using RNAi," *Nature*, vol. 421, no. 6920, pp. 231–237, 2003.

[7] E. A. Winzeler, D. D. Shoemaker, A. Astromoff et al., "Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis," *Science*, vol. 285, no. 5429, pp. 901–906, 1999.

[8] T.-L. Hsiao and D. Vitkup, "Role of duplicate genes in robustness against deleterious human mutations," *PLoS Genetics*, vol. 4, no. 3, Article ID e1000014, 2008.

[9] H. Liang and W. Li, "Gene essentiality, gene duplicability and protein connectivity in human and mouse," *Trends in Genetics*, vol. 23, no. 8, pp. 375–378, 2007.

[10] B.-Y. Liao and J. Zhang, "Mouse duplicate genes are as essential as singletons," *Trends in Genetics*, vol. 23, no. 8, pp. 378–381, 2007.

[11] Z. Su and X. Gu, "Predicting the proportion of essential genes in mouse duplicates based on biased mouse knockout genes," *Journal of Molecular Evolution*, vol. 67, no. 6, pp. 705–709, 2008.

[12] T. Makino, K. Hokamp, and A. McLysaght, "The complex relationship of gene duplication and essentiality," *Trends in Genetics*, vol. 25, no. 4, pp. 152–155, 2009.

[13] K. Hannay, E. M. Marcotte, and C. Vogel, "Buffering by gene duplicates: an analysis of molecular correlates and evolutionary conservation," *BMC Genomics*, vol. 9, article 609, 2008.

[14] A. Prachumwat and W. Li, "Protein function, connectivity, and duplicability in yeast," *Molecular Biology and Evolution*, vol. 23, no. 1, pp. 30–39, 2006.

[15] D. O. Wiemers, L. J. Shao, R. Ain, G. Dai, and M. J. Soares, "The mouse prolactin gene family locus," *Endocrinology*, vol. 144, no. 1, pp. 313–325, 2003.

[16] X. Gu, Y. Wang, and J. Gu, "Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution," *Nature Genetics*, vol. 31, no. 2, pp. 205–209, 2002.

[17] G. Panopoulou, S. Hennig, D. Groth et al., "New evidence for genome-wide duplications at the origin of vertebrates using an amphioxus gene set and completed animal genomes," *Genome Research*, vol. 13, no. 6, pp. 1056–1066, 2003.

[18] K. Vandepoele, W. de Vos, J. S. Taylor, A. Meyer, and Y. van de Peer, "Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 6, pp. 1638–1643, 2004.

[19] W. H. Li, *Molecular Evolution*, Sinauer Associates, Sunderland, Mass, USA, 1997.

[20] M. Nei and S. Kumar, *Molecular Evolution and Phylogenetics*, Oxford University Press, New York, NY, USA, 2000.

[21] S. Ohno, *Evolution by Gene Duplication*, Springer, Berlin, Germany, 1970.

[22] A. McLysaght, K. Hokamp, and K. H. Wolfe, "Extensive genomic duplication during early chordate evolution," *Nature Genetics*, vol. 31, no. 2, pp. 200–204, 2002.

[23] S. Chen, Y. E. Zhang, and M. Long, "New genes in Drosophila quickly become essential," *Science*, vol. 330, no. 6011, pp. 1682–1685, 2010.

[24] 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.

[25] A. E. Hirsh and H. B. Fraser, "Protein dispensability and rate of evolution," *Nature*, vol. 411, no. 6841, pp. 1040–1049, 2001.

[26] J.-F. Rual, K. Venkatesan, T. Hao et al., "Towards a proteome-scale map of the human protein-protein interaction network," *Nature*, vol. 437, no. 7062, pp. 1173–1178, 2005.

[27] T. Blomme, K. Vandepoele, S. De Bodt, C. Simillion, S. Maere, and Y. Van de Peer, "The gain and loss of genes during 600 million years of vertebrate evolution," *Genome Biology*, vol. 7, no. 5, article R43, 2006.

[28] Y. Nakatani, H. Takeda, Y. Kohara, and S. Morishita, "Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates," *Genome Research*, vol. 17, no. 9, pp. 1254–1265, 2007.

[29] J. Ihmels, S. R. Collins, M. Schuldiner, N. J. Krogan, and J. S. Weissman, "Backup without redundancy: genetic interactions reveal the cost of duplicate gene loss," *Molecular Systems Biology*, vol. 3, 2007.

[30] R. Harrison, B. Papp, C. Pál, S. G. Oliver, and D. Delneri, "Plasticity of genetic interactions in metabolic networks of yeast," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 7, pp. 2307–2312, 2007.

[31] H. H. Gan, R. A. Perlow, S. Roy et al., "Analysis of protein sequence/structure similarity relationships," *Biophysical Journal*, vol. 83, no. 5, pp. 2781–2791, 2002.

[32] T. Baba, T. Ara, M. Hasegawa et al., "Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection," *Molecular Systems Biology*, vol. 2, article 0008, 2006.

[33] X. He and J. Zhang, "Why do hubs tend to be essential in protein networks?" *PLoS Genetics*, vol. 2, no. 6, p. e88, 2006.

[34] M. Kellis, B. W. Birren, and E. S. Lander, "Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae," *Nature*, vol. 428, no. 6983, pp. 617–624, 2004.

[35] A. M. Heimberg, L. F. Sempere, V. N. Moy, P. C. J. Donoghue, and K. J. Peterson, "MicroRNAs and the advent of vertebrate morphological complexity," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 8, pp. 2946–2950, 2008.

[36] O. Lespinet, Y. I. Wolf, E. V. Koonin, and L. Aravind, "The role of lineage-specific gene family expansion in the evolution of eukaryotes," *Genome Research*, vol. 12, no. 7, pp. 1048–1059, 2002.

[37] N. Lopez-Bigas, S. de, and S. A. Teichmann, "Functional protein divergence in the evolution of *Homo sapiens*," *Genome Biology*, vol. 9, no. 2, article R33, 2008.

[38] A. Prachumwat and W. Li, "Gene number expansion and contraction in vertebrate genomes with respect to invertebrate genomes," *Genome Research*, vol. 18, no. 2, pp. 221–232, 2008.

[39] C. Vogel and C. Chothia, "Protein family expansions and biological complexity," *PLoS Computational Biology*, vol. 2, no. 5, p. e48, 2006.

[40] G. Liu, Y. Zou, Q. Cheng, Y. Zeng, X. Gu, and Z. Su, "Age distribution patterns of human gene families: divergent for Gene Ontology categories and concordant between different subcellular localizations," *Molecular Genetics and Genomics*, vol. 289, no. 2, pp. 137–147, 2014.

[41] X. Wang, W. E. Grus, and J. Zhang, "Gene losses during human origins," *PLoS Biology*, vol. 4, no. 3, article e52, 2006.

[42] J. Zhu, J. Z. Sanborn, M. Diekhans, C. B. Lowe, T. H. Pringle, and D. Haussler, "Comparative genomics search for losses of long-established genes on the human lineage," *PLoS Computational Biology*, vol. 3, no. 12, e247, 2007.

[43] B. Yngvadottir, Y. Xue, S. Searle et al., "A genome-wide survey of the prevalence and evolutionary forces acting on human nonsense SNPs," *American Journal of Human Genetics*, vol. 84, no. 2, pp. 224–234, 2009.

[44] W. H. Chen, K. Trachana, M. J. Lercher, and P. Bork, "Younger genes are less likely to be essential than older genes, and duplicates are less likely to be essential than singletons of the same age," *Molecular Biology and Evolution*, vol. 29, no. 7, pp. 1703–1706, 2012.

[45] B. Liao and J. Zhang, "Null mutations in human and mouse orthologs frequently result in different phenotypes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 19, pp. 6987–6992, 2008.

[46] X. Gu, "Evolutionary framework for protein sequence evolution and gene pleiotropy," *Genetics*, vol. 175, no. 4, pp. 1813–1822, 2007.

[47] J. F. Wendel, "Genome evolution in polyploids," *Plant Molecular Biology*, vol. 42, no. 1, pp. 225–249, 2000.

[48] Z. Zhang, N. Carriero, and M. Gerstein, "Comparative analysis of processed pseudogenes in the mouse and human genomes," *Trends in Genetics*, vol. 20, no. 2, pp. 62–67, 2004.

[49] J. T. Eppig, C. J. Bult, J. A. Kadin et al., "The Mouse Genome Database (MGD): from genes to mice—a community resource for mouse biology," *Nucleic Acids Research*, vol. 33, supplement 1, pp. D471–D475, 2005.

[50] Z. Yang, "PAML: a program package for phylogenetic analysis by maximum likelihood," *Computer Applications in the Biosciences*, vol. 13, no. 5, pp. 555–556, 1997.

[51] Z. Gu, A. Cavalcanti, F. C. Chen, P. Bouman, and W. H. Li, "Extent of gene duplication in the genomes of Drosophila, nematode, and yeast," *Molecular Biology and Evolution*, vol. 19, no. 3, pp. 256–262, 2002.

[52] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.

[53] M. Remm, C. E. V. Storm, and E. L. L. Sonnhammer, "Automatic clustering of orthologs and in-paralogs from pairwise species comparisons," *Journal of Molecular Biology*, vol. 314, no. 5, pp. 1041–1052, 2001.

[54] C. Notredame, D. G. Higgins, and J. Heringa, "T-coffee: a novel method for fast and accurate multiple sequence alignment," *Journal of Molecular Biology*, vol. 302, no. 1, pp. 205–217, 2000.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 272950, 8 pages
http://dx.doi.org/10.1155/2014/272950

Research Article

Designing Peptide-Based HIV Vaccine for Chinese

Jiayi Shu,[1,2] Xiaojuan Fan,[3] Jie Ping,[3] Xia Jin,[1,2] and Pei Hao[3]

[1] *Viral Disease and Vaccine Translational Research Unit, Institut Pasteur of Shanghai, Chinese Academy of Sciences, Room 507, Building B, Life Science Research Building, 320 Yueyang Road, Shanghai 200031, China*
[2] *Vaccine Centre, Institut Pasteur of Shanghai, Chinese Academy of Sciences, 320 Yueyang Road, Shanghai 200031, China*
[3] *Bioinformatics Platform, Institut Pasteur of Shanghai, Chinese Academy of Sciences, Room 405, Building B, Life Science Research Building, 320 Yueyang Road, Shanghai 200031, China*

Correspondence should be addressed to Xia Jin; xjin@ips.ac.cn and Pei Hao; phao@ips.ac.cn

Received 15 April 2014; Accepted 16 June 2014; Published 6 July 2014

Academic Editor: Siyuan Zheng

CD4+ T cells are central to the induction and maintenance of CD8+ T cell and antibody-producing B cell responses, and the latter are essential for the protection against disease in subjects with HIV infection. How to elicit HIV-specific CD4+ T cell responses in a given population using vaccines is one of the major areas of current HIV vaccine research. To design vaccine that targets specifically Chinese, we assembled a database that is comprised of sequences from 821 Chinese HIV isolates and 46 human leukocyte antigen (HLA) DR alleles identified in Chinese population. We then predicted 20 potential HIV epitopes using bioinformatics approaches. The combination of these 20 epitopes has a theoretical coverage of 98.1% of the population for both the prevalent HIV genotypes and also Chinese HLA-DR types. We suggest that testing this vaccine experimentally will facilitate the development of a CD4+ T cell vaccine especially catered for Chinese.

1. Introduction

Over 30 million people have died from HIV/AIDS related illnesses since HIV was discovered in the 1980s. There are currently 33 million of HIV carriers [1]. The rate of new infection is still on the rise globally. In China, HIV infection is a great concern, especially in southern part of China, for example, Yunnan, Sichuan, Guangxi, and Xinjiang Provinces, where a large number of infected people are drug users. Additionally, in the regions of Henan, Hubei Provinces where people were infected through illicit blood collection, the rate of infection reached up to 60% of blood donors [2]. Highly active antiretroviral therapy (HAART), a combination of three or more antiretroviral drugs, is routinely used to treat individuals with HIV infection [3]. It significantly extends the lifespan and improves the quality of life of people infected with HIV but cannot eradicate the virus [4]. The course of treatment is life-long and the medicines are expensive. In developing countries, available antiretroviral drugs are still limited. Therefore, a preventive HIV vaccine is especially needed.

HIV genome is comprised of nine structural (*Env*, *Gag*, and *Pol*) and regulatory (*Tat*, *Rev*, *Nef*, *Vif*, *Vpr*, and *Vpu*) genes. The *pol* gene encodes for reverse transcriptase which is error prone. This leads to high mutation rate, 15–20% divergence between the nucleic acid sequences of different HIV clades, and 7–12% variability within each clade [5]. Although the base composition of HIV genome is stable [6], host immune response further increases the HIV nucleotide diversity.

Due to the extreme sequence diversity and high mutation rate of HIV, it has been difficult to develop an efficacious HIV vaccine. A successful HIV vaccine requires inducing neutralizing antibodies and cytotoxic T cell responses, both of which can only be optimally induced and maintained in the presence of a concurrent CD4+ T helper cell response [7]. Despite many years of basic and clinical research, to date, there are only three major human HIV vaccine clinical trials completed. Set up in 1998, AIDSVAX gp120 protein vaccine is the first HIV vaccine going through Phase III trial in human and targeted to induce neutralizing antibody activity. Although antibodies to homologous virus were elicited, they

failed to neutralize heterologous viruses [8]. In 2004, a Phase IIb trial with Merck's MRKAd5, which is a trivalent vaccine including *gag*, *pol*, and *nef* genes in an adenovirus 5 vector, is designed for inducing cytotoxic T cell responses [9, 10]. Despite the induction of significant level of IFN gamma-producing T cells, the MRKAd5 has increased the risk of HIV acquisition in vaccine recipients and failed to reduce viral load after HIV infection [11]. Later in 2009, a Phase III trial of RV144 HIV-1 vaccine was completed in Thailand, which is a vaccine combination comprised of ALVAC (a vaccine containing genetically engineered versions of *gag*, *env*, and *pol* inserted in canarypox vector) and AIDSVAX (a bivalent gp120 envelope protein vaccine). These vaccines are theoretically capable of eliciting both CD8+ T cell response and neutralizing antibody response. Despite neither vaccine worked alone, in the combination, they unexpectedly lowered the HIV incidence by 31.2% in vaccine recipients; however, they did not reduce viral load [12]. These large clinical trials have opened new questions and revealed new opportunities for HIV vaccine research, including a rethinking of the need for a vaccine for CD4+ T helper cells.

In order to stimulate a CD4+ T helper cell response, antigens need to be processed and presented through MHC class II molecules. The form of antigen could be either whole protein or peptide epitopes. A previous study with a subunit vaccine comprised of 18 CD4+ T helper cell epitopes has demonstrated an efficient induction of robust helper T cell response in a Phase I clinical trial in Caucasian population [13]. Whether a similar strategy works in Chinese population requires to be tested.

To select antigenic epitopes for a vaccine, one must address several issues. One, HIV exhibits high mutation rates, and thus conserved sequences may be needed to cover a given population. Two, the human leukocyte antigen (HLA) is highly polymorphic, and it restricts the proportion of individuals who will respond to a particular antigen [14, 15]. To overcome these problems, promising T cell epitopes that bind to several HLA alleles for maximal population coverage should be selected [16], and a large variety of HIV sequences should be considered in the design of a HIV vaccine.

MHC class II is a heterodimer that is comprised of a monomorphic α and a highly polymorphic β chain. There are over 400 class II alleles identified, spreading among HLA-DM, HLA-DO, HLA-DP, HLA-DQ, and HLA-DR loci. Among them DRB1 is the most polymorphic gene, consists of 221 alleles; followed by DPB1 and DQB1 that has 84 and 39 alleles, respectively. Whereas other gene loci may have only 1 or 2 alleles [17]. Therefore, DRB1 is the best choice to optimize MHC II coverage. The frequency of HLA-DRB1 serotype differs among ethnic groups. Within DRB1 allotype, DRB1∗11 and 13 serotypes present in 16% and 14% of black population, whereas, in Caucasoid and Chinese, DRB1∗07 and DRB1∗11 and DRB1∗12 and DRB1∗15 appear in the highest percentage [17]. The above evidences support the development of a new HIV vaccine specifically for Chinese population. Such a vaccine should have higher probability in dealing with circulating HIV serotypes in China.

To overcome these complex issues of vaccine design, bioinformatics methods may help to determine common features of vaccine antigens that have potential to deal with divergent population and HIV quasispecies. Specifically, bioinformatics-based approach is the most feasible method in screening a large set of peptide epitopes and selection of promising vaccine antigens. In this study, we extracted 821 HIV sequence and 46 Chinese DRB1 alleles from public information and compiled a database. A combination of 7 public available epitope prediction algorithms was used to screen the database and identify CD4+ T cell epitopes as HIV vaccine antigens. We selected a set of 20 epitopes, which in combination could cover more than 98% of our target population.

2. Materials and Methods

2.1. Data Collection and Methods for Epitope Prediction. In total, 821 HIV whole genome sequences of Chinese population were retrieved from HIV Database (http://www.hiv.lanl.gov/) [27], and the distribution of 46 HLA-DR alleles (Table 1) was extracted from The Allele Frequency Net Database (AFND) (http://www.allelefrequencies.net/) [28].

Seven existing methods available in Immune Epitope Database (IEDB) [29] for MHC class II binding were used to predict HIV epitopes based on binding affinity between HLA DR types and HIV epitopes. These methods included Consensus method [30], NN-align (netMHCII-2.2) [31], stabilization matrix alignment method (SMM-align) [32], Sturniolo [33], average relative binding (ARB) [34], NetMHCIIpan [35], and Combinatorial library (ComLib) [30].

2.2. Epitope Selection. All epitopes are 15 amino acids in length. To be a potential epitope, it must have a MHC binding affinity threshold of IC_{50} = 500 nM or below. A selected epitope was removed from the epitope pool before the next prediction. The process is repeated until all epitopes were selected. All calculations of epitope selection process were conducted in INFORSENSE Knowledge Discovery Environment (KDS) software platform [36]. The mathematical model used to calculate the predictive score for each DR allele of known coverage (as listed in Table 1) is the following equations:

$$S(\alpha) = \sum_{i=1}^{821} \sum_{j=1}^{46} \delta_{i,j}(\alpha) \times C(j), \quad \text{(I)}$$

$$\delta_{i,j}(\alpha) = \begin{cases} 1, & \text{if } \alpha \text{ in the combination of HIV} \\ & \text{sequence } i \text{ and DR allele } j \\ 0, & \text{Otherwise.} \end{cases} \quad \text{(II)}$$

In the first equation (I), α represents the epitope; $C(j)$ is the percentage coverage of number j DRB1 allele; $\delta(\alpha)$ is the function to indicate whether the epitope exists in the combination of HIV sequence and DR allele, existence scored 1, and absence scored 0. $S(\alpha)$ is the sum of number of times of the binding of HIV sequence i and DR allele j after being standardized to the proportion of DR allele j in all DRB1

Table 1: The DRB1 allele coverage in Chinese population.

Number	Alleles	Coverage
1	DRB1*01:01	0.0145
2	DRB1*01:02	0.0014
3	DRB1*03:01	0.0514
4	DRB1*03:07	0.0009
5	DRB1*04:01	0.0120
6	DRB1*04:02	0.0024
7	DRB1*04:03	0.0238
8	DRB1*04:04	0.0082
9	DRB1*04:05	0.0413
10	DRB1*04:06	0.0233
11	DRB1*04:07	0.0041
12	DRB1*04:08	0.0075
13	DRB1*04:10	0.0030
14	DRB1*04:17	0.0018
15	DRB1*07:01	0.0677
16	DRB1*08:01	0.0018
17	DRB1*08:02	0.0076
18	DRB1*08:03	0.0512
19	DRB1*08:04	0.0029
20	DRB1*08:09	0.001
21	DRB1*08:12	0.0011
22	DRB1*09:01	0.0490
23	DRB1*10:01	0.0149
24	DRB1*11:01	0.0669
25	DRB1*11:03	0.0015
26	DRB1*11:04	0.0154
27	DRB1*11:06	0.0013
28	DRB1*12:01	0.0518
29	DRB1*12:02	0.1048
30	DRB1*13:01	0.0227
31	DRB1*13:02	0.0233
32	DRB1*13:03	0.0029
33	DRB1*13:12	0.0025
34	DRB1*14:01	0.0214
35	DRB1*14:02	0.0013
36	DRB1*14:03	0.0091
37	DRB1*14:04	0.0078
38	DRB1*14:05	0.0193
39	DRB1*14:07	0.0023
40	DRB1*15:01	0.1139
41	DRB1*15:02	0.0418
42	DRB1*15:04	0.0013
43	DRB1*15:05	0.0018
44	DRB1*16:01	0.0029
45	DRB1*16:02	0.0401
46	DRB1*16:05	0.0032
Total		0.9520

alleles. All DRB1 alleles included in the study cover 95.2% all Chinese HLA-DR alleles [28].

We selected epitopes from a combined pool of epitopes through KDS platform using 7 prediction methods from IEDB with a dataset that consisted of 821 circulating HIV genome sequences in China and 46 Chinese HLA-DRB1 alleles. The epitopes bind to MHC class I molecules that were removed first, and then the value of IC_{50} was considered. Next, we ranked all epitopes based on the coverage score (the higher the better coverage in HIV genome and DR-HLA alleles). After an epitope has been selected, it was removed from the database before next selection. This process was repeated until 20 epitopes were selected. The workflow diagram of this procedure was illustrated in Figure 1.

3. Results

3.1. The Coverage Distribution of HLA-DR of Chinese Population. A total of 46 HLA-DR alleles were identified from AFND (Table 1). The alleles were listed and its coverage in Chinese population was given. The table showed the coverage ranged from 0.1% (DRB1∗08:09) to 6.77% (DRB1∗07:01) and in a total of 95.2% of the Chinese population. The sample population comprises 1704 individuals of the Han ethnicity. This information was obtained from ten regions within the mainland, China, and two other regions, Hong Kong and Singapore, where Chinese ethnicity dominates. Among them, the DRB1-02, -05, and -06 genotypes were not detected.

3.2. The Diversity of Epitope Coverage. With a combination of 7 existing epitope prediction methods in IEDB, using database comprised of 46 different DRB1 alleles and 821 full genome sequences of HIV isolates circulating in China, we then predicted 38,460,402 potential epitopes. After duplicates were removed, 21,007,527 potential epitopes remained. We scored these epitopes based on the allele coverage and total coverage score, which was in general normally distributed. As shown in Figure 2, most epitopes displayed low coverage scores, 0.1 or lower; the highest epitope count reached approximately 3000.

3.3. HIV Epitopes Specifically for Chinese Population. By using the methods described above, we obtained 20 epitopes, which in theory covered all 46 DRB1 allelic genotypes and 821 Chinese HIV sequences (Table 2). All 20 epitopes were selected for binding to MHC class II and absence of binding to MHC class I. Table 2 listed the amino acid sequences of the 20 epitopes, their location in HIV-1 gene, their percentage of coverage in HIV-1 genome sequences from 4% to 43%, the proportion in the HLA-DR allele sequences between 52% and 100%, and the total coverage in both sequences as low as 4% and the highest at 41%. One single epitope WIILGLNKIVRMYSP covered 41% of both DRB1 and Chinese specific HIV-1 genome sequences, which is of note. This epitope had been reported before [18]. In fact, 4 other predicted epitopes (LNKIVRMYSPT-SILD, GFPVRPQVPLRPMTY, VDRFYKTLRAEQASQ, and LYKYKVVKIEPLGVA) have also been published previously [22–24, 26] and 4 peptide sequences (PVVSTQLLLNGSLAE,

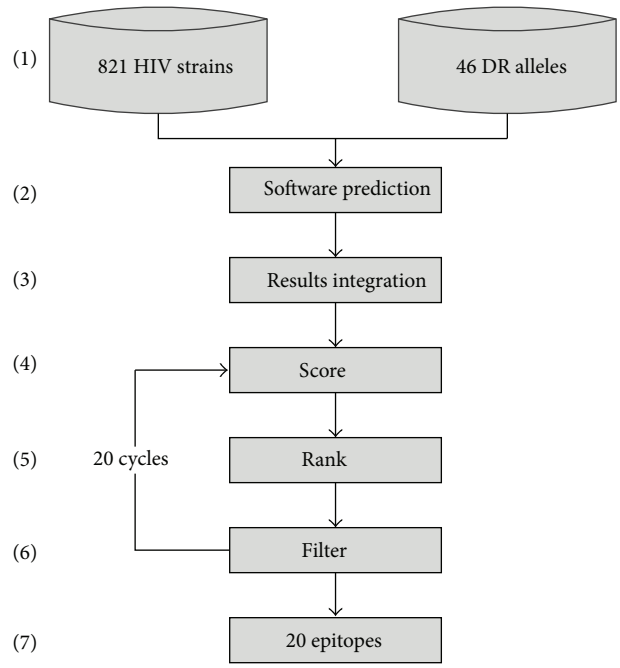


Figure 1: A flowchart illustrates procedures for CD4+T cell epitope prediction. (1) Using KDS platform with datasets of 821 circulating HIV-1 strains and 46 HLA-DRB1 alleles in Chinese population; (2) the software predicted possible epitopes by 7 known methods from the IEDB database; (3) all results were combined and scored using (I) and (II); (4) the epitopes were ranked according to the score; (5) the epitope with the top score and the lowest IC_{50} value was selected; (6) the selected epitope was then removed from the epitope pool; (7) steps 4–6 were repeated until all 20 epitopes fulfilled the criteria that were selected.

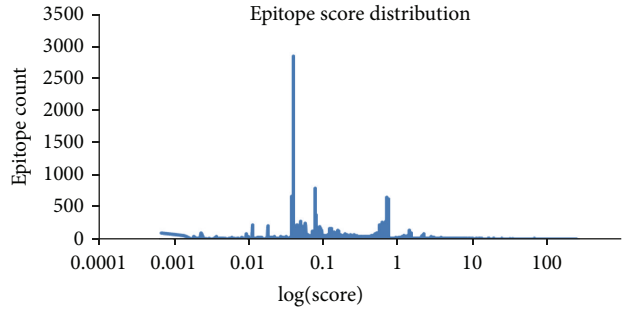


Figure 2: The distribution of epitope coverage score. The epitope coverage scores (log-transformed) were plotted on the horizontal axis against the frequency of epitope count on the vertical axis. Most of the log score localized to the region between 0.01 and 1.

LRIIFAVLSIVNRVR, ILDLWVYHTQGYFPD, and YKRWIILGLNKIVRM) were reported in patents before [19–21, 25], whereas the remaining 11 epitopes have never been reported. All 20 epitopes together provided 98.1% coverage in HIV genome and HLA-DR alleles. These predicted epitopes were found in HIV-1 *gag*, *env*, *pol*, and *nef* genes. Six of them were in *gag* gene, 6 in *env*, 2 in *pol*, and 6 in *nef*.

We then applied the new method to a previously published HIV vaccine comprised of T helper epitopes and tested in clinical trial [13]. The table listed 17 epitopes, from *gag*, *pol*, *env,* and *vpu* genes. One published epitope that has a HIA binding IC_{50} above our threshold of 500, Env 566 IKQFINMWQEVKAMY, was not listed. For these epitopes, HIV coverage is from 2% to 43%, DR coverage is between 35% and 98%, and specific coverage is at highest of 41% and in sum of 69%.

4. Discussion

In this paper, we described a novel method for designing a peptide-based T helper cell vaccine for HIV, which is specific for Chinese HIV strains and Chinese MHC class II genotypes. The current method has several advantages. First, our methodology of epitope prediction is easily accessible to public use. In fact, it is a combination of all seven existing methods publically available in IEDB. The IEDB database comprises a series of most up-to-date and evidence based methods specifically created for the prediction of MHC restricted T cell epitopes. In contrast to other studies that only used one of the methods, we used them all for more accurate prediction of MHC class II restricted T helper epitopes.

So far, there are three major types of bioinformatics methods for the prediction of MHC class II restricted T helper cell epitopes. One is called matrix alignment algorithm, and these are SMM, ARB, and Sturniolo methods. This algorithm uses published T cell epitopes and their respective binding affinity to MHC class II, in terms of the IC_{50} value, to determine epitopes. The other relies on machine learning, and NN-align and NetMHCIIpan methods belong to this category. New sequences are subjected to computer simulated models to predict whether any epitopes can bind to a particular MHC II to high enough affinity. The third type combines several methods together to predict epitopes. These include Consensus method and ComLib method.

Consensus method was reported to provide highest true positive rate, followed by NN-align and ARB [37]. NetMHCIIpan performed the best among all other pan-specific methods for MHC class II with varied experimental settings [38]. NN-align performs especially well in handling large dataset among all other machine learning methods and in combination with ARB outperforms the use of NN-align alone [30].

In this study, we used all above seven methods simultaneously, scored the potential epitopes independently, and then used IC_{50} value as a filter to select T cell epitopes that have the broadest population coverage. Our method did not use all 8 IEDB recommended methods but integrated 7 of the IEDB methods because the 8th IEDB method is an integration of the other seven and thus not an independent measurement. The method we used could be considered as "greedy" algorithm in the bioinformatics field, which predicts the best epitope among all in a pool of potential epitopes. Thus, we believe an integrated method that uses a combination of all seven original algorithms might be the best to predict more accurately MHC class II epitopes.

Another unique feature of our study is that we designed candidate helper T cell vaccine targets specifically to the

Table 2: Predicted HIV T helper cell epitopes for Chinese population.

Amino acid sequences	Protein destination[1]	HIV%[2]	DR%[3]	Total coverage[4]	Specific coverage[5]	Reference[6]
WIILGLNKIVRMYSP	Gag 265	43%	89%	41%	40.72%	Younes et al., 2003 [18]
PVVSTQLLLNGSLAE	Env 262	38%	74%	34%	27.73%	August et al., 2013 [19]
VQMAVFIHNFKRKGG	Pol 892	24%	93%	23%	9.79%	NA (IEDB)
LRIIFAVLSIVNRVR	Env 702	24%	96%	26%	4.39%	Sette et al., 2005 [20]
ILDLWVYHTQGYFPD	Nef 127	12%	63%	10%	5.37%	Sette et al., 2002 [21]
LNKIVRMYSPTSILD	Gag 284	25%	100%	25%	1.81%	Korber et al., 2001 [22]
WGIKQLQARVLAVER	Env 588	22%	87%	20%	1.25%	NA
GAFDLSFFLKEKGGL	Nef 91	4%	63%	4%	1.20%	NA
VDRFYKTLRAEQATQ	Gag 297T	15%	98%	15%	1.17%	NA
GFPVRPQVPLRPMTY	Nef 85	9%	65%	6%	0.84%	Korber et al., 2002 [23]
TPGIRYQYNVLPQGW	Pol 295	23%	93%	22%	0.78%	NA
VDRFYKTLRAEQASQ	Gag 297S	15%	98%	15%	0.71%	Bozzacco et al., 2012 [24]
RQLLSGIVQQQSNLL	Env 549	27%	83%	26%	0.56%	NA
GLIYSKKRQEILDLW	Nef 117	6%	67%	5%	0.50%	NA
KPCVKLTPLCVTLNC	Env 126	17%	89%	16%	0.28%	NA
YKRWIILGLNKIVRM	Gag 272	43%	89%	41%	0.24%	Sette et al., 2002 [25]
PLTFGWCFKLVPVDP	Nef 144	11%	52%	11%	0.21%	NA
FGWCFKLVPVDPREV	Nef 147	4%	93%	4%	0.24%	NA
CKQIIKQLQPALQTG	Gag 67	8%	98%	8%	0.16%	NA
LYKYKVVKIEPLGVA	Env 489	6%	100%	6%	0.14%	Dzuris et al., 2001 [26]
Total specific coverage[7]					98.1%	

[1]The location of epitopes on HIV viral gene products and the first amino acid of the viral gene product.
[2]The epitope sequence presented in the proportion of 821 HIV genome sequences.
[3]The epitope sequence presented in the proportion of 46 DR alleles.
[4]The ratio of the epitope appeared in both 821 HIV genome and DR allele sequences.
[5]Calculated based on the coverage of the epitope in the rest of the dataset after removing the preceding epitope.
[6]Reference where the epitope had been published. NA: not available in published literature.
[7]Sum of specific coverage for all 20 epitopes.

Chinese population. Most common world circulating HIV subtypes are B and C, and recombinant forms are AE and AG. In contrast, the common subtypes are B and recombinant forms are BC and AE in China [27, 39, 40]. We extracted all 821 subtypes of HIV-1 strains which are mostly subtypes B and C for developing a highly specific vaccine for Chinese population. As T helper cell epitopes are recognized through MHC class II, and that Chinese exhibit divergence DRB1 alleles, we also included 46 published Chinese HLA-DRB1 genotypes into our prediction.

In comparison to a previous paper that selected MHC class II binders according to the binding affinity to multiple HLA-DR subtypes [13], we focused on DRB1 alleles which are most polymorphic among human MHC class II loci and thus directed our study to be more specific and increased possibility to induce T cell responses specifically for Chinese.

One limitation in our study, as shown in Table 1, is that DRB1 genotypes 2, 5, and 6 were not included. This is due to a lack of publication of any information on DRB1∗02, 05, and 06. Therefore, our dataset represents what is currently available; that is, there are only 46 DRB1 alleles in Chinese population.

By using our method, we obtained 20 helper T cell epitopes which covered 98.1% of HIV strains known to have been circulating in China and all Chinese HLA-DR genotypes. There are limited studies that have tested designed peptide T helper vaccine in humans. In a published paper that contains 18 T helper epitopes [13], our combination of epitope predication methods found that these epitopes covered 69% of Chinese HIV genomes (Table 3). In a different population that is predominantly Caucasian, these epitopes combined have a 100% coverage. Thus, the difference in the coverage may suggest our predicting method is more specific for Chinese population, and our epitopes are better potential HIV vaccine candidate for Chinese. Furthermore, 9 epitopes we obtained have been published before and 11 are not. Thus, we both have the empirical evidence to support that our allelic specific peptides have the potential to stimulate T cell responses and new epitopes to suggest that our prediction is innovative.

Table 3: Using novel algorism to calculate the coverage of epitopes in a published T helper vaccine for Chinese population.

Amino acid sequence	Protein destination[1]	HIV%[2]	DR%[3]	Specific coverage[4]
FRKYTAFTIPSINNE	Pol 303	14%	98%	13%
EKVYLAWVPAHKGIG	Pol 711	3%	98%	3%
GEIYKRWIILGLNKI	Gag 294	20%	87%	18%
KRWIILGLNKIVRMY	Gag 298	43%	89%	41%
GAVVIQDNSDIKVVP	Pol 989	21%	57%	12%
YRKILRQRKIDRLID	Vpu 31	2%	89%	2%
QKQITKIQNFRVYYR	Pol 956	19%	98%	19%
SPAIFQSSMTKILEP	Pol 335	11%	93%	11%
QHLLQLTVWGIKQLQ	Env 729	23%	83%	21%
AETFYVDGAANRETK	Pol 619	7%	41%	2%
QGQMVHQAISPRTLN	Gag 171	3%	85%	3%
WAGIKQEFGIPYNPQ	Pol 874	3%	35%	1%
KVYLAWVPAHKGIGG	Pol 712	3%	93%	2%
KTAVQMAVFIHNFKR	Pol 915	24%	83%	22%
EVNIVTDSQYALGII	Pol 674	24%	57%	16%
WEFVNTPPLVKLWYQ	Pol 596	22%	91%	22%
HSNWRAMASDFNLPP	Pol 758	11%	57%	7%
Total specific coverage[5]				69%

[1]The epitopes were selected from a published paper.
Data in columns 2–5 were calculated using the same method as in Table 2.

There was one core epitope WIILGLNKIVRMY, appeared in both studies, showing very high HIV, HLA-DR, and specific coverage. The Gag epitope with two amino acids modification WIILGLNKIVRMYSP was reported to stimulate strong CD4+ T responses [27]; another variant of the same epitope KRWIILGLNKIVRMY exhibited superior HLA-DR binding capacity [13]. Another difference between our study and that published is that our epitopes consisted of those in *nef* gene but not *vpu* gene, whereas Walker's study did not cover *nef* but *vpu*. These comparisons suggest that a vaccine designed predominantly for Caucasian may not be optimal for Chinese population. One epitope, for instance, Env 566 (IKQFINMWQEVKAMY) [13], given in Walker's paper, was not picked up in our study.

Our method predicted epitopes, in theory, together covered 98.1% of HIV-1 genome and Chinese specific DRB1 alleles. In comparison, Walker's study reported 18 T helper cell epitopes that cover 100% of the global population. By using a prediction algorithm which based mostly on HLA supertypes [13]. However, when submitted to our new prediction method, the same epitopes only achieved 69% of coverage of the Chinese population. The discrepancy in methods for prediction may give different results. Further experimental evidence is required to find out whether our method is more accurate.

The allele coverage of DRB1 for Chinese was based on 1704 subjects of whom 1569 were from mainland China and 135 were from Hong Kong and Singapore. All Chinese allele data regarding DRB1 frequencies were extracted from AFND, and all 1704 subjects were Chinese Han ethnics. There is no information on other minor national groups in China available. This may lead to inaccuracy in prediction of helper T cell epitopes for the Chinese. Larger sample size may improve the quality of our prediction.

5. Conclusions

In this study, we report a novel bioinformatics method for designing peptide epitope based T helper vaccine for HIV. We suggest further in vitro and in vivo experiments to be performed to test the immunogenicity of this vaccine and improvement of method of prediction to be made when necessary.

Conflict of Interests

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

Authors' Contribution

Jiayi Shu and Xiaojuan Fan contributed equally to this work.

Acknowledgment

This work is supported by the Grant of National Major Scientific and Technological Special Project for "Significant New Drugs Development" during the Twelfth Five-Year Plan Period (2013ZX10001002002002).

References

[1] Joint, "Global report fact sheet: The global AIDS epidemic," Edited by HIV/AIDS UNPo, 2010, http://www.unaids.org/documents/20101123_FS_Global_em_en.pdf.

[2] Y. X. Yan, Y. Q. Gao, X. Sun et al., "Prevalence of hepatitis C virus and hepatitis B virus infections in HIV-positive Chinese patients," *Epidemiology and Infection*, vol. 139, no. 3, pp. 354–360, 2011.

[3] D. D. Ho, A. U. Neumann, A. S. Perelson, W. Chen, J. M. Leonard, and M. Markowitz, "Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection," *Nature*, vol. 373, no. 6510, pp. 123–126, 1995.

[4] F. J. Palella Jr., K. M. Delaney, A. C. Moorman et al., "Declining morbidity and mortality among patients with advanced human immunodeficiency virus infection," *The New England Journal of Medicine*, vol. 338, no. 13, pp. 853–860, 1998.

[5] A. S. de Groot, B. Jesdale, W. Martin et al., "Mapping cross-clade HIV-1 vaccine epitopes using a bioinformatics approach," *Vaccine*, vol. 21, pp. 27–30, 2003.

[6] A. C. van der Kuyl and B. Berkhout, "The biased nucleotide composition of the HIV genome: a constant factor in a highly variable virus," *Retrovirology*, vol. 9, article 92, 2012.

[7] B. D. Walker and D. R. Burton, "Toward an AIDS vaccine," *Science*, vol. 320, no. 5877, pp. 760–764, 2008.

[8] N. M. Flynn, D. N. Forthal, C. D. Harro et al., "Placebo-controlled phase 3 trial of a recombinant glycoprotein 120 vaccine to prevent HIV-1 infection," *Journal of Infectious Diseases*, vol. 191, no. 5, pp. 654–665, 2005.

[9] S. P. Buchbinder, D. V. Mehrotra, A. Duerr et al., "Efficacy assessment of a cell-mediated immunity HIV-1 vaccine (the Step Study): a double-blind, randomised, placebo-controlled, test-of-concept trial," *The Lancet*, vol. 372, no. 9653, pp. 1881–1893, 2008.

[10] M. J. McElrath, S. C. de Rosa, Z. Moodie et al., "HIV-1 vaccine-induced immunity in the test-of-concept Step Study: a case-cohort analysis," *The Lancet*, vol. 372, no. 9653, pp. 1894–1905, 2008.

[11] A. S. Fauci, M. I. Johnston, C. W. Dieffenbach et al., "HIV vaccine research: the way forward," *Science*, vol. 321, no. 5888, pp. 530–532, 2008.

[12] S. Rerks-Ngarm, P. Pitisuttithum, S. Nitayaphan et al., "Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in Thailand," *The New England Journal of Medicine*, vol. 361, no. 23, pp. 2209–2220, 2009.

[13] L. E. Walker, L. Vang, X. Shen et al., "Design and preclinical development of a recombinant protein and DNA plasmid mixed format vaccine to deliver HIV-derived T-lymphocyte epitopes," *Vaccine*, vol. 27, no. 50, pp. 7087–7095, 2009.

[14] V. Brusic and J. T. August, "The changing field of vaccine development in the genomics era," *Pharmacogenomics*, vol. 5, no. 6, pp. 597–600, 2004.

[15] I. G. Ovsyannikova, R. M. Jacobson, and G. A. Poland, "Variation in vaccine response in normal populations," *Pharmacogenomics*, vol. 5, no. 4, pp. 417–427, 2004.

[16] N. C. Toussaint and O. Kohlbacher, "OptiTope—a web server for the selection of an optimal set of peptides for epitope-based vaccines," *Nucleic Acids Research*, vol. 37, no. 2, pp. W617–W622, 2009.

[17] S. G. E. Marsh, P. Parham, and L. D. Barber, "No. Part 11. HLA-DR," in *The HLA Facts Book*, vol. 1st, pp. 330–390, Academic Press, San Diego, Calif, USA, 2000.

[18] S. Younes, B. Yassine-Diab, A. R. Dumont et al., "HIV-1 viremia prevents the establishment of interleukin 2-producing HIV-specific memory $CD4^+$ T cells endowed with proliferative capacity," *The Journal of Experimental Medicine*, vol. 198, no. 12, pp. 1909–1922, 2003.

[19] J. T. August, G. G. Simon, T. W. Tan, A. M. Khan, and Y. Hu, "Human immunodeficientcy virus (HIV-1) highly conserved and low variant sequences as targets for vaccine and diagnostic applications," United States Patent Application US2013/0195904 A1, National University of Singapore, The Johns Hopkins University, edited by Office USP, 2013.

[20] A. Sette, J. Sidney, S. Southwood et al., *Inducint Cellular Immune Responses to Human Immunodeficiency Virus-1 Using Peptide and Uncleic Acid Compositions*, vol. 20050271676, Epimmune Inc, San Diego, Calif, USA, 2005, Edited by Office USP.

[21] A. Sette, J. Sidney, and S. Southwood, *HLA Class I and II Binding Peptides and Their Uses*, vol. WO2003040165A2, Epimmune Inc., New York, NY, USA, 2002.

[22] B. T. M. Korber, R. Koup, B. D. Walker et al., *HIV Molecular Immunology*, vol. LA-UR 02, Theoretical Biology and Biophysics, Los Alamos, NM, USA, 2001, edited by J. A. Bradac.

[23] B. T. M. Korber, R. Koup, B. D. Walker et al., "HIV molecular immunology," in *Theoretical Biology and Biophysics*, J. A. Bradac, Ed., 2002.

[24] L. Bozzacco, H. Q. Yu, J. Dengjel et al., "Strategy for identifying dendritic cell-processed CD4+ T cell epitopes from the HIV Gag p24 protein," *PLoS ONE*, vol. 7, no. 7, Article ID e41897, 2012.

[25] A. Sette, J. Sidney, and S. Southwood, "Identification of broadly reactive DR restricted epitopes," Tech. Rep. WO 1999061916 A1, Epimmune, San Diego, Calif, USA, 2002.

[26] J. L. Dzuris, J. Sidney, H. Horton et al., "Molecular determinants of peptide binding to two common rhesus macaque major histocompatibility complex class II molecules," *Journal of Virology*, vol. 75, no. 22, pp. 10958–10968, 2001.

[27] C. Kuiken, B. Korber, and R. W. Shafer, "HIV sequence databases," *AIDS Reviews*, vol. 5, no. 1, pp. 52–61, 2003.

[28] F. F. Gonzalez-Galarza, S. Christmas, D. Middleton, and A. R. Jones, "Allele frequency net: A database and online repository for immune gene frequencies in worldwide populations," *Nucleic Acids Research*, vol. 39, no. 1, pp. D913–D919, 2011.

[29] R. Vita, L. Zarebski, J. A. Greenbaum et al., "The immune epitope database 2.0," *Nucleic Acids Research*, vol. 38, pp. D854–D862, 2010.

[30] P. Wang, J. Sidney, Y. Kim et al., "Peptide binding predictions for HLA DR, DP and DQ molecules," *BMC Bioinformatics*, vol. 11, article 568, 2010.

[31] M. Nielsen and O. Lund, "NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction," *BMC Bioinformatics*, vol. 10, article 296, 2009.

[32] M. Nielsen, C. Lundegaard, and O. Lund, "Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method," *BMC Bioinformatics*, vol. 8, article 238, 2007.

[33] T. Sturniolo, E. Bono, J. Ding et al., "Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices," *Nature Biotechnology*, vol. 17, no. 6, pp. 555–561, 1999.

[34] H. H. Bui, J. Sidney, B. Peters et al., "Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications," *Immunogenetics*, vol. 57, no. 5, pp. 304–314, 2005.

[35] M. Nielsen, C. Lundegaard, T. Blicher et al., "Quantitative predictions of peptide binding to any HLA-DR molecule of known sequence: NetMHCIIpan," *PLoS Computational Biology*, vol. 4, no. 7, Article ID e1000107, 2008.

[36] Q. Lu, P. Hao, V. Curcin et al., "KDE bioscience: platform for bioinformatics analysis workflows," *Journal of Biomedical Informatics*, vol. 39, no. 4, pp. 440–450, 2006.

[37] P. Wang, J. Sidney, C. Dow, B. Mothé, A. Sette, and B. Peters, "A systematic assessment of MHC class II peptide binding predictions and evaluation of a consensus approach," *PLoS Computational Biology*, vol. 4, no. 4, Article ID e1000048, 2008.

[38] L. M. Zhang, K. Udaka, H. Mamitsuka, and S. F. Zhu, "Toward more accurate pan-specific MHC-peptide binding prediction: a review of current methods and tools," *Briefings in Bioinformatics*, vol. 13, no. 3, pp. 350–364, 2011.

[39] X. L. Yu, L. Yuan, Y. Huang et al., "Susceptibility of HIV-1 subtypes B′, CRF07_BC and CRF01_AE that are predominantly circulating in China to HIV-1 entry inhibitors," *PLoS ONE*, vol. 6, no. 3, Article ID e17605, 2011.

[40] W. Wang, S. Jiang, S. Li et al., "Identification of subtype B, multiple circulating recombinant forms and unique recombinants of HIV type 1 in an MSM cohort in China," *AIDS Research and Human Retroviruses*, vol. 24, no. 10, pp. 1245–1254, 2008.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 529271, 11 pages
http://dx.doi.org/10.1155/2014/529271

Research Article

RNA-Seq Identifies Key Reproductive Gene Expression Alterations in Response to Cadmium Exposure

Hanyang Hu, Xing Lu, Xiang Cen, Xiaohua Chen, Feng Li, and Shan Zhong

Department of Medical Genetics, School of Basic Medical Science, Wuhan University, Wuhan 430071, China

Correspondence should be addressed to Shan Zhong; zhongshan@whu.edu.cn

Received 14 April 2014; Revised 7 May 2014; Accepted 7 May 2014; Published 27 May 2014

Academic Editor: Leng Han

Cadmium is a common toxicant that is detrimental to many tissues. Although a number of transcriptional signatures have been revealed in different tissues after cadmium treatment, the genes involved in the cadmium caused male reproductive toxicity, and the underlying molecular mechanism remains unclear. Here we observed that the mice treated with different amount of cadmium in their rodent chow for six months exhibited reduced serum testosterone. We then performed RNA-seq to comprehensively investigate the mice testicular transcriptome to further elucidate the mechanism. Our results showed that hundreds of genes expression altered significantly in response to cadmium treatment. In particular, we found several transcriptional signatures closely related to the biological processes of regulation of hormone, gamete generation, and sexual reproduction, respectively. The expression of several testosterone synthetic key enzyme genes, such as Star, Cyp11a1, and Cyp17a1, were inhibited by the cadmium exposure. For better understanding of the cadmium-mediated transcriptional regulatory mechanism of the genes, we computationally analyzed the transcription factors binding sites and the mircoRNAs targets of the differentially expressed genes. Our findings suggest that the reproductive toxicity by cadmium exposure is implicated in multiple layers of deregulation of several biological processes and transcriptional regulation in mice.

1. Introduction

Cadmium is an environmental and occupational toxic heavy metal that is widely used in industrial process and consumer products. The usual pattern of the nonoccupational cadmium intake is mainly from food, drinking water, and smoking [1], which caused several diseases by toxically targeting the lung, liver, kidney, bone, and the cardiovascular system as well as the immune and the reproductive system [2]. The multiple mechanisms involved in response to cadmium include modulating cell cycle, DNA repair process, DNA methylation status, altering gene expression, and several signaling pathways in carcinogenesis and other diseases [3–5].

It has been well documented that cadmium exposure leads to the impairment of male and female reproductive system both in human and animals. The high level cadmium in the serum and seminal fluid positively correlated to the azoospermia in the infertile Nigerian males [6]. In the female reproductive system, cadmium exposure leads to failure to ovulate, defective steroidogenesis, suppressed oocyte maturation, and failure of developmental progression in preimplantation embryo and implantation. Moreover, by using the established animal model, the cadmium exposure causing the reproductive system damage is associated with a series of abnormalities, including disruption of blood-testis barrier, testicular necrosis, disruption of blood-epididymis barrier, and abnormal sperm morphology [7]. In addition to reproductive system, recent study also indicated that prenatal cadmium exposure perturbs the vascular and immune system of the murine offspring [8, 9], implicating the role of cadmium exposure in offspring health. A number of mechanisms of reproductive toxicity of cadmium have been suggested, including ionic and molecular mimicry, interference with cell adhesion and signaling, oxidative stress, genotoxicity, and cell cycle disturbance [7]. Although the DNA microarray was employed to study the transcriptional gene alternation in cell lines and peripheral blood cells exposed to cadmium and identified several differentially expressed genes as well as signaling pathways [10, 11], the comprehensive understanding

of the mechanisms that are responsible for the toxicity of chronic cadmium exposure in testis is still lacking.

In this study, we performed the RNA-seq to profile the alterations of gene expression in response to chronic cadmium exposure. By analyzing the transcriptome between the cadmium treated and untreated mice, we identified a number of transcriptional signatures, which provided mechanistic insight into the mechanism of how the male reproductive system is affected by chronic cadmium exposure.

2. Materials and Methods

2.1. Chemicals. Cadmium chloride ($CdCl_2$) was purchased from Sigma Chemical Co. (St. Louis, MO).

2.2. Animals and Experimental Design. Thirty six-week-old male Chinese Kun Ming (KM) mouse weighing about 30–32 g were used in the experiment. The animals were obtained from Wuhan University Center for Animal Experiments/A3-Lab. All animals were housed in a laboratory-controlled environment (25°C, 50% humidity, and light : dark cycle 12 h : 12 h). The animals were permitted free access to food and drinking water *ad libitum*. The food for mice was purchased from Wuhan University Center for Animal Experiment. All animal experiments were carried out in strict accordance with the recommendations in the Regulations for the Administration of Affairs Concerning Experimental Animals of China. The protocol was approved by the Wuhan University Center for Animal Experiment (approved permit number: SCXK 2008-0004). All surgery was performed under sodium pentobarbital anesthesia, and all efforts were made to minimize suffering.

After acclimatization and one-week observation, we found the daily food consumption per mouse was about 6–8 g. Then all animals were randomly divided equally into three groups and every five mice were housed in a cage. To calculate the consumption of food containing cadmium, we make high-cadmium food as 0.3 mg $CdCl_2$/g and low-cadmium food as 0.15 mg $CdCl_2$/g. Then, each cage of high level cadmium-exposed group was supplied with 50 g high-cadmium food daily, and low level cadmium-exposed group was supplied with 50 g low-cadmium food as well. On the following day, we collected and removed the remaining food and residue and new 50 g food was given. By deducting the weights of remaining food and residue, we calculated that the food intake for cadmium treated mice was 6.5 ± 0.8 g. Therefore, the intake of cadmium in the high group was considered as 1.95 ± 0.24 g per mouse daily and in the low group was considered as 0.975 ± 0.12 g per mouse daily. Subsequently, all animals were treated with the according food for 6 months, in which the mice of control group were treated with the same amount of food without $CdCl_2$. We chose these specific dose and period of cadmium exposure as our preliminary results showed that cadmium accumulation in serum and testis reached the highest value at six months, and the mice treated with such dose of cadmium did not show abnormalities or other health problems. Then the blood samples of 5 different mice in each group were randomly collected. Serum were separated by centrifugation from the blood samples above and stored at −80°C until assay for the cadmium concentration using the graphite furnace atomic absorption spectrometry (GFAAS) method. Testosterone in serum was measured using Testosterone Parameter Assay kit (R&D, USA).

2.3. RNA Isolation and Preparation for Next-Generation Sequencing. A total of 9 testis samples (three samples from each group) were selected for RNA isolation. Total RNA was isolated using Trizol Reagent (Invitrogen) according to the manufacturer's instructions. Then these RNA samples were sent to Analytical & Testing Center at Institute of Hydrobiology, Chinese Academy of Sciences (http://www.ihb.ac.cn/fxcszx/) for the verification of RNA integrity. Then one RNA sample from each group was collected for pair-ends transcriptome sequencing under the Illumina Genome Analyzer IIx platform. The sequencing data have been deposited in NCBI Sequence Read Archive (SRA, http://www.ncbi.nlm.nih.gov/Traces/sra/) with the accession number SRP032958.

2.4. Read Alignment with TopHat and RNASEQR. Raw data were mapped to the mouse reference genome (mm10 downloaded from UCSC) using TopHat (version 2.0.3) and RNASEQR (version 1.0.2) software, respectively [12, 13]. Prebuilt genomic indices were created by bowtie and provided to the alignment software for reads mapping. TopHat removes a few low-quality score reads then aligns the reads that are directly mapped to the reference genome. It then determines the possible location of gaps in the alignment based on splice junctions flanking the aligned reads and uses gapped alignments to align the reads that were not aligned by Bowtie in the first step. Compared with other alignment tools, RNASEQR takes advantages of annotated transcripts and genomic reference sequences to obtain high quality mapping result by the three sequential processing steps. It aligns the RNA-seq sequences to a transcriptomic reference firstly, then detects novel exons and identifies novel junctions using an anchor-and-align strategy finally. The output of Tophat can be the input of Cufflinks, while the output of RNASEQR can be converted as the input for Cufflinks using the convert_RNASEQRSAM_to_CufflinkSAM.py script.

2.5. Transcriptome Reconstruction. Aligned reads from TopHat and RNASEQR were assembled by Cufflinks (version 2.0.2), an *ab-initio* tancscriptome assembler that reconstructs the transcriptome based on RNA-seq reads aligned to the genome with a spliced read aligner. To obtain transcriptome assemblies from the aligned reads, we run Cufflinks with default parameters. After that, normalized expression levels were estimated and reported as FPKM (Fragments Per Kilobase of exon per Million fragments mapped). As several assembled transcripts were obtained from each sample, we used Cuffmerge to assemble them into a comprehensive set of transcripts for further downstream differential expression analysis.

Table 1: Primers used for qPCR validation.

Gene	Primer sequence (5′-3′)	Target size (bp)	Tm (°C)
Actin, beta	Forward: CTGTCGAGTCGCGTCCACCC Reverse: ACATGCCGGAGCCGTTGTCG	128	59
Cyp11a1	Forward: AGATCCCTTCCCCTGGTGACAATG Reverse: CGCATGAGAAGAGTATCGACGCATC	192	60
Cyp17a1	Forward: CCAGGACCCAAGTGTGTTCT Reverse: CCTGATACGAAGCACTTCTCG	250	59
Prm2	Forward: CAAGAGGCGTCGGTCA Reverse: TGGCTCCAGGCAGAATG	167	59
Tex15	Forward: ATTTGAGTGGCACAGAC Reverse: AGTATTGGGATTTGGAG	194	59
Adam9	Forward: CGCTTAGCAAACTACCTG Reverse: TTCCCGCCACTGAACAA	147	59
Dazl	Forward: GGAGGCCAGCACTCAGTCTTC Reverse: AGCCCTTCGACACACCAGTTC	184	60

2.6. Differential Expression Analysis. In order to determine the differentially expressed transcripts within the dataset, we used Cuffdiff, a separate program included in Cufflinks, to calculate expression in two or more samples and test the statistical significance of each observed change in expression between them. Cuffdiff reports numerous output files containing the results of the differential analysis of the samples, including genes and transcripts expression level changes, familiar statistics such as $\log_2$-fold change, P values (both raw and corrected for multiple testing), and gene-related attributes such as common name and genome location. The differentially expressed genes were clustered and visualized by TreeView [14]. Functional annotations of these genes were performed by the DAVID bioinformatics resources [15] and WEB-based GEne SeT AnaLysis Toolkit (WebGestalt) [16]. Downstream enrichment analyses such as TF binding sites in promoter regions and microRNA sites in 3′-UTRs were performed using Expander (version 6.0.5) [17], and the miRNA-target gene network was constructed by Cytoscape [18].

2.7. qPCR. qPCR analyses were performed to validate the results of RNA-seq. The reverse transcription is synthesized using RevertAidTM First Strand cDNA Synthesis Kit from Fermentas according to the manufacturer's instructions. The PCR primers were designed with Primer Premier 5.0 software and β-Actin was used as a reference gene. The primer sequences, melting temperatures, and product sizes are shown in Table 1. qPCR was performed on iQ5 Real Time PCR Detection System (Bio-Rad) (Bio-Rad, USA) using SYBR Green Realtime PCR Master Mix (TOYOBO CO., LTD, Japan) as the readout. Three independent biological replicates of the control and cadmium treated groups were included in the analysis and all reactions were carried out in triplicates. Data was analyzed by the 2-ΔΔCT method [19].

2.8. Statistical Analysis. Besides those statistical tools embedded in the bioinformatics software and resources, additional statistical analyses were performed using GraphPad Prism (Version 5.00). Cadmium treated groups were compared with the control groups by unpaired Student's t-test. $P < 0.05$ was considered statistically significant.

3. Results

3.1. Cadmium Accumulated in Mouse and Inhibited the Testosterone. After six-month cadmium-exposure treatment, all animals survived with the slight loss of body weight for the treated groups (Figure 1(a)) and did not show abnormalities or other health problems. We first tested the concentration of cadmium in serum and testis for each of the groups. As expected, cadmium concentration was significantly increased in proportion to the exposure level both in serum and testis (Figures 1(b)-1(c)). We examined the level of serum testosterone, which is essential for normal spermatogenesis and other reproductive processes. Results showed that cadmium exposure significantly reduced the level of serum testosterone in high dose cadmium treated mouse compared with the low and control mouse (Figure 1(d)).

3.2. The Next-Generation Transcriptome Sequencing. To determine the molecular events during cadmium exposure, we performed RNA-seq on testis samples of treated and untreated mice. Raw data of 80 million, 36-bp pair-ends reads were obtained and aligned to the mouse reference genome by TopHat and RNASEQR software [12, 13], resulting in, on average, 71% of the reads mapped to the reference genome with 57% unique mapped reads. Then, both the unique and multiple mapped reads were kept as Cufflinks can handle the multimapped reads by uniformly dividing each multimapped read to all of the positions it maps to and calculating initial abundance. Then the mapped reads were assembled for reconstructing transcriptome and estimating expression abundance by Cufflinks. The results of estimated normalized expression levels were reported as FPKM (Fragments Per Kilobase of exon per Million fragments mapped). Overall, a total of 27600 transcripts on average were successfully reconstructed from each

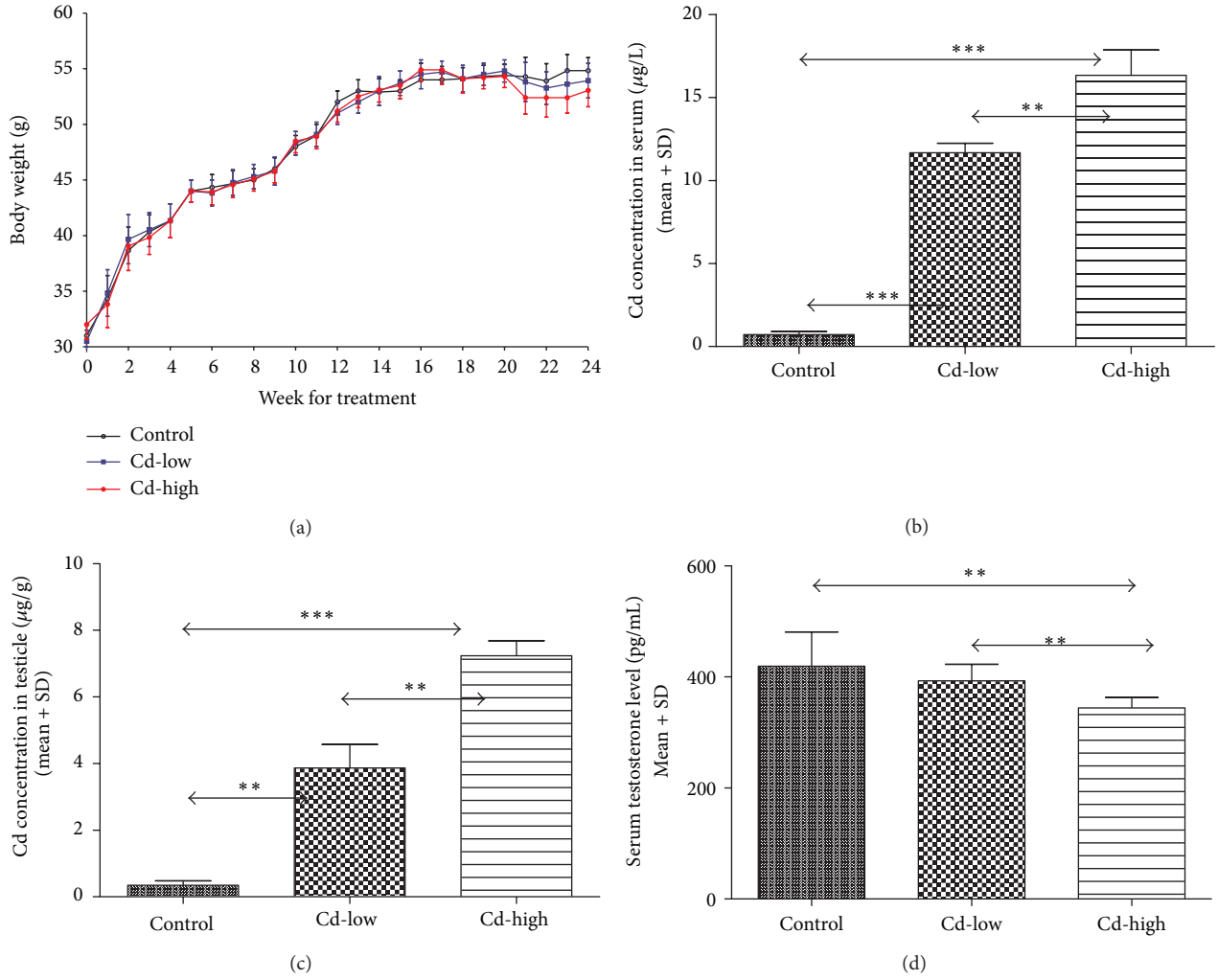


FIGURE 1: Body weights, level of cadmium in serum and testis, and serum testosterone. Mice were treated with different level of $CdCl_2$ in their rodent chow for six months. (a) Body weights were measured per week during cadmium treatment. The level of accumulated cadmium in serum (b) and testis (c) were measured using the graphite furnace atomic absorption spectrometry (GFAAS) method. The levels of serum testosterone (d) were measured by ELISA. Data was shown as mean ± standard deviation ($n = 5$). ($^{**}P < 0.01$, $^{***}P < 0.001$, Student's t-test).

group and mapped to the annotated genomic loci. For example, the genes on chr11:69,594,778–70,305,335 were reconstructed (Figure 2(a)) and the intergenic transcribed regions were pervasively detected compared with the mouse gene reference annotations (Figure 2(b)). Figure 2(c) shows a representative example histogram of read coverage versus a genomic loci containing the adam24 gene.

3.3. Differentially Expressed Genes Analysis. After reads mapping, transcripts assembling, and expression level calculating, we next sought to identify differentially expressed genes between samples with different treatment. By using Cuffdiff, two expression profiles were obtained from different mapping software. We then compared them and combined the overlapped differentially expressed genes. Genes with q value < 0.05 were considered to be differentially expressed. Finally, a total of 830 genes were identified as differentially expressed (Table S1 available online at http://dx.doi.org/10.1155/2014/529271). In detail, there were 283 differentially expressed genes between high and low groups (103 upregulated and 180 downregulated), 401 genes between high and control groups (137 upregulated and 264 downregulated), and 145 genes between low and control groups (61 upregulated and 84 downregulated), respectively (Figure 3(a)). All the differentially expressed transcripts were hierarchically clustered and the results showed that the distinct gene expression pattern was associated with cadmium exposure level, although the low and control groups exhibited a more similar expression pattern than the high group, probably because we used quite low dose of cadmium to treat the low group (Figure 3(b)).

In order to gain a comprehensive impact assessment of cadmium exposure on testicular gene expression, all biochemical pathways that altered in response to cadmium exposure were identified by comparing the ontology of all the

Figure 2: Testicular transcriptome reconstruction of RNA-seq under mouse reference annotation. (a) The genes on chr11:69,594,778–70,305,335 were reconstructed as examples. (b) The intergenic transcription was detected beyond the reference annotation. (c) Read coverage of Adam24 gene on chr8 was shown.

genes differentially expressed between samples. Here, 373 differentially expressed genes annotated by UCSC and Ensembl of the 830 genes were performed with gene ontology enrichment analysis and functional classification. These genes were classified into several ontology categories according to their function in various biological processes (Figure 3(c)). Consistent with previous reports, some ontology categories that are implicated in cadmium toxicity to testis were confirmed in our study, such as genes involved in immunity, cell cycle, toxin, oxidation reduction, and metabolism [7].

Notably, the most enriched ontology category contains the genes associated with regulation of transcription. Genes that are involved in many classical signaling transduction pathways are modulated, such as Nfat5, E2f2, Fos, Junb, Notch1, and Stat4. In addition, we observed that abnormal epigenetic regulation occurred during cadmium exposure. Some of the differentially expressed genes involved in DNA methylation and histone modification are those with DNA methyltransferase, histone methyltransferase, acetyltransferase, or deacetylase activities, including Crebbp, Dmap1, Prdm9, Setd2, Prmt7, and Hdac2. Thus, both the transcriptional program and epigenetic patterns are supposed to be misregulated and implicated in cadmium caused reproductive toxicity.

Importantly, we also identified several novel and specific pathways modulated by cadmium exposure, including homeostasis of hormone (Table 2), gamete generation, and sexual reproduction (Table 3). Among these pathways, we noticed that similar functional categories shared the same differentially expressed candidate genes between them.

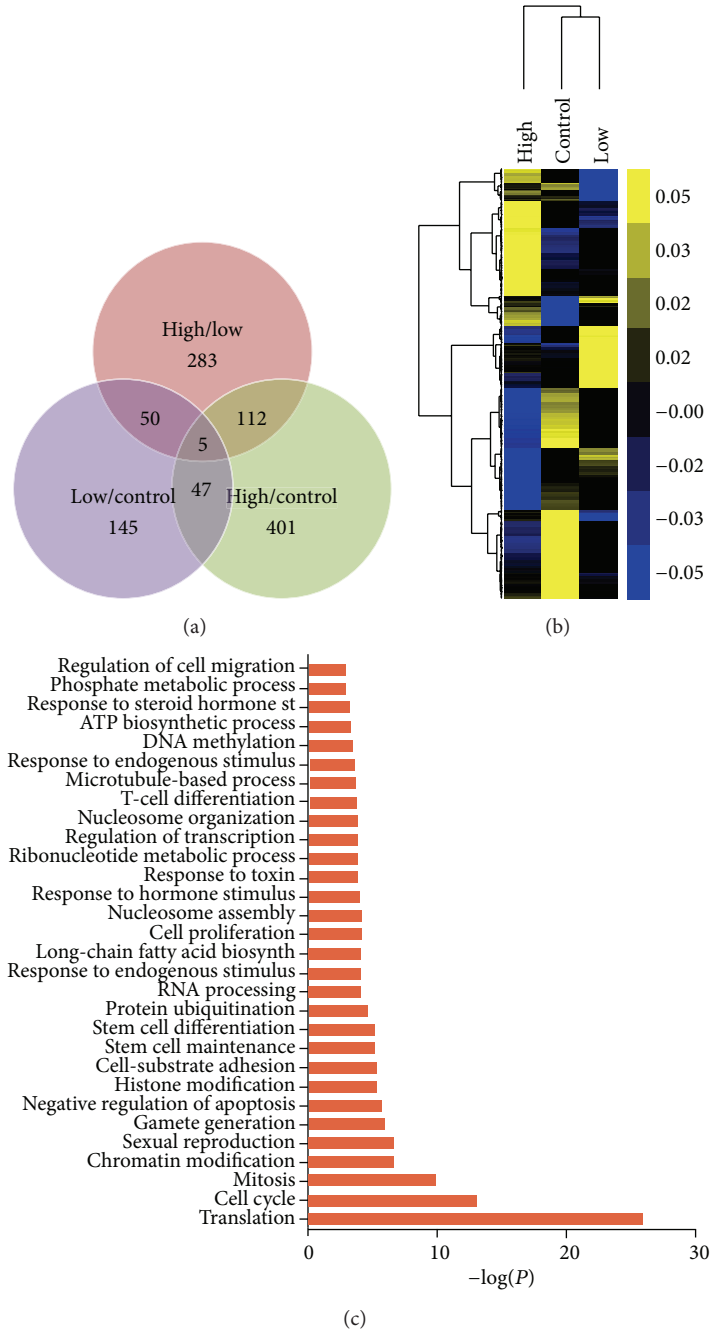


Figure 3: Cadmium modulated differentially expressed genes analysis. (a) The number of genes differentially expressed between the high level cadmium treated, low level cadmium treated, and control groups. (b) Hierarchical clustering analysis of gene expression profiles. Each column represents one mouse, and each horizontal line refers to a gene. Color legend is on the top-left of the figure. Yellow indicates genes with a greater expression relative to the geometrical means; blue indicates genes with a lower expression relative to the geometrical means. (c) Biological process Gene Ontology (GO) analysis of differentially expressed genes.

The reproduction associated functional categories comprise 16 annotated genes, most of which were inhibited due to cadmium exposure. For example, Fndc3a, Dazl, Kitl, Tex15, and Zfx were downregulated with the fold changes ranging from −2.6 to −4.5. Since these genes are critical for spermatogenesis, germ cell development, or junctions between Sertoli cells [20–26], we speculated that cadmium induced testicular toxicity through targeting and downregulating these genes. The hormone related categories contained six genes, half of which were induced. In particular, the rest of three downregulated genes, named Star, Cyp11a1, and Cyp17a1, were specifically responsible for testosterone synthetic. The downregulation of these genes may contribute to the reduction of serum testosterone in response to cadmium exposure. Collectively, these findings suggested that cadmium impaired the reproductive process and spermatogenesisand also potentially modulated

Table 2: Regulation of hormone level related genes.

Gene symbol	Description	Fold change	Q value
Adh1	Alcohol dehydrogenase 1 (class I)	3.806345	0.005
Cyp11a1	Cytochrome P450, family 11, subfamily a, polypeptide 1	−7.197441	$5.57E-08$
Cyp17a1	Cytochrome P450, family 17, subfamily a, polypeptide 1	−4.438219	0.0001
Ren1, Ren2	Renin 1 structural; similar to renin 2 tandem duplication of Ren1; renin 2 tandem duplication of Ren1	7.678866	0.016
Retsat	Retinol saturase (all trans retinol 13, 14 reductase)	4.603697	0.025
Star	Steroidogenic acute regulatory protein	−5.377734	$5.57E-05$

Table 3: Regulation of reproductive process related genes.

Gene symbol	Description	Fold change	Q value
Adam24	A disintegrin and metallopeptidase domain 24 (testase 1)	−2.872645	0.038
Adam25	A disintegrin and metallopeptidase domain 25 (testase 2)	−2.736529	0.048
Adam26a	A disintegrin and metallopeptidase domain 26A (testase 3)	−3.429332	0.023
Dazl	Deleted in azoospermia-like	−3.38705	0.01
Fndc3a	Fibronectin type III domain containing 3A	−2.636543	0.044
Kitl	Kit ligand	−4.289398	0.044
Prm2	Protamine 2	1.169034	0
Qk	Similar to Quaking protein; quaking	−3.974115	0.006
Sycp2	Synaptonemal complex protein 2	−2.977031	0.031
Tex15	Testis expressed gene 15	−2.926283	0.0298951
Txndc3	Thioredoxin domain containing 3 (spermatozoa)	−2.973916	0.017
Zbtb16	Zinc finger and BTB domain containing 16	−6.408031	0.011
Zfp37	Zinc finger protein 37	−3.430625	0.004
Zfx	Zinc finger protein X-linked; similar to zinc finger protein	−4.58343	0.027347
Zan	Zonadhesin	−3.605458	0.00578172

the normal hormone levels during the toxic response to cadmium in testis.

Besides, 9 KEGG signaling pathways were affected by cadmium exposure by mapping the differentially expressed genes to the KEGG database. Those modulated signaling pathways were comprised of ribosome, Alzheimer's disease, asthma, oxidative phosphorylation, focal adhesion, ECM-receptor interaction, C21-Steroid hormone metabolism, metabolic pathways, and prostate cancer (Table 4). Among these over-represented pathways, according to functional hierarchies in KEGG, pathways of asthma are associated with the immune system. ECM-receptor interaction corresponds tosignaling molecules and interaction. Focal adhesion and ribosome are associated with cell communication and translation, respectively. Three modulated pathways, C21-Steroid hormone metabolism, oxidative phosphorylation, and metabolic pathways, are associated with metabolism, indicating the basal metabolism of cells in testis are affected. Altogether, our results reflected that multiple cellular and molecular mechanismsare modulated during cadmium exposure.

Table 4: Modulated KEGG pathways.

Pathway name	Number of genes	Corrected P value
Ribosome	17	$5.54E-12$
Alzheimer's disease	11	0.0246
Asthma	2	0.0486
Oxidative phosphorylation	9	0.0419
Focal adhesion	10	$1.73E-4$
ECM-receptor interaction	7	$1.15E-4$
C21-steroid hormone metabolism	2	0.0106
Prostate cancer	6	0.00868
Metabolic pathways	37	0.0059

3.4. Validation of RNA-Seq Data by Quantitative Real Time PCR (qPCR). To confirm the changes in gene expression observed by RNA-seq, we performed qPCR analysis on three reproductions (Prm2, Tex15, and Dazl), two hormones (Cyp11a1 and Cyp17a1) associated with functional categories genes, and a randomly selected gene named Adam9. qPCR results showed that these genes are significantly differentially expressed ($P < 0.05$) and exhibited the similar expression status compared to RNA-seq and conformed that Tex15, Dazl, Cyp11a1, and Cyp17a1 were inhibited by cadmium treatment (Figure 4).

3.5. Transcriptional and Posttranscriptional Control of Cadmium Modulated Genes. In an effort to uncover the potential regulatory mechanism underlying the transcription of

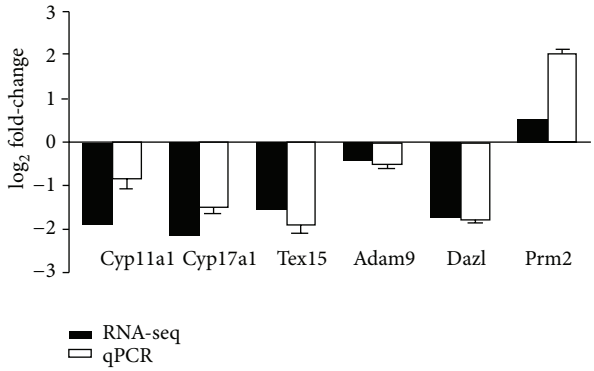


Figure 4: qPCR validation of the RNA-seq data. $\log_2$-fold change determined from the relative Ct values of six genes were compared to those detected by RNA-seq. Replicates ($n = 3$) of each sample were run and the Ct values averaged. All Ct values were normalized to beta-actin. P values of the Q-PCR data are 0.002 (Cyp11a1), 0.02 (Cyp17a1), 0.012 (Tex15), 0.014 (GLRX2), 0.008 (Adam9), 0.016 (Dazl), and 0.014 (Prm2), respectively.

Table 5: Enrichment of transcription factors across the promoter regions of differentially expressed genes.

Transcriptional factors	Number of genes	Corrected P value
ETF	193	$4.37E-16$
NKX3A	178	0.007
Nrf-1	170	$6.53E-4$
HMGIY	140	0.009
SRY	328	$3.49E-4$
ZF5	199	$2.05E-7$
FOXJ2	153	$7.62E-5$
OCT-1	229	$1.12E-4$
E2F-1	226	$2.24E-7$
LUN-1	35	0.043
FOXP1	319	$7.31E-7$
AP2	169	0.014

the cadmium modulated gene sets, we performed transcription factor binding sites analysis within the promoters and microRNA targets analysis of the cadmium modulated genes. Promoter regions for positions of −1000–+200 of the TSS across the cadmium modulated genes were predicted for the binding sites enrichment of several transcription factors ($P < 0.05$, Bonferroni test) (Table 5). Gene ontology analysis of these transcription factors revealed that they were involved in multiple biological processes containing regulation of cell cycle, hormone secretion, organ morphogenesis, reproductive process, neurogenesis, and response to stimulus, which were in accordance with the biological processes associated with the differentially expressed genes regulated by these transcription factors (Figure 5(a)).

We next performed microRNA targets analysis of the differentially expressed genes for further investigating the posttranscriptional control of them. A total of 10 microRNAs were identified as significantly enriched at 3′-UTR region of the differentially expressed genes (FDR < 0.05) (Table 6). By

Table 6: MicroRNAs enriched at 3′-UTR region of the differentially expressed genes.

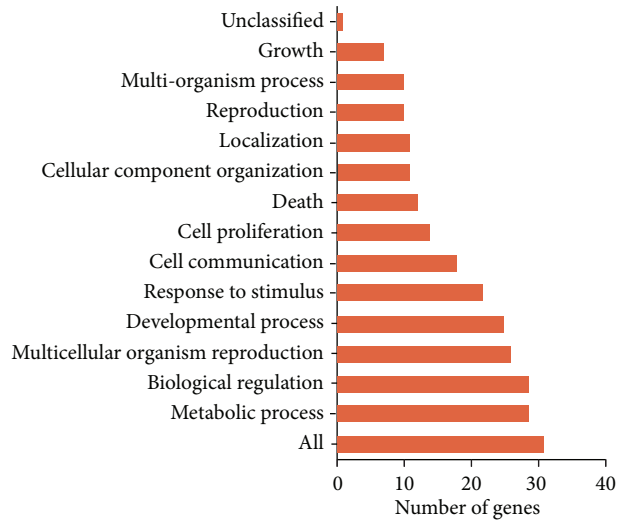
MicroRNA	Number of targeted genes	FDR
Mir-142-3p	6	0.001
Mir-342/342-3p	5	0.0015
Mir-196ab	5	0.0025
Mir-874	4	0.003
Mir-124/506	5	0.0135
Mir-30a	3	0.0075
Mir-124/506	11	0.0205
Mir-153	4	0.0265
Mir-25/32/92/92ab/363/367	3	0.0195
Mir-448	4	0.006

evaluating the microRNAs-Target-Network generated from the above information, it is implicated that a number of altered genes expression in this study may be regulated by the common microRNAs (Figure 5(b)), indicating their potential roles in regulating the reproductive process in response to cadmium exposure.

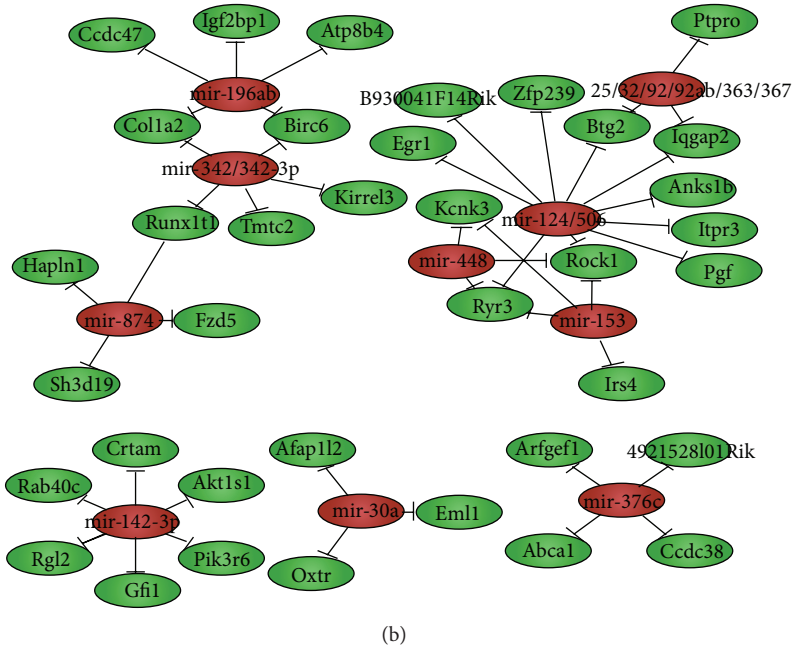
4. Discussion

Cadmium has been suggested to be anenvironmental and occupational toxic heavy metal that causes several diseases and toxically targets the lung, the liver, the kidney, the bone, the cardiovascular system, the immune system, and the reproductive system [2]. Here, in order to uncover the exclusive molecular mechanism of the mouse reproductive toxicity caused by chronic cadmium exposure, we simulated the conditions of cadmium pollution in human by treating the mouse with different doses of cadmium for 6 months. We observed cadmium exposure significantly reduced the level of serum testosterone. As a member of androgens, testosterone brings about its biological functions through associations with androgen receptor (AR), leading to AR transactivation which then results in the modulation of AR downstream gene expressions [27]. While the difference of testosterone seems to be small, such difference may lead to significantly downstream effects through the cascade signal transduction and transcriptional regulation of many genes by androgen receptor.

We next used RNA-seq to analyze the transcriptome of mouse testis affected by cadmium. We found a total of 830 genes and transcripts that were differentially expressed. Gene Ontology analysis revealed that these genes were enriched in several biological processes, in which the genes related to the reproductive process were paid more attention. For example, Fndc3a was reported to be required for adhesion between spermatids and Sertoli cells during mouse spermatogenesis [20]. Loss of Tex15 function causes early meiotic arrest in male mice, and Tex15-deficient spermatocytes exhibit a failure in chromosomal synapsis and DNA double-strand breaks are impaired [21]. In human, the deletion of DAZ



(a)



(b)

Figure 5: Transcriptional regulation analysis of the differentially expressed genes. (a) Biological process gene ontology analysis of the transcription factors that regulate the gene expression. (b) MicroRNA-target gene network. Red circles represent microRNAs; green circles represent the target genes.

cluster is associated with azoospermia and oligospermia in 5–10% of infertile men [22], and disruption of the Dazla gene leads to loss of germ cells and complete absence of gamete production [23]. In mouse, Dazl is required for embryonic development and survival of XY germ cells [24]. Kitl mutant mice exhibited reduced testis size due to aberrant spermatogonial proliferation, affecting the formation of tight junctions between Sertoli cells during postnatal development [25]. The zinc-finger proteins ZFX mutant mice were smaller and less viable and had fewer germ cells than wild-type mice [26]. Altogether, we supposed that the cadmium could cause the impairment of spermatogenesis and testicular toxicity through the repression of these genes. In addition, we identified six hormone related genes that were modulated by cadmium. We observed the decrease of expression for Star, Cyp11a1, and Cyp17a1, which were in accordance with a previous study [28]. As these genes encode the primarily testosterone synthetic enzymes, it is likely that the cadmium perturbed the spermatogenesis through repressing the synthesis of testicular testosterone as well. Combined with other modulated functional categories such as immunity, cell cycle, transcription, epigenetic regulation, and metabolism, the molecular mechanisms of cadmium caused male reproductive toxicity are implicated in multiple layers of deregulation of several biological processes.

Further, we computationally analyzed the transcriptional and posttranscriptional control of the differentially expressed genes. We found several transcriptional factors were enriched

with the binding sites at the promoter regions of some gene sets. These binding events should be verified by further ChIP experiments. While these transcriptional factors were unable to be detected as statistically significantly differentially expressed between the samples, it is likely that the slight change of expression ultimately led to the significant expression change of their targets. We also predicted the microRNAs with the binding possibility of some sets of cadmium modulated genes. We identified 10 microRNAs targeted to the differentially expressed genes, the regulatory roles of which in testis response to cadmium could be explored by their expression patterns and the gain- or loss-of-function studies in the future.

In summary, our study demonstrated that many genes in testis were modulated due to chronic cadmium exposure. In particular, aside from the genes related to the functional categories previously reported, we identified novel pathways and the potential transcriptional regulatory mechanism on the cadmium modulated genes. These findings provide evidence for the elucidation of the molecular mechanism linking the chronic cadmium exposure to the impairment of male reproductive system and the clues for future studies of potential biomarkers and therapeutic targets for cadmium exposure.

Conflict of Interests

The authors declare no conflict of interests.

Acknowledgments

The authors thank Jie Li at the Department of Pharmacology, Basic Medical School of Wuhan University for technical help. They also thank all members in Dr. Zhong's laboratory for helpful suggestions and technical assistance. This work was financially supported by the National Natural Science Foundation of China (Grant no. 31172395) and Key Technologies Research and Development Program of China (Grant no. 2013BAI12B01-3).

References

[1] S. Satarug and M. R. Moore, "Adverse health effects of chronic exposure to low-level cadmium in foodstuffs and cigarette smoke," *Environmental Health Perspectives*, vol. 112, no. 10, pp. 1099–1103, 2004.

[2] B. A. Fowler, "Monitoring of human populations for early markers of cadmium toxicity: a review," *Toxicology and Applied Pharmacology*, vol. 238, no. 3, pp. 294–300, 2009.

[3] M. Waisberg, P. Joseph, B. Hale, and D. Beyersmann, "Molecular and cellular mechanisms of cadmium carcinogenesis," *Toxicology*, vol. 192, no. 2-3, pp. 95–117, 2003.

[4] Y. H. Jin, A. B. Clark, R. J. C. Slebos et al., "Cadmium is a mutagen that acts by inhibiting mismatch repair," *Nature Genetics*, vol. 34, no. 3, pp. 326–329, 2003.

[5] B. Wang, Y. Li, Y. Tan et al., "Low-dose Cd induces hepatic gene hypermethylation, along with the persistent reduction of cell death and increase of cell proliferation in rats and mice," *PLoS ONE*, vol. 7, no. 3, Article ID e33853, 2012.

[6] O. Akinloye, A. O. Arowojolu, O. B. Shittu, and J. I. Anetor, "Cadmium toxicity: a possible cause of male infertility in Nigeria," *Reproductive Biology*, vol. 6, no. 1, pp. 17–30, 2006.

[7] J. Thompson and J. Bannigan, "Cadmium: toxic effects on the reproductive system and the embryo," *Reproductive Toxicology*, vol. 25, no. 3, pp. 304–315, 2008.

[8] A. M. Ronco, M. Montenegro, P. Castillo et al., "Maternal exposure to cadmium during gestation perturbs the vascular system of the adult rat offspring," *Toxicology and Applied Pharmacology*, vol. 251, no. 2, pp. 137–145, 2011.

[9] M. L. Hanson, I. Holásková, M. Elliott, K. M. Brundage, R. Schafer, and J. B. Barnett, "Prenatal cadmium exposure alters postnatal immune cell development and function," *Toxicology and Applied Pharmacology*, vol. 261, no. 2, pp. 196–203, 2012.

[10] S. Dakeshita, T. Kawai, H. Uemura et al., "Gene expression signatures in peripheral blood cells from Japanese women exposed to environmental cadmium," *Toxicology*, vol. 257, no. 1-2, pp. 25–32, 2009.

[11] G.-Y. Li, M. Kim, J.-H. Kim, M.-O. Lee, J.-H. Chung, and B.-H. Lee, "Gene expression profiling in human lung fibroblast following cadmium exposure," *Food and Chemical Toxicology*, vol. 46, no. 3, pp. 1131–1137, 2008.

[12] 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, no. 3, pp. 562–578, 2012.

[13] L. Y. Chen, K.-C. Wei, A. C.-Y. Huang et al., "RNASEQR-a streamlined and accurate RNA-seq sequence analysis program," *Nucleic Acids Research*, vol. 40, no. 6, article e42, 2012.

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

[15] 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.

[16] B. Zhang, S. Kirov, and J. Snoddy, "WebGestalt: an integrated system for exploring gene sets in various biological contexts," *Nucleic Acids Research*, vol. 33, no. 2, pp. W741–W748, 2005.

[17] R. Shamir, A. Maron-Katz, A. Tanay et al., "EXPANDER—an integrative program suite for microarray data analysis," *BMC Bioinformatics*, vol. 6, article 232, 2005.

[18] 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.

[19] 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.

[20] K. L. Obholz, A. Akopyan, K. G. Waymire, and G. R. MacGregor, "FNDC3A is required for adhesion between spermatids and Sertoli cells," *Developmental Biology*, vol. 298, no. 2, pp. 498–513, 2006.

[21] F. Yang, S. Eckardt, N. A. Leu, K. J. McLaughlin, and P. J. Wang, "Mouse TEX15 is essential for DNA double-strand break repair and chromosomal synapsis during male meiosis," *Journal of Cell Biology*, vol. 180, no. 4, pp. 673–679, 2008.

[22] R. Reijo, T.-Y. Lee, P. Salo et al., "Diverse spermatogenic defects in humans caused by Y chromosome deletions encompassing a novel RNA-binding protein gene," *Nature Genetics*, vol. 10, no. 4, pp. 383–393, 1995.

[23] M. Ruggiu, R. Speed, M. Taggart et al., "The mouse Dazla gene encodes a cytoplasmic protein essential for gametogenesis," *Nature*, vol. 389, no. 6646, pp. 73–77, 1997.

[24] Y. Lin and D. C. Page, "Dazl deficiency leads to embryonic arrest of germ cell development in XY C57BL/6 mice," *Developmental Biology*, vol. 288, no. 2, pp. 309–316, 2005.

[25] S. Deshpande, V. Agosti, K. Manova, M. A. S. Moore, M. P. Hardy, and P. Besmer, "Kit ligand cytoplasmic domain is essential for basolateral sorting in vivo and has roles in spermatogenesis and hematopoiesis," *Developmental Biology*, vol. 337, no. 2, pp. 199–210, 2010.

[26] S.-W. Luoh, P. A. Bain, R. D. Polakiewicz et al., "Zfx mutation results in small animal size and reduced germ cell number in male and female mice," *Development*, vol. 124, no. 11, pp. 2275–2284, 1997.

[27] H. V. Heemers and D. J. Tindall, "Androgen receptor (AR) coregulators: a diversity of functions converging on and regulating the AR transcriptional complex," *Endocrine Reviews*, vol. 28, no. 7, pp. 778–808, 2007.

[28] Y.-L. Ji, H. Wang, P. Liu et al., "Pubertal cadmium exposure impairs testicular development and spermatogenesis via disrupting testicular testosterone synthesis in adult mice," *Reproductive Toxicology*, vol. 29, no. 2, pp. 176–183, 2010.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 969768, 8 pages
http://dx.doi.org/10.1155/2014/969768

Research Article

Stratification of Gene Coexpression Patterns and GO Function Mining for a RNA-Seq Data Series

Hui Zhao,[1,2,3,4] Fenglin Cao,[1,2,3] Yonghui Gong,[4] Huafeng Xu,[5] Yiping Fei,[1,2,3] Longyue Wu,[1,2,3] Xiangmei Ye,[1,2,3] Dongguang Yang,[1,2,3] Xiuhua Liu,[1,2,3] Xia Li,[4] and Jin Zhou[1,2,3]

[1] *Department of Hematology, The First Affiliated Hospital, Harbin Medical University, Harbin 150001, China*
[2] *Health Ministry Key Lab of Cell Transplantation, Harbin 150001, China*
[3] *Heilongjiang Institute of Hematology and Oncology, Harbin 150001, China*
[4] *College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, China*
[5] *College of Life Science, Heilongjiang University, Harbin 150080, China*

Correspondence should be addressed to Xia Li; lixia@hrbmu.edu.cn and Jin Zhou; zhoujin1111@126.com

Received 16 February 2014; Revised 5 April 2014; Accepted 6 April 2014; Published 19 May 2014

Academic Editor: Leng Han

RNA-Seq is emerging as an increasingly important tool in biological research, and it provides the most direct evidence of the relationship between the physiological state and molecular changes in cells. A large amount of RNA-Seq data across diverse experimental conditions have been generated and deposited in public databases. However, most developed approaches for coexpression analyses focus on the coexpression pattern mining of the transcriptome, thereby ignoring the magnitude of gene differences in one pattern. Furthermore, the functional relationships of genes in one pattern, and notably among patterns, were not always recognized. In this study, we developed an integrated strategy to identify differential coexpression patterns of genes and probed the functional mechanisms of the modules. Two real datasets were used to validate the method and allow comparisons with other methods. One of the datasets was selected to illustrate the flow of a typical analysis. In summary, we present an approach to robustly detect coexpression patterns in transcriptomes and to stratify patterns according to their relative differences. Furthermore, a global relationship between patterns and biological functions was constructed. In addition, a freely accessible web toolkit "coexpression pattern mining and GO functional analysis" (COGO) was developed.

1. Introduction

High-throughput RNA sequencing (RNA-Seq) is a revolutionary technology in the postgenome era. RNA-Seq rapidly generates transcript sequences and provides more detailed information than microarray-based technologies. RNA-Seq has the ability to reconstruct a complete map of the transcriptome in different cell types or physiological conditions [1, 2]. The dynamic transcriptome of cells is an important molecular signature that can represent the physiological state of different tissues, facilitating an understanding of the mechanism of gene regulation. RNA-Seq technology is becoming increasingly common as the sequencing cost is reduced and the accuracy is improved. More studies use RNA-seq technology, resulting in a series of RNA-Seq datasets across multiple related experimental conditions, such as in comparisons of multiple tumor subtypes or the effect of the concentration of a drug. Genes that exhibit similar responses to external stimuli are potentially controlled by similar regulatory mechanisms [3]. Therefore, it is important to monitor the expression pattern of genes and to discover the genes that are coexpressed among multiple conditions. These coexpression patters could describe the biological regulatory relationships of these genes.

Since the emergence of RNA-Seq technology, many differential expression (DE) analysis methods based on RNA-Seq data have been developed, such as Cuffdiff [4], DESeq [5], edgeR [6], and SAMseq [7]. These methods have been extensively used for differential expression analysis between

two conditions. Numerous genes related to specific biological functions have been found by these bioinformatics methods and confirmed by follow-up biological experiments [8, 9]. However, the DE methods described above were developed for pairwise comparisons, creating cumbersome, and confusing analyses when processing data from more than two conditions. In addition, a functional analysis was performed for only the DE genes that were isolated from the whole transcriptome, overlooking useful additional gene expression information. Because of the gene dosage effect, genes that are only slightly differently expressed may still provide useful information as a measure of functional status [10, 11]. Even the overlooked stably-expressed genes may be more essential for the survival of an organism [12].

Therefore, we developed an integrated strategy for differential coexpression pattern and GO function mining (COGO) for a RNA-Seq data series. The COGO strategy enables the biologist to view the data from a global perspective (Figure 1). First, the characteristic attributes should be extracted from the expression values of a series of RNA-Seq datasets. Second, the expression patterns can be established and stratified according to feature attributes that were extracted. Finally, functional enrichment analyses are performed for each category to determine significant function terms and the functional relationships of different GO terms that are obtained by measuring their functional semantic similarity [13]. The algorithms used in COGO are detailed in Section 2 and in Figure 1.

To illustrate a typical analysis, we applied a published RNA-Seq dataset obtained from the Gene Expression Omnibus (GEO) that contains three biological conditions [14]. The results indicated that genes coexpressed in specific categories could represent the response and stability of biological functions to the experimental conditions. In addition, a web toolkit, "COGO", was developed based on this method (http://202.97.205.74:8080/COGO). Users of this toolkit submit a profile of RNA-Seq data and receive stratified gene coexpression categories and the affected functional modules.

2. Methods

2.1. Differences in Gene Expression among Multiple Groups. Gene expression levels were quantified and normalized as FPKM/RPKM measurements. The Cufflinks package was used to calculate gene expression values using default settings [15]. Then, the average gene expression level was calculated for the experimental replicates. To identify coexpression patterns of a series of RNA-Seq libraries with $M(M \geq 3)$ experimental conditions in one study, we first quantified gene expression differences among multiple conditions. We defined $e_{i,j}$ as the expression value of gene $i = \{1, \ldots, N\}$ of condition $j = \{1, \ldots, M\}$, where N is the number of genes in the dataset. We adopted a method that was based on Shannon's Entropy (SE). SE has been used previously to identify DE genes and alternative splicing in gene expression data [16]. In this procedure, SE was introduced to measure the differences in gene expression values across experimental conditions and was defined as follows:

$$\mathrm{SE}_i = -\sum_{j=1}^{M} \frac{ae_{i,j}}{S_i} \log_2 \left(\frac{ae_{i,j}}{S_i} \right). \quad (1)$$

A tiny value α was added to the expression value $e_{i,j}$ to avoid 0 values. The new expression value was $ae_{i,j} = e_{i,j} + \alpha$, and the sum of the expression value of gene i among M experimental conditions was calculated as $S_i = \sum_{j=1}^{M} ae_{i,j}$.

2.2. Attributes Extraction according to Gene Expression Trends. SE could measure differences in variable elements, but was unable to determine the specific expression patterns within a calculation unit. Therefore, we introduced a pattern mining method based on a derivation method of polynomial curve fitting (DPCF) to describe the expression patterns of a specific gene among multiple conditions [17]. To facilitate the pattern mining of genes, the gene expression values were normalized because the polynomial fitting coefficients and fitted values are positively correlated. We defined a new dimensionless expression value, $en_{i,j} = ae_{i,j}M/S_i$, as the gene relative-expression level among multiple conditions. Then, the polynomial fitting formula was defined as $y = f_i(x)$, $x \in (1 \cdots j \cdots M)$. The derivative is a measure of how a function changes and the response of the curve trend as the inputs change. Therefore, the derivative function value of each experimental point was obtained from the following clustering attribute formula:

$$\mathrm{Der}_i = f_i'(x), \quad x \in (1 \cdots j \cdots M). \quad (2)$$

The changes in the gene expression trend between successions of conditions could be represented by Der_i. The arrangements of data should influence the discovery of the effect of expression patterns. Therefore, the order of the data must be consistent with the properties of the study, for example, sorting data according to a drug concentration gradient or tumor stages of development.

2.3. Clustering to Mine Coexpression Patterns. The determination of DE genes was obscured by the fact that a 2-fold-change may not be more meaningful than a 1.5-fold-change at the level of biological function. Therefore, we aimed to discover the expression patterns that led to different phenotypes. A hierarchical clustering method was applied, which sought to create a hierarchy of clusters in an unsupervised classifier [18]. To decide which genes should be combined in a cluster, a measure of dissimilarity in the sets of attributes was obtained. A distance matrix was constructed with $M + 1$ attributes and N genes, and then the hierarchical clusters were built by progressively merging clusters. To construct a relatively objective map of the transcriptome, the default value for the cluster number (CN) was defined as follows:

$$\mathrm{CN} = \lfloor (M + 1) \rfloor \log_{10} N. \quad (3)$$

However, we zoomed in/out of the map by changing the value of $\mathrm{CN}(1 < \mathrm{CN} < N)$ if rigorous expression patterns were

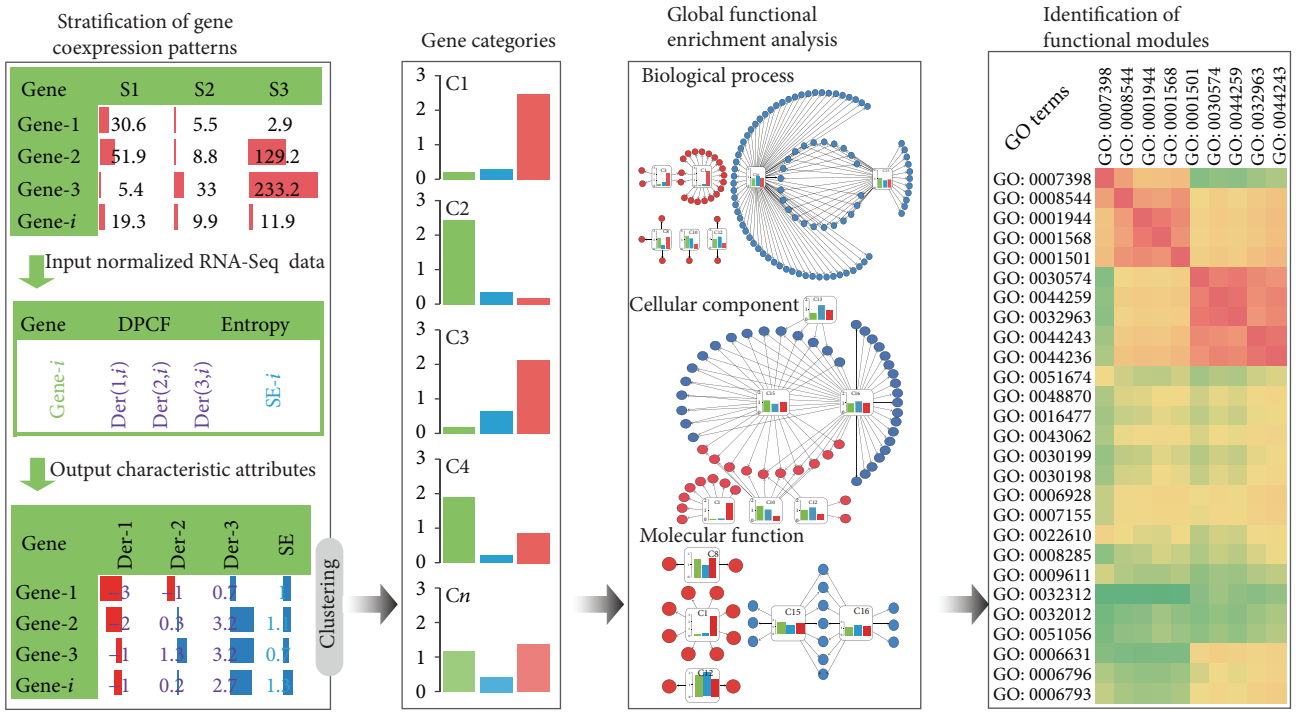


FIGURE 1: A schematic overview of COGO. A series of RNA-Seq data with three conditions was selected to illustrate the analysis process. The characteristic attributes "Der" and "SE" were extracted by a derivation method of polynomial curve fitting (DPCF) and by Shannon's Entropy (SE) models, respectively. Gene categories can then be established through clustering. A functional enrichment analysis was then performed for the categories to determine significant functions. Finally, the semantic similarity measurement was conducted to identify functional modules.

needed for detailed analysis. Then, categories of coexpression genes were obtained and represented as Cn ($1 < n < \mathrm{CN}$), and the gene number of category Cn is N_n. The gene expression patterns of categories were represented by the average of the gene relative-expression level, which is defined as $Aen_{Cn,j} = \sum_1^{N_n} en_{i,j}/N_n$, $j \in (1 \cdots M)$. Therefore, stably expressed and unstably expressed categories among multiple conditions were divided by the following criteria:

$$\frac{\max\left(Aen_{Cn,j}\right)}{\min\left(Aen_{Cn,j}\right)} \begin{cases} \leq \beta, \text{Stable expressed}, \\ > \beta, \text{Unstable expressed}, \end{cases} \quad j \in (1 \cdots M), \tag{4}$$

where β was defined as the Relative Average Expression Difference (RAED) and was set to 1.2 as default, which is more stringent than the fold-change cutoff value of "2" and can be defined by users [19].

2.4. Global Functional Enrichment Analysis. To explore the biological relationships of genes in the categories obtained by our method, a functional enrichment analysis (FEA) was introduced for the gene categories using DAVID [20]. The goal of the enrichment analysis was to determine which biological functions might be predominantly affected in the set of genes with identical expression patterns among different experimental conditions [21]. We established the Gene Ontology categories as the background knowledgebase of the FEA to acquire the functional annotating concepts for each gene category and arrive at a profile of the biological function or mechanisms. Performing a FEA on all categories was meaningful because we were able to explore the effect of external factors or physiological state on the stability of gene expression or on the biological function. To elucidate the mechanisms of regulation, a semantic similarity measurement in GO terms was conducted to identify functional modules [13].

3. Application and Results

3.1. Data Acquisition. To examine the newly developed expression pattern classifying method, published RNA-Seq data were obtained from the GEO (http://www.ncbi.nlm.nih.gov/geo/, accession number GSE33782). The data contains three RNA-Seq libraries from a colorectal cancer patient: cancer (C), paracancer (P), and distant normal tissues (N). To avoid potential biases, the datasets were filtered according to the status code provided by the Cufflinks and the FPKM value; all expression levels for a specific gene among samples were reliable (status code is OK), and the average of the gene's FPKM among samples was greater than 2. In total, 11,969 genes were detected as expressed in at least one of the samples (see Table S1).

3.2. Coexpression Pattern Mining. The characteristic attributes were computed and genes were clustered into 16 categories using the defined formula (see Section 2).

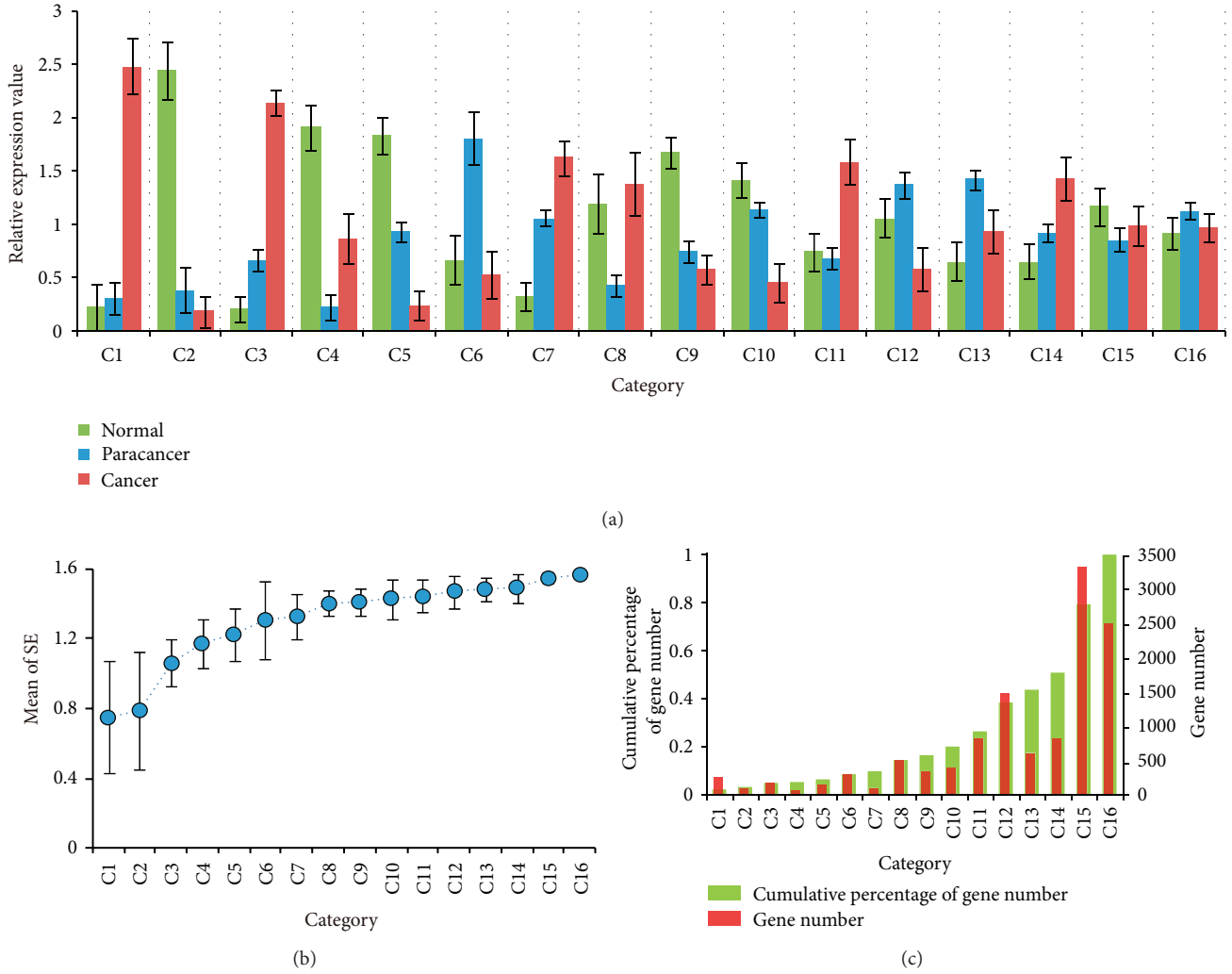


FIGURE 2: Gene expression pattern classification results of the colon RNA-Seq dataset. (a) A chart showing gene expression patterns among different tissues for each cluster category. The y-axis is dimensionless and represents the mean gene relative expression level; error bars show the standard deviation. (b) The hollow dots represent the mean of SE for each category; error bars show the standard deviation. (c) The number of genes in each category and the cumulative percentage of the number of genes from C1 to C16.

The results showed that the genes with similar expression patterns among different types of tissues clustered into identical categories (Figure S1(a)). For example, transcripts in C1 were absent in or at a very low level in normal tissues and paracancer tissues; however, these transcripts were expressed at relatively high levels in cancer tissues (Figure 2(a)). Similarly, genes in C2 were expressed at low levels in paracancer tissues and cancer tissues, but were expressed at high levels in normal tissues. In general, the gene expression differences among the three types of tissues gradually reduced from C1 to C16 (Figure 2(a)). We conducted a statistical analysis of the number of gene and average entropy of each category and then calculated the category frequency over the total number of genes. The mean of the SE of the categories gradually increased, which represents a decrease in the expression difference treads from C1 to C16 (Figure 2(b)). The majority of the genes were gathered in higher-numbered categories, which was in agreement with real biological situations (Figure 2(c)) [22]. The gene expression differences of the categories were determined using a stringent default value. The results showed that the top 14 categories accounted for 51.2% of the total genes and had differences in various degrees, and 48.8% of genes in the last two categories were stably expressed. These results indicated that the expression levels of most genes were relatively stable among different physiological states; this finding is consistent with the assumption that most genes are equivalently expressed at different conditions [22, 23] (Figure 2(c)). Furthermore, the number of significant upregulated genes exceeded the downregulated genes. The overrepresentation of upregulated gene transcripts is likely because of the metabolic exuberant state of cancer cells promoting related genes to be upregulated. Therefore, upregulated genes may be

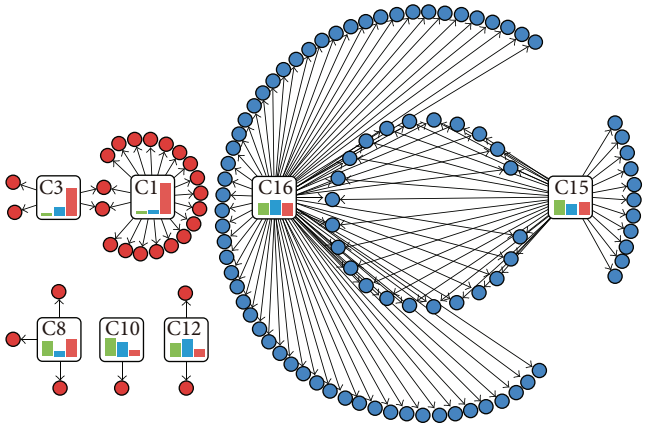


Figure 3: The functional relationship network of categories and enriched GO terms for the biological process category. The enriched GO terms of C15 and C16 are indicated by blue circles, and the other categories are indicated by red circles. The bar charts represent the expression pattern of the category. This figure was constructed to show the overall relationship of GO functions to gene patterns and gene patterns to gene patterns. More detailed GO terms are presented in Table S2 in Supplementary Material available online at http://dx.doi.org/10.1155/2014/969768.

more involved in the process of tumor formation compared to downregulated genes (Figures 2(a) and 2(c)).

3.3. Functional Enrichment Analysis. Most cancers, including colon cancer, are complex and can be caused by multiple genes and interactions. With the advance of high throughput technologies, it is now feasible to reverse engineer the underlying genetic networks that describe the interplay of molecular elements that lead to complex diseases. To explore the biological relationship of coexpressed genes obtained by our method, a FEA was performed for the gene categories using DAVID [20]. The gene ontology (GO) analysis revealed that not every category was significantly enriched for GO terms, but the number of GO terms that were significantly enriched in C15 and C16 substantially exceeded the other categories (Figure 3, Figure S2, and Figure S3). This finding suggested that the majority of the core physiological function of the cell remains stable, such as "cell death" and the "cell cycle." The FEA identified 23% of the significantly enriched terms in the biological process category to be associated with dysfunctional terms (see Table S2). However, the percent of dysfunctional terms (23%) is not proportional to the percent of differentially coexpressed genes (51.2%). This indicated that the abnormality of colon cell proliferation is because of the abnormal expression of related genes, but there were differentially expressed genes independent of experimental factors. To elucidate the mechanism of gene regulation, a functional relationship in enriched GO terms can be discovered by measuring their functional semantic similarity. Although we chose a relatively lenient cluster number by default, we still discovered enriched GO terms consistent with previous studies, such as "ectoderm development," "collagen catabolic process," and "cell migration" [24, 25]. Some functional modules were identified for specific categories, such as functions related to "development," "metabolic process," and "migration" (Figure 4). For example, the significant GO terms in the biological process category in C1 can be classified into 5 functional modules by the GO semantic similarity method and summarized by keywords; the "development" subtype, including "ectoderm development," "epidermis development," "vasculature development," "blood vessel development," and "skeletal system development," is relevant to cancer development (Figure 4) [26–28]. Therefore, causative agents of cellular state can be deduced from the subset of differentially coexpressed genes.

3.4. Comparisons of Methods and Performance Evaluations

3.4.1. The Results Obtained from the above Analysis Were Compared by a Pairwise Differential Analysis Method. Wu et al. used Cuffdiff to identify the differentially expressed genes (DEGs) of the dataset described above (GSE33782) [14]. In total, 1660, 1528, and 941 genes were extracted as significantly DE between the C-P tissue pairs, the C-N tissue pairs, and the P-N tissue pairs, respectively. Each of these groups contains upregulated and downregulated genes, thus making subsequent functional analysis more complicated. In our approach, genes were classified into 16 categories according to their expression patterns and further stratified based on differences (Figure 2(a)). Finally, the results of the FEA of the two methods were compared (Table S3). According to Wu et al., 31 GO terms in the biological process category were enriched. In total, 17 of 31 GO terms were significantly enriched in our method (FDR $\leq$ 5.0, Table S3), which were highly relevant to cancer development, such as "collagen metabolic process," "cell migration," and "ectoderm development" [29–31]. Little direct evidence was present linking the other categories to cancer; these categories included "heart development," "regulation of system process," and "muscle organ development," which were not significantly enriched in our results (FDR $>$ 5.0, Table S3). Additionally, we discovered some extra categories significantly related to cancer development, such as "blood vessel development," "collagen metabolic process," and "cell adhesion" (Figure 4) [32, 33]. The COGO method is based on specifying coexpression patterns to identify function and disease relationships. Therefore, our approach may correctly identify more biological functions than approaches based on pairwise DE methods.

3.4.2. A Comparison with the Direct Clustering Method. A comparative study was performed to evaluate with a direct clustering method using the colon cancer dataset. We first log-transformed (base 2) the gene expression values [34]. A hierarchical cluster analysis with identical settings to the method we developed was applied, and the genes were clustered into 16 categories using the default cluster number formula described above (see Section 2). Our approach displayed a better mining of the coexpression patterns of the transcriptome by reporting a smaller average variable

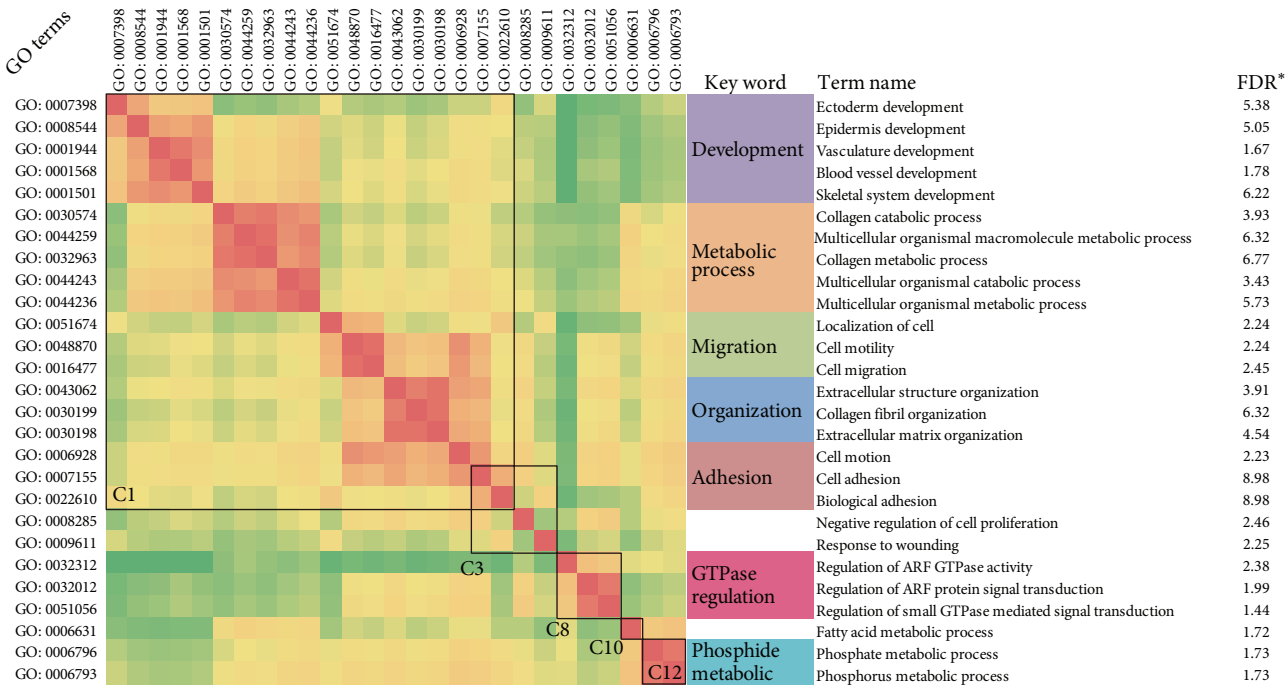


FIGURE 4: The functional similarity of the GO terms in the biological process category enriched in C1, C3, C8, C10, and C12 are displayed as a heatmap, and the similarity scores are indicated by color intensity, with red representing high similarity and green representing low similarity ($FDR^* = -\log_{10}(FDR)$).

coefficient (CV = 0.24) compared to direct cluster method (CV = 2.10) (Figure S1 and S5).

3.4.3. Comparison with STEM. Simultaneously, we compared our results with the STEM method [35]. The STEM method was developed for short time series microarray datasets and is widely used. The colon dataset was applied under standard procedures of STEM. Notably, genes were also clustered into 16 categories (SC0–SC15, Figure S4), and the average CV of the relative expression of patterns had no significant difference from our method (COGO: 0.24 versus STEM: 0.21) (Figure S5). However, our method provided clearer and more specific coexpression patterns for downstream analysis (Figure S4).

3.4.4. A Time Series Dataset. To further illustrate the performance and the application of our method, a rat pineal gland RNA-Seq dataset with 6 sampling time points was analyzed (GSE46069) [36]. In total, 8,250 genes were obtained after preprocessing, and 27 coexpression patterns were identified by COGO using default settings. A comparative study was provided to compare our method to the direct clustering method. The chart in Figure S6 shows that our results described the data better than the direct clustering method (COGO: CV = 0.27 versus direct clustering: CV = 1.92). One category containing the timekeeping AANAT gene was mainly enriched in the two-function model (Figure S7). One of the functions was related to "cytokine response," including "response to hormone stimulus" and "response to inorganic substance," and the other function was related to "neuron function," including "neuron development" and "axonogenesis." Both of these functions are associated with the circadian clock, and the findings are consistent with previous studies [37–39].

4. Discussion and Conclusions

The transcriptome reveals the status and functional mechanism of the cell as the cell responds to external stimuli. In the presence of various confounders, such as the technical deviation between runs and biological variability, one of the challenges in RNA-Seq data analysis is to extract real biological responses from substantial amounts of transcriptomic expression data. Most of the RNA-Seq data analysis methods have been developed to determine the lists of genes with significant differential expression [40]. In addition, evidence has shown that genes with similar expression patterns are likely to be regulated through similar mechanisms [3]. Alterations in the biological function can be detected by identifying gene expression patterns among a series of RNA-Seq data.

In general, analyses of the transcriptome should be performed on three levels: probe the tendency of macroscopic expression changes, such as in a functional enrichment analysis; analyze captured genes with fluctuations among conditions; and state information based hypotheses and confirm with biological experiments or literature. This research design is a continuously exploring process that cyclically considers the entire dataset to individual members. In this study, all of the detectable genes are stratified into categories according to their expression pattern. A GO enrichment analysis

was then performed on each category. We downplayed the importance of DE genes and rediscovered significant gene sets at the integral level. Therefore, the map reflecting biological functional changes is objectively structured on total detectable genes. Genes with different expression patterns exhibit different functional orientations. Therefore, GO terms enriched from categories with large gene-expression differences among conditions may reflect biological dysfunction, and GO terms enriched from categories with little gene-expression differences among conditions may also provide important biological information and may be important for cell survival. Therefore, all of the enriched functional results promote a comprehensive understanding of the molecular mechanisms involved in a specific biological process or disease.

Furthermore, not every category displayed enriched GO terms. Confounding genes may display similar expression patterns and lead to an indeterminate functional orientation or a strong relationship between genes and experimental factors is absent. Our research strategy removes distractions to focus on the notable genes and biological functions. However, meaningful genes can be retrieved through an analysis of significant biological functions or pathways, even in the presence of the unannotated genes.

In this study, we provide an integrated global strategy for coexpression pattern stratification and GO functional analysis for a RNA-Seq data series. We globally clustered genes in RNA-Seq data according to their expression patterns and gene expression differences. The results showed that genes with similar expression patterns clustered into categories in multiple characteristic attribute strategies. This creates opportunities for integrated genomic analyses of unprecedented scope and scale. Global functional analyses can be conducted, and the resulting functional modules provide a diverse repertoire of biological states of different cell types that cannot be captured by analyzing differentially expressed genes alone. Additionally, genes of a specific function can be clustered into categories to explore the expression patterns and regulatory relationships of the functional unit, providing insights into the response of functional mechanisms.

We believe that our method provides a new perspective that downplays the importance of DE genes and rediscovers significant gene sets at an integral level. We provide more useful scenarios for biologists to further explore mechanisms of biological functions and gene regulation.

Conflict of Interests

The authors declare that there is no conflict of interests.

Authors' Contribution

Hui Zhao and Fenglin Cao contributed equally to this work.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant no. 81070439) and the National High-tech Research and Development Program of China (Grant no. SS2012AA020203).

References

[1] V. Costa, M. Aprile, R. Esposito, and A. Ciccodicola, "RNA-Seq and human complex diseases: recent accomplishments and future perspectives," *European Journal of Human Genetics*, vol. 21, no. 2, pp. 134–142, 2013.

[2] S. Oh, S. Song, G. Grabowski, H. Zhao, and J. P. Noonan, "Time series expression analyses using RNA-seq: a statistical approach," *BioMed Research International*, vol. 2013, Article ID 203681, 16 pages, 2013.

[3] J. L. DeRisi, V. R. Iyer, and P. O. Brown, "Exploring the metabolic and genetic control of gene expression on a genomic scale," *Science*, vol. 278, no. 5338, pp. 680–686, 1997.

[4] 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.

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

[6] 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.

[7] 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, 2013.

[8] 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.

[9] S. Marguerat and J. Bähler, "RNA-seq: from technology to biology," *Cellular and Molecular Life Sciences*, vol. 67, no. 4, pp. 569–579, 2010.

[10] A. S. Nord, W. Roeb, D. E. Dickel et al., "Reduced transcript expression of genes affected by inherited and de novo CNVs in autism," *European Journal of Human Genetics*, vol. 19, no. 6, pp. 727–731, 2011.

[11] G. Klein, "The role of gene dosage and genetic transpositions in carcinogenesis," *Nature*, vol. 294, no. 5839, pp. 313–318, 1981.

[12] M. Juhas, L. Eberl, and J. I. Glass, "Essence of life: essential genes of minimal genomes," *Trends in Cell Biology*, vol. 21, no. 10, pp. 562–568, 2011.

[13] J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu, and C.-F. Chen, "A new method to measure the semantic similarity of GO terms," *Bioinformatics*, vol. 23, no. 10, pp. 1274–1281, 2007.

[14] Y. Wu, X. Wang, F. Wu et al., "Transcriptome profiling of the cancer, adjacent non-tumor and distant normal tissues from a colorectal cancer patient by deep sequencing," *PLoS ONE*, vol. 7, no. 8, Article ID e41001, 2012.

[15] 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, no. 3, pp. 562–578, 2012.

[16] P. F. F. de Arruda, M. Gatti, F. N. F. Junior et al., "Quantification of fractal dimension and Shannon's entropy in histological diagnosis of prostate cancer," *BMC Clinical Pathology*, vol. 13, no. 1, 6 pages, 2013.

[17] S. L. Arlinghaus, *Practical Handbook of Curve Fitting*, CRC Press, Boca Raton, Fla, USA, 1994.

[18] R. Xu, D. C. Wunsch, and IEEE Computational Intelligence Society, *Clustering*, IEEE Press Series on Computational Intelligence, Hoboken, NJ, USA, 2009.

[19] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 9, pp. 5116–5121, 2001.

[20] 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.

[21] 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.

[22] M. A. Dillies, A. Rau, J. Aubert et al., "A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis," *Brief Bioinformatics*, vol. 14, no. 6, pp. 671–683, 2013.

[23] S. Jiao and S. Zhang, "Estimating the proportion of equivalently expressed genes in microarray data based on transformed test statistics," *Journal of Computational Biology*, vol. 17, no. 2, pp. 177–187, 2010.

[24] J. P. Thiery, "Epithelial-mesenchymal transitions in development and pathologies," *Current Opinion in Cell Biology*, vol. 15, no. 6, pp. 740–746, 2003.

[25] A. Klein, C. Olendrowitz, R. Schmutzler et al., "Identification of brain- and bone-specific breast cancer metastasis genes," *Cancer Letters*, vol. 276, no. 2, pp. 212–220, 2009.

[26] N. Bessodes, "Reciprocal signaling between the ectoderm and a mesendodermal left-right organizer directs left-right determination in the sea urchin embryo," *PLoS Genetics*, vol. 8, no. 12, Article ID e1003121, 2012.

[27] D. Liu and P. J. Hornsby, "Fibroblast stimulation of blood vessel development and cancer cell invasion in a subrenal capsule xenograft model: stress-induced premature senescence does not increase effect," *Neoplasia*, vol. 9, no. 5, pp. 418–426, 2007.

[28] J. Dutkowski and T. Ideker, "Protein networks as logic functions in development and cancer," *PLoS Computational Biology*, vol. 7, no. 9, Article ID e1002180, 2011.

[29] M. K. Bode, T. J. Karttunen, J. Makela, L. Risteli, and J. Risteli, "Type I and III collagens in human colon cancer and diverticulosis," *Scandinavian Journal of Gastroenterology*, vol. 35, no. 7, pp. 747–752, 2000.

[30] M. E. Minard, L. M. Ellis, and G. E. Gallick, "Tiam1 regulates cell adhesion, migration and apoptosis in colon tumor cells," *Clinical and Experimental Metastasis*, vol. 23, no. 5-6, pp. 301–313, 2006.

[31] C. D. House, C. J. Vaske, A. M. Schwartz et al., "Voltage-gated Na+ channel SCN5A is a key regulator of a gene transcriptional network that controls colon cancer invasion," *Cancer Research*, vol. 70, no. 17, pp. 6957–6967, 2010.

[32] J. Haier and G. L. Nicolson, "The role of tumor cell adhesion as an important factor in formation of distant colorectal metastasis," *Diseases of the Colon and Rectum*, vol. 44, no. 6, pp. 876–884, 2001.

[33] S. Patan, S. Tanda, S. Roberge, R. C. Jones, R. K. Jain, and L. L. Munn, "Vascular morphogenesis and remodeling in a human tumor xenograft: blood vessel formation and growth after ovariectomy and tumor implantation," *Circulation Research*, vol. 89, no. 8, pp. 732–739, 2001.

[34] M. D. Robinson and A. Oshlack, "A scaling normalization method for differential expression analysis of RNA-seq data," *Genome Biology*, vol. 11, no. 3, article r25, 2010.

[35] J. Ernst and Z. Bar-Joseph, "STEM: a tool for the analysis of short time series gene expression data," *BMC Bioinformatics*, vol. 7, article 191, 2006.

[36] J. Falcon, S. L. Coon, L. Besseau et al., "Drastic neofunctionalization associated with evolution of the timezyme AANAT 500 Mya," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 111, no. 1, pp. 314–319, 2014.

[37] B. Blömeke, K. Golka, B. Griefahn, and H. C. Roemer, "Arylalkylamine N-acetyltransferase (AANAT) genotype as a personal trait in melatonin synthesis," *Journal of Toxicology and Environmental Health—Part A: Current Issues*, vol. 71, no. 13-14, pp. 874–876, 2008.

[38] C. Sandu, D. Hicks, and M.-P. Felder-Schmittbuhl, "Rat photoreceptor circadian oscillator strongly relies on lighting conditions," *European Journal of Neuroscience*, vol. 34, no. 3, pp. 507–516, 2011.

[39] M. Seth and S. K. Maitra, "Importance of light in temporal organization of photoreceptor proteins and melatonin-producing system in the pineal of carp catla catla," *Chronobiology International*, vol. 27, no. 3, pp. 463–486, 2010.

[40] J. H. Kim, "Chapter 8: biological knowledge assembly and interpretation," *PLoS Computational Biology*, vol. 8, no. 12, Article ID e1002858, 2012.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 730814, 7 pages
http://dx.doi.org/10.1155/2014/730814

Research Article

BLAT-Based Comparative Analysis for Transposable Elements: BLATCAT

Sangbum Lee,[1] Sumin Oh,[2] Keunsoo Kang,[3] and Kyudong Han[2,4]

[1] *Department of Computer Science, Dankook University, Cheonan 330-714, Republic of Korea*
[2] *Department of Nanobiomedical Science and BK21 PLUS NBM Global Research Center for Regenerative Medicine, Dankook University, Cheonan, 330-714, Republic of Korea*
[3] *Department of Microbiology, Dankook University, Cheonan 330-714, Republic of Korea*
[4] *DKU-Theragen Institute for NGS Analysis (DTiNa), Cheonan 330-714, Republic of Korea*

Correspondence should be addressed to Kyudong Han; kyudong.han@gmail.com

Received 2 April 2014; Accepted 28 April 2014; Published 18 May 2014

Academic Editor: Zhixiang Lu

The availability of several whole genome sequences makes comparative analyses possible. In primate genomes, the priority of transposable elements (TEs) is significantly increased because they account for ~45% of the primate genomes, they can regulate the gene expression level, and they are associated with genomic fluidity in their host genomes. Here, we developed the BLAST-like alignment tool (BLAT) based comparative analysis for transposable elements (BLATCAT) program. The BLATCAT program can compare specific regions of six representative primate genome sequences (human, chimpanzee, gorilla, orangutan, gibbon, and rhesus macaque) on the basis of BLAT and simultaneously carry out RepeatMasker and/or Censor functions, which are widely used Windows-based web-server functions to detect TEs. All results can be stored as a HTML file for manual inspection of a specific locus. BLATCAT will be very convenient and efficient for comparative analyses of TEs in various primate genomes.

1. Introduction

The advancement of DNA sequencing technology and bioinformatics has tremendously accelerated whole genome sequencing and comparative genomic analysis. Currently, 88 genome sequences are available in the University of California, Santa Cruz (UCSC) Genome Brower website (http://www.genome.ucsc.edu/) [1]. Although the genome database is easily accessible for genome research, data analysis and interpretation still remain challenging due to the amount of sequence data and various research areas within genomics. The UCSC Genome Browser was produced in the early stage of the human genome project and provides optical effects and precise sequence alignments on query sequences [1, 2]. Users can obtain a variety of information including gene tracks, genome conservation, single nucleotide polymorphisms (SNPs), and transposable elements (TEs) from the UCSC Genome Browser [3].

In the human genome, the protein coding regions only account for about 2% of the genome, whereas TEs consist of ~50% of the primate genomes within intragenic and intergenic sequences, which are called noncoding regions [4, 5]. Most studies have focused on the protein coding regions to understand their roles in human health and disease. However, noncoding regions have been emphasized since the ENCyclopedia of DNA Elements (ENCODE) project, which aims to detect new functional sources in the human genomes [6, 7].

To screen TEs in the eukaryote genomes, RepeatMasker (http://www.repeatmasker.org) [8] and Censor (http://www.girinst.org/censor/) [9] web servers have been commonly used. These software tools provide accurate and rapid repetitive DNA annotation results; the UCSC Genome Browser is also connected with them. In the comparative genomic study between six primate whole genome sequences (human, chimpanzee, gorilla, orangutan, gibbon, and rhesus macaque) [10–14], the BLAST-like alignment tool (BLAT) [15] provides an index to find homologous regions from query sequences and allows the manual retrieved alignment of query sequences from the UCSC webpage [3]. However, these processes of

manually comparing and retrieving aligned sequences from query sequences are time consuming and difficult to use for novice users.

Here, we propose a handy Windows-based program, BLAT-based comparative analysis for transposable elements (BLATCAT; http://hanlab.dankook.ac.kr/gnu/data/file/Utility/765016963_ExyIiut9_BLATCAT.exe), which automatically and simultaneously performs BLAT, RepeatMasker, and Censor [8, 9, 15]. BLATCAT was developed to detect orthologous regions between the primate genomes. Since other nonprimate species have more genomic diversity and low-quality sequences, it is not accurate to compare with orthologous regions in other nonprimate species. Therefore, BLATCAT compares only six primate genome sequences (human, chimpanzee, gorilla, orangutan, gibbon, and rhesus macaque). These primate genomes are adequate to analyze the evolution of closely related species. The BLATCAT program can significantly reduce serial steps in comparing specific regions of six representative primate genome sequences and support both position and sequence based approach. With these features, the BLACAT program is competitive for comparative analysis of the TE in various primate species.

2. Materials and Methods

Sources. To obtain comprehensive results, the BLATCAT program utilizes the outputs of the following four popular applications.

2.1. UCSC Genome Browser. The UCSC Genome Browser is an interactive website providing useful sequenced-based tools along with a variety of genome sequence data [3]. This website offers useful browsing service for retrieving locations of DNA sequences, gene structures, and distribution of TEs in the genomes by using genomic positions or gene search terms. It currently covers genome sequences of 88 species including the human genome [1].

2.2. BLAT Search. BLAT is a pairwise DNA-sequence alignment algorithm that is widely used in comparative genomics [15]. BLAT rapidly identifies similar sequences to a query with high accuracy (>95%). The total limit of multiple query sequences is up to 75,000 letters. BLAT search results display a lot of information as follows: score (calculated according to aligned length and sequence similarity), start (position of first match on the query), end (position of last match on the query), query size (the size of input sequence), identity (sequence similarity), genomic coordinates (genomic positions of the matched sequence), and strand (orientation of the matched sequence in the genome).

2.3. RepeatMasker. RepeatMasker [8] is a TE search tool characterizing TEs in given query sequences or genomes. This program uses the Smith-Waterman-Gotoh algorithm, developed by Phil Green (unpublished data). As an input, it accepts both FASTA-formatted sequences and files.

2.4. Censor. Censor [9] is also a web-based tool that scans DNA sequences for TEs against a reference dataset of TEs and delivers an abridged annotation of TEs. The major classes of TEs annotated by Censor are 40 subfamilies of DNA transposon and LTR and non-LTR retrotransposons including retroviruses and simple repeats. Censor is also available to screen TEs in other species besides human TEs [16]. It uses the same algorithm with RepeatMasker and supports FASTA, GenBank, and EMBL formats for query sequence.

Table 1: List of developmental libraries implemented in BLATCAT.

Development tool	Eclipse Indigo version Java EE IDE
Development language	Java (JDK 1.6)
Used library	Jsoup, Windowbuilder, and Jsmooth

2.5. Development Environment. BLATCAT was developed in the environment as described below (see also Table 1). Since it was implemented in Java (it requires Java Virtual Machine version 1.6 or above) [17], the current executable version of BLATCAT only supports Windows. BLATCAT is implemented with three open libraries called Jsoup, Windowbuilder, and Jsmooth. Briefly, Jsoup (http://jsoup.org) is responsible for interacting with the UCSC genome browser. Windowbuilder (https://www.eclipse.org/windowbuilder) is used to design user interface. An executable version of the BLATCAT program was packed with Jsmooth (http://jsmooth.sourceforge.net).

3. Results and Discussion

3.1. BLATCAT Workflow. BLATCAT accepts two types of input: genomic position or DNA sequence (Figures 1 and 2). Users can choose species and different versions of genome assembly for analysis (Figure 2(d)). In addition, the users can extend range of searching regions up to three times by adjusting "DNA option" placed at the bottom (Figure 2(e)). When the user selects the "position" tab for a query with options (Figure 2(a)), BLATCAT first extracts DNA sequences of the given positions (Figure 2(b)) and searches selected genomes for homologous sequences via the UCSC Genome Browser [1]. On the other hand, if the user provides genome sequences instead of the genome positions without any options on the "sequence" tab (Figure 3), the program directly performs pairwise sequence alignment using the BLAT algorithm [15]. Only the most similar sequence is selected and used as a query for searching homologous sequences. Once the homologous sequences are extracted, repetitive DNA sequences in all homologous sequences are identified using RepeatMasker as default [8]. Subsequently, Censor marks TEs in the homologous sequences for visualization [9].

3.2. BLATCAT Output. The BLATCAT output provides the following useful information for researchers. It shows the homologous sequence and its genomic coordinate in each species (Figure 4). BLATCAT maintains color of strings or formats acquired from other programs, such as the UCSC genome browser, BLAT (Figure 4), RepeatMasker (Table 4),

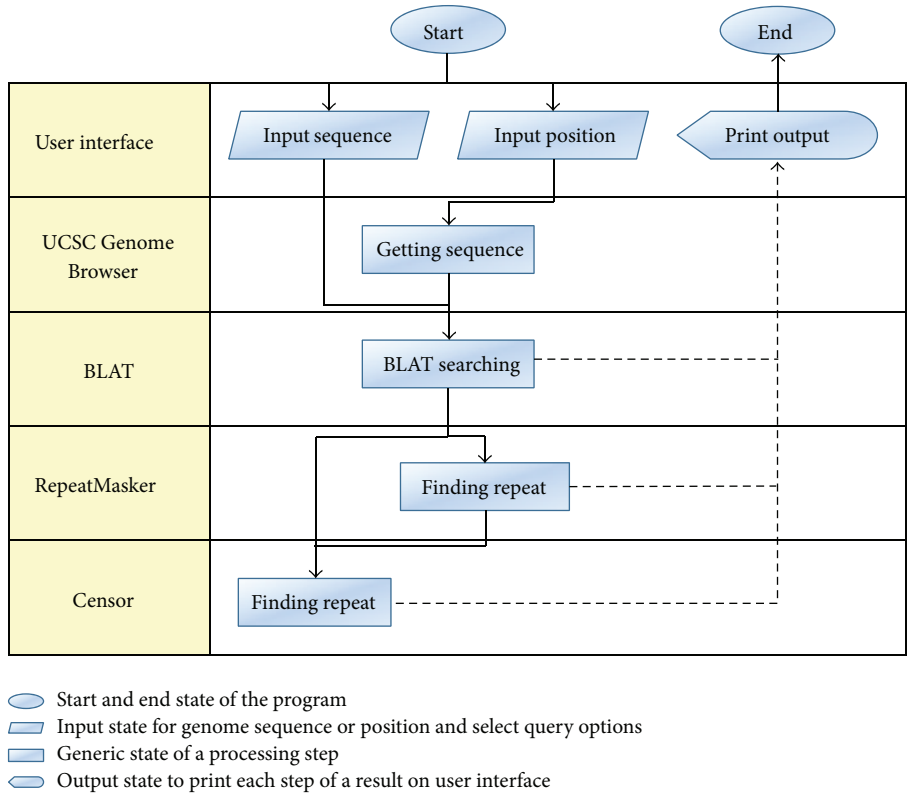


Figure 1: BLATCAT flowchart. BLATCAT runs several programs sequentially and utilizes outputs of the programs. The arrows indicate the flow of the BLATCAT algorithm.

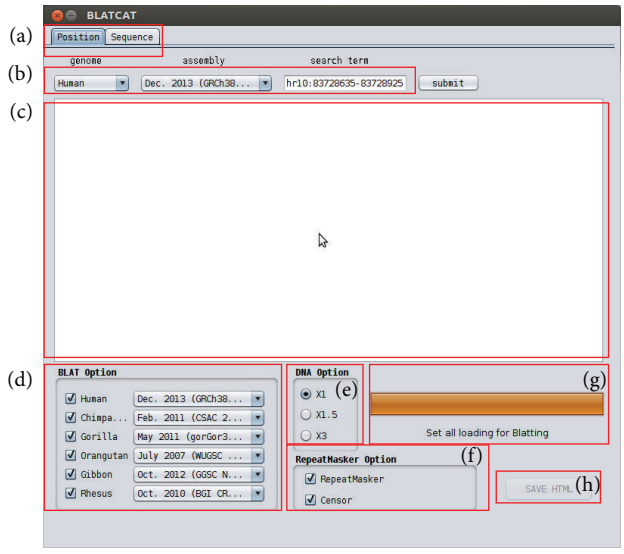


Figure 2: BLATCAT user interface for genomic position. (a) Two types of input tabs are shown. (b) Genome and its assembly version can be changed. Users can put position information in the search term field. (c) Result appears in this field. (d) Selectable species and their genome assembly are shown. (e) The length of a given input sequence can be extended up to three times (×3). Selectable RepeatMasker options (f) and a progress bar (g) are shown. (h) The output can be saved as a HTML file.

and Censor (Table 5) [1, 8, 9, 15]. These results are merged and displayed at the same time upon submission (Figure 5). Excluding the user interface, all results of previous steps can be stored as a HTML file (Figure 2(h)) if the user clicks the "save HTML" button (Figure 5). Descriptions of attributes of RepeatMasker and Censor can be found in Tables 2 and 3 [8, 9]. The user can easily "copy and paste" any part of the output to other software applications.

3.3. Comparison of BLATCAT with the UCSC-BLAT-RepeatMasker-Censor Procedure. Previous studies [18–25] that examined species-specific insertions/deletions mediated by TEs should inspect orthologous primate sequences at each locus using manual methods (UCSC, BLAT, and RepeatMasker/Censor). BLATCAT is a user-friendly program optimized for identifying TEs in homologous sequences of six primate species. The one-step procedure of BLATCAT allows researchers to perform comparative identification of TEs. To obtain TEs in homologous sequences of six species manually, users have to go through several steps. First, the users have to extract DNA sequence of interest from genome browsers, such as UCSC and Ensembl genome (see Figure S1 in Supplementary Material available online at http://dx.doi.org/10.1155/2014/730814) [1, 26]. Then, homologous sequences are identified by aligning the extracted sequence to the genome of interest by using BLAT or similar programs (Figure S2).

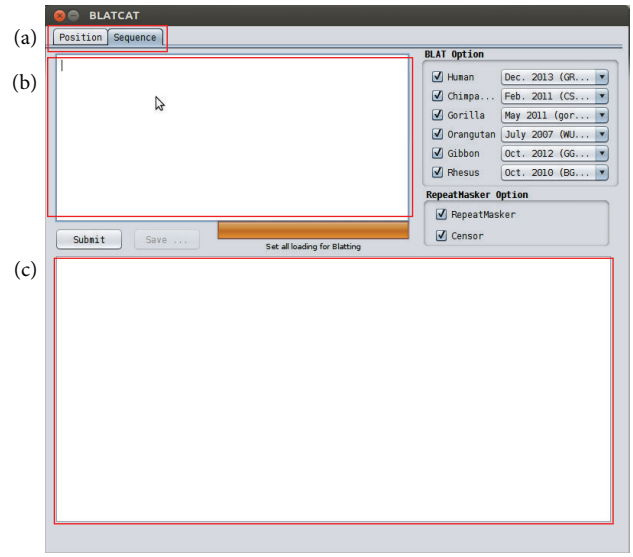


Figure 3: BLATCAT user interface for DNA sequence. (a) DNA sequence can be used as an input for analysis. (b) DNA sequence should be placed in the empty field. (c) Result appears in this empty field.

>Human chr10: 83728636-83728925
aatctgctctaaaaaaaaaggtctgtttttaaaattatcaggttgagatattgtatttttaaaccacacatttcaatattggcatctattgcctacttcTGCTCCATAATATGTGAGAAAA...

>Chimpanzee chr10: 83107742-83108031
aatctgctcttaaaaaaaaggtctgtttttaaaattatcaggttgagatattgtatttttaaaccacacatttcaatattggcatctattgcctacttcTGCTCCATAATATGTGAGAAAAT...

>Gorilla chr10: 96758224-96758634
catcagttttaacaatgtaccgtctggtgggggatgtcaatagtgaggaaggttatgcatatgtggggctgaggagcatattggaactttctgtacttttaTGCTCaatttttctgtaagtct...

>Orangutan chr10: 51231208-51231497
aatctgctctaaaaaaaaaggtctgtttttaaaattatcaggttgagatattgtatttttaaaccacacatttcaatattggcatctattgcctacttcTGCTCCATAATATGTGAGAAAAT...

>Gibbon chr18: 40838618-40838906
aaatctgctctaagaaaaaggtctgtttttaaaattatcaggttgagatattgtatttttaaaccacacatttcaatattggcatctattgcctatttcTGCTCCATAATATGTGAGAAAAT...

>Rhesus chr9: 51426848-51427124
ctagaaaaaataggtctgtttttaaaaattatcaggttgagatatcgtacttttaaaaccacacatttcaatattggcatctattgtctatttctattctATATGTGAGAAAATTGaCATTTC...

Figure 4: The result of BLAT searching within BLATCAT. Homologous sequence of each species is displayed as FASTA format. Genomic position (red) and repeat sequence (blue) are marked with different colors.

Table 2: Description of the RepeatMasker attributes.

Attribute	Description
SW score	Smith-Waterman score of the match, usually complexity adjusted
Perc div.	Percentage of substitutions in matching region compared to the consensus
Perc del.	Percentage of bases opposite a gap in the query sequence (deleted bp)
Perc ins.	Percentage of bases opposite a gap in the repeat sequence (inserted bp)
Query sequence	Name of query sequence
Position in query	
Begin	Starting position of match in query sequence
End	Ending position of match in query sequence
(Left)	Number of bases in query sequence past the ending position of match
Matching repeat	Match is with the complement of the consensus sequence in the database
Repeat class/family	Name of the matching interspersed repeat
Position in repeat	
Begin	The class of the repeat
End	Number of bases in (complement of) the repeat consensus sequence prior to beginning of the match
(Left)	Starting position of match in database sequence (using top-strand numbering)
ID	Ending position of match in database sequence

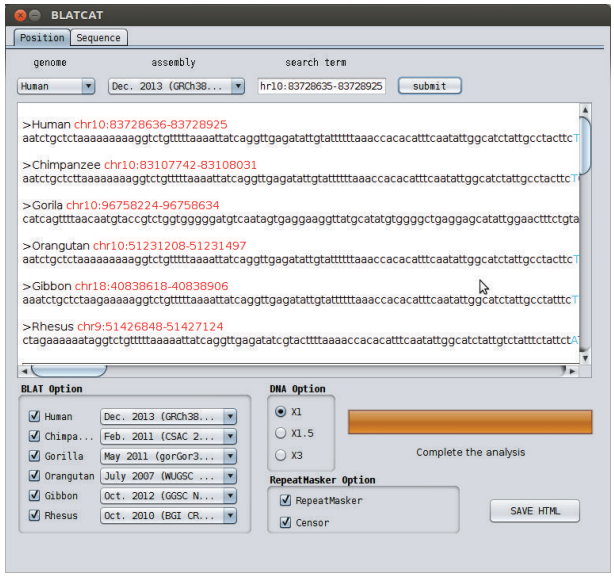


Figure 5: Screenshot of the BLATCAT output. All the results (Figure 4 and Tables 4 and 5, results of BLAT, RepeatMaster, and Censor) are merged and displayed in the user interface at the same time. Other contexts are identical to Figure 2.

Table 3: Description of the Censor attributes.

Attribute	Description
Name	Column Name contains locus names of submitted query sequences (first column) and Repbase library sequences (fourth column). Repbase names are hyperlinked to their sequences.
From/To	Column From/To contains beginning/ending of positions of fragment on corresponding sequence.
Class	This is class/subclass of repeat as specified in repeat annotation.
Dir	Values in column Dir indicate orientation ("d" for direct and "c" for complementary) of repeat fragment—columns 4–6.
Sim	Column Sim contains value of similarity between 2 aligned fragments.
Pos	Column Pos is roughly the ratio of positives to alignment length.
Mn : Ts	Column Mm : Ts is a ratio of mismatches to transitions in nucleotide alignment. The closer this number is to 1 the more likely is that mutations are evolutionary.
Score	This column contains the alignment score obtained from blast.

To identify TEs in these sequences, the users have to run RepeatMasker and/or Censor with each homologous sequence as a query repeatedly (Figures S3 and S4) [8, 9]. These sequential analyses require certain knowledge of algorithms and are time-consuming tasks. Our application explicitly shortens the steps for comparative TE analysis and is easy to use.

To estimate the efficiency of BLATCAT, we compared manual method and BLATCAT in the human position as a query (chr18: 40,208,090–40,208,390). The result indicates that BLATCAT (processing time: 65 sec) works five times faster than that of the manual method (processing time: 356 sec).

3.4. The Weaknesses of BLATCAT. Although BLATCAT is a straightforward approach to identify TEs in homolog regions, it also has some weaknesses due to the algorithm. First, BLATCAT requires an Internet connection since it interacts with several web applications. Second, the current version of BLATCAT only runs on the Windows operating system. Third, if the size of input sequence is more than 75,000 bases, it cannot be processed due to the size limitation of the BLAT website. However, most computers are connected to the Internet these days and the typical size of input sequence should be around several kilobases. Fourth, BLATCAT only returns the top-scoring locus of homology found by BLAT, even if there is one or more homologous loci with scores nearly as high as the top hit. Therefore, BLATCAT is comparable to other genomic tools.

4. Conclusions

BLAT only finds an orthologous region between a query sequence and another single genome. However, we developed the Windows-based BLATCAT program to simultaneously compare a query sequence with its corresponding sequences from five other primates. In addition, this tool is linked to RepeatMasker and/or Censor to identify full spectrum TEs in

Table 4: The result of RepeatMasker within BLATCAT.

SW	perc	perc	perc	query	position in query				Matching repeat		Position in repeat			ID
Score	Div.	Del.	Ins.	Sequence	Begin	End	(Left)		Repeat	Class/family	Begin	End	(Left)	
510	28.2	6.4	4.5	Human	10	355	(135)	C	HAL1b	LINE/L1	(406)	2015	1664	5
475	28.7	6.4	4.2	Chimpanzee	10	355	(135)	C	HAL1b	LINE/L1	(405)	2016	1664	1
792	20.5	1.3	0.0	Gorilla	1	151	(460)	+	L1MC1	LINE/L1	6176	6328	(5)	3
402	29.3	7.3	4.8	Gorilla	133	476	(135)	C	HAL1b	LINE/L1	(406)	2015	1664	4*
478	28.6	6.7	4.8	Orangutan	10	355	(135)	C	HAL1b	LINE/L1	(406)	2015	1664	6
465	29.1	6.6	4.3	Gibbon	11	373	(116)	C	HAL1b	LINE/L1	(406)	2015	1645	2
319	32.7	6.6	2.1	Rhesus	24	342	(135)	C	HAL1b	LINE/L1	(425)	1996	1664	7

The RepeatMasker output is displayed. Descriptions of the attributes can be found in Table 1.
*indicates that there is a higher-scoring match whose domain partly (<80%) includes the domain of this match [8].

Table 5: The result of Censor within BLATCAT.

Name	From	To	Name	From	To	Class	Dir	Sim	Pos/Mm : Ts	Score
Human (SVG plot; alignments; masked)										
Human	10	368	HAL1B	610	973	NonLTR/L1	c	0.7003	2.0667	774
Chimpanzee (SVG plot; alignments; masked)										
Chimpanzee	10	368	HAL1B	610	974	NonLTR/L1	c	0.6955	2.0652	745
Chimpanzee	386	434	Gypsy-2_HMM-I	5194	5247	LTR/Gypsy	c	0.8039	1.6	209
Gorilla (SVG plot; alignments; masked)										
Gorilla	1	151	L1MC1	923	1075	NonLTR/L1	d	0.7843	1.3478	757
Gorilla	154	489	HAL1B	610	953	NonLTR/L1	c	0.6907	1.8936	674
Orangutan (SVG plot; alignments; masked)										
Orangutan	10	361	HAL1B	617	973	NonLTR/L1	c	0.7064	1.8298	761
Orangutan	386	434	Gypsy-2_HMM-I	5194	5247	LTR/Gypsy	c	0.8039	1.6	209
Gibbon (SVG plot; alignments; masked)										
Gibbon	11	367	HAL1B	610	973	NonLTR/L1	c	0.6966	1.9375	765
Gibbon	385	433	Gypsy-2_HMM-I	5194	5247	LTR/Gypsy	c	0.8039	1.6	209
Rhesus (SVG plot; alignments; masked)										
Rhesus	24	355	HAL1B	610	954	NonLTR/L1	c	0.6677	1.9231	606

The Censor output is shown. Each table shows the result of each species obtained from the Censor analysis.

the primate genomes. BLATCAT is an easy-to-use tool and is more effective than manual work. Therefore, we believe that BLATCAT is a valuable tool for a comparative analysis of TEs in primate genomes.

Conflict of Interests

The authors declare that no conflict of interests exists in this paper.

Acknowledgment

The present work was conducted with funding from the Research Fund of Dankook University in 2013.

References

[1] D. Karolchik, G. P. Barber, J. Casper et al., "The UCSC genome browser database: 2014 update," *Nucleic Acids Research*, vol. 42, pp. D764–D770, 2014.

[2] W. J. Kent, C. W. Sugnet, T. S. Furey et al., "The human genome browser at UCSC," *Genome Research*, vol. 12, no. 6, pp. 996–1006, 2002.

[3] R. M. Kuhn, D. Haussler, and W. J. Kent, "The UCSC genome browser and associated tools," *Briefings in Bioinformatics*, vol. 14, no. 2, pp. 144–161, 2013.

[4] E. S. Lander, L. M. Linton, B. Birren et al., "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, pp. 860–921, 2001.

[5] Y. J. Kim, J. Lee, and K. Han, "Transposable elements: no more 'Junk DNA'," *Genomics & Informatics*, vol. 10, no. 4, pp. 226–233, 2012.

[6] E. P. Consortium, B. E. Bernstein, E. Birney et al., "An integrated encyclopedia of DNA elements in the human genome," *Nature*, vol. 489, no. 7414, pp. 57–74, 2012.

[7] E. P. Consortium, "The ENCODE (ENCyclopedia Of DNA elements) project," *Science*, vol. 306, no. 5696, pp. 636–640, 2004.

[8] A. F. A. Smit, R. Hubley, and P. Green, "RepeatMasker Open-3.0," 1996–2010, http://www.repeatmasker.org.

[9] O. Kohany, A. J. Gentles, L. Hankus, and J. Jurka, "Annotation, submission and screening of repetitive elements in Repbase:

RepbaseSubmitter and Censor," *BMC Bioinformatics*, vol. 7, article 474, 2006.

[10] The Chimpanzee Sequencing and Analysis Consortium, "Initial sequence of the chimpanzee genome and comparison with the human genome," *Nature*, vol. 437, no. 7055, pp. 69–87, 2005.

[11] E. S. Lander, L. M. Linton, B. Birren et al., "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, pp. 860–921, 2001.

[12] D. P. Locke, L. W. Hillier, W. C. Warren et al., "Comparative and demographic analysis of orang-utan genomes," *Nature*, vol. 469, no. 7331, pp. 529–533, 2011.

[13] Rhesus Macaque Genome Sequencing and Analysis Consortium, R. A. Gibbs, J. Rogers et al., "Evolutionary and biomedical insights from the rhesus macaque genome," *Science*, vol. 316, no. 5822, pp. 222–234, 2007.

[14] A. Scally, J. Y. Dutheil, L. W. Hillier et al., "Insights into hominid evolution from the gorilla genome sequence," *Nature*, vol. 483, no. 7388, pp. 169–175, 2012.

[15] W. J. Kent, "BLAT—the BLAST-like alignment tool," *Genome Research*, vol. 12, no. 4, pp. 656–664, 2002.

[16] J. Jurka, V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, and J. Walichiewicz, "Repbase update, a database of eukaryotic repetitive elements," *Cytogenetic and Genome Research*, vol. 110, no. 1–4, pp. 462–467, 2005.

[17] J. Gosling, "Feel of Java," *Computer*, vol. 30, no. 6, pp. 53–57, 1997.

[18] A. B. Carter, A. H. Salem, D. J. Hedges et al., "Genome-wide analysis of the human Alu Yb-lineage," *Human Genomics*, vol. 1, no. 3, pp. 167–178, 2004.

[19] K. Han, J. Lee, T. J. Meyer, P. Remedios, L. Goodwin, and M. A. Batzer, "L1 recombination-associated deletions generate human genomic variation," *Proceedings of the National Academy of Sciences*, vol. 105, no. 49, pp. 19366–19371, 2008.

[20] K. Han, J. Lee, T. J. Meyer et al., "Alu recombination-mediated structural deletions in the chimpanzee genome," *PLoS Genetics*, vol. 3, no. 10, pp. 1939–1949, 2007.

[21] J. Lee, R. Cordaux, K. Han et al., "Different evolutionary fates of recently integrated human and chimpanzee LINE-1 retrotransposons," *Gene*, vol. 390, no. 1-2, pp. 18–27, 2007.

[22] J. Lee, K. Han, T. J. Meyer, H.-S. Kim, and M. A. Batzer, "Chromosomal inversions between human and chimpanzee lineages caused by retrotransposons," *PLoS ONE*, vol. 3, no. 12, Article ID e4047, 2008.

[23] A. C. Otieno, A. B. Carter, D. J. Hedges et al., "Analysis of the human Alu Ya-lineage," *Journal of Molecular Biology*, vol. 342, no. 1, pp. 109–118, 2004.

[24] S. K. Sen, K. Han, J. Wang et al., "Human genomic deletions mediated by recombination between Alu elements," *The American Journal of Human Genetics*, vol. 79, no. 1, pp. 41–53, 2006.

[25] H. Wang, J. Xing, D. Grover, D. J. Hedges, J. A. Walker, and M. A. Batzer, "SVA elements: a hominid-specific retroposon family," *Journal of Molecular Biology*, vol. 354, no. 4, pp. 994–1007, 2005.

[26] T. Hubbard, D. Barker, E. Birney et al., "The Ensembl genome database project," *Nucleic Acids Research*, vol. 30, no. 1, pp. 38–41, 2002.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 291280, 8 pages
http://dx.doi.org/10.1155/2014/291280

Research Article

A Systematic Analysis of miRNA-mRNA Paired Variations Reveals Widespread miRNA Misregulation in Breast Cancer

Lei Zhong,[1] Kuixi Zhu,[2] Nana Jin,[2] Deng Wu,[2] Jianguo Zhang,[1] Baoliang Guo,[1] Zhaoqi Yan,[1] and Qingyuan Zhang[3]

[1] *Department of General Surgery, Second Affiliated Hospital of Harbin Medical University, Harbin 150086, China*
[2] *Department of Bioinformatics, Harbin Medical University, Harbin 150081, China*
[3] *Department of Internal Medicine, Cancer Hospital Affiliated to Harbin Medical University, Harbin 150040, China*

Correspondence should be addressed to Qingyuan Zhang; zhangqy_med@163.com

Received 16 March 2014; Accepted 16 April 2014; Published 18 May 2014

Academic Editor: Zhixi Su

MicroRNAs (miRNAs) are a class of small noncoding RNAs that can regulate gene expression by binding to target mRNAs and induce translation repression or RNA degradation. There have been many studies indicating that both miRNAs and mRNAs display aberrant expression in breast cancer. Previously, most researches into the molecular mechanism of breast cancer examined miRNA expression patterns and mRNA expression patterns separately. In this study, we systematically analysed miRNA-mRNA paired variations (MMPVs), which are miRNA-mRNA pairs whose pattern of regulation can vary in association with biopathological features, such as the oestrogen receptor (ER), TP53 and human epidermal growth factor receptor 2 (HER2) genes, survival time, and breast cancer subtypes. We demonstrated that the existence of MMPVs is general and widespread but that there is a general unbalance in the distribution of MMPVs among the different biopathological features. Furthermore, based on studying MMPVs that are related to multiple biopathological features, we propose a potential crosstalk mechanism between ER and HER2.

1. Introduction

MicroRNAs (miRNAs) are a class of naturally occurring small noncoding RNAs. Mature miRNAs are 19- to 25-nucleotide-long molecules that are cleaved from 70- to 100-nucleotide hairpin pre-miRNA precursors [1, 2]. miRNAs regulate the expression of genes and play a vital role in almost every biological process, including cell differentiation, turning signalling pathways on/off, apoptosis, and cell proliferation [2, 3]. Although several models have been proposed for the mechanism underlying miRNA regulation, it is generally accepted that miRNAs regulate gene expression by binding to their target mRNAs [4, 5]. In vertebrate animals, most miRNAs bind to the 3′ untranslated region (3′UTR) of a target mRNA sequence at a partially complementary sequence and induce translation repression or mRNA degradation [6]. Interestingly, a recent study indicated that miRNAs can shift from acting as a repressor to an activator of gene translation during the cell cycle arrest period [7, 8].

Increasing numbers of microRNAs and mRNAs have been found to be related to the development of breast cancer. In contrast to previous studies based only on miRNA or mRNA expression profiles, examining both miRNA and mRNA expression profiles enables us not only to study miRNA and mRNA expression profiles separately but also to examine miRNA-mRNA regulatory pairs together [8–12]. Nevertheless, in many cancer studies based on miRNA and mRNA expression profiles, instead of considering miRNA-mRNA regulatory pairs together, the tendency is to examine either an miRNA or mRNA first and then apply strategies such as computational miRNA target gene prediction algorithms, sequence homology analysis, or expression correlation indexes to identify the corresponding counterpart of the miRNA (mRNA) and, hence, accomplish the integration of the miRNA-mRNA pair [12, 13]. Interestingly, many of these studies share the common assumption that the regulatory relationship between an miRNA and its target mRNAs is negative, and a great deal of research is therefore based on

this assumption [8–12]. For example, to identify the target mRNAs of a specific miRNA from hundreds of candidate mRNAs predicted by a computational algorithm, many scientists prefer to choose those mRNAs whose expression is significantly negatively correlated with that of the miRNA. However, this hypothesis of an miRNA negatively regulating its target mRNA conflicts with the results of a recent study showing that, in some cases, miRNAs can activate the translation of their target mRNAs [7, 8]. Moreover, the aberrant expression of miRNAs and mRNAs in breast cancer gives rise to the question of whether the regulatory pattern of miRNA-mRNA pairs varies with the development of this disease [14, 15]. Thus, we attempt to answer this question by studying the possible effects of several breast cancer-related biopathological features on the regulatory pattern of miRNA-mRNA pairs, and we consider the answer to this question to represent the cutting edge of the exploration of the molecular mechanisms of breast cancer.

Here, we propose MMPV as a term that indicates miRNA-mRNA pairs whose pattern of regulation can vary in association with different statuses of biopathological features. We reveal that the distribution of MMPVs is widespread. Moreover, we find that the miRNAs of the MMPVs that are associated with a particular biopathological feature tend to display a significant regulatory effect on the target mRNAs related to a specific status of the biopathological feature and tend to display no significant regulatory effect on the target mRNAs related to different statuses. Furthermore, based on studying MMPVs associated with multiple biopathological features, we propose the existence of a potential crosstalk mechanism between ER and HER2. Importantly, this study demonstrates that the pattern of miRNA-mRNA regulation can be altered in the context of different statuses of biopathological features, and this discovery will benefit further research exploring the molecular mechanisms underlying breast cancer.

2. Materials and Methods

2.1. miRNA and mRNA Expression Data. Both miRNA and mRNA expression data were obtained from PMID: 21364938 [16]. The data were derived from the expression profiling of 799 miRNAs and 30,981 mRNAs in 101 primary human breast tumours. Five biopathological features of each sample were available. We classified each biopathological feature as showing one of two different statuses: oestrogen receptor positive (ER+)/oestrogen receptor negative (ER−); mutant TP53 (TP53+)/wild type TP53 (TP53−); survival greater than five years (survial5+)/survival less than five years (survival5−); HER2 positive (HER+)/HER2 negative (HER2−); and basal-like breast cancer (basal)/no basal-like breast cancer (non-basal). The miRNA and mRNA expression data have been submitted to the Gene Expression Omnibus (GEO) under accession numbers GSE19536 and GSE19783, respectively.

2.2. miRNA-mRNA Targeting Pairs. We obtained experimentally validated miRNA-mRNA targeting pairs from Tarbase 6.0 [17]. Among the healthy population, the regulatory pattern of 293 miRNA-mRNA pairs indicated positive regulation, while that of 3,628 miRNA-mRNA pairs showed negative regulation.

2.3. Computation of the miRNA-mRNA Regulatory Patterns. To examine the regulatory pattern of miRNA-mRNA pairs, which could vary with different statuses of biopathological features, we must quantify the regulatory patterns of the miRNA-mRNA pairs associated with a certain status of a biopathological pattern. For illustrating, here we calculate the regulatory pattern of each miRNA-mRNA pair in ER+ and ER− specimens first. Gene expression with samples in both ER+ and ER− was compiled first, and then the Pearson correlation coefficient (PCC) was adapted to measure the regulatory pattern of miRNA-mRNA pairs associated with a specific status of a biopathological feature. If the PCC is greater than zero, then a positive regulatory pattern corresponds to this PCC and vice versa.

2.4. Choosing the miRNA-mRNA Pairs Whose Regulatory Pattern Varies Significantly with Each Biopathological Status to the Normal Condition. Each miRNA-mRNA pair receives 2 PCCs, corresponding to ER+ and ER− statuses, representing its regulatory pattern in ER+ and ER− specimens, respectively. Based on the quantified results regarding the regulatory pattern of each miRNA-mRNA pair, we prefer those miRNA-mRNA pairs whose 2 PCCs showed opposite algebraic signs (sign change pairs) and those whose 2 PCCs showed the same algebraic sign but displayed the ratio of the PCC of each condition to the normal greater than 2 or smaller than 0.5 (fold change pairs). So that our results would have greater biological importance, we later removed miRNA-mRNA pairs whose 2 PCCs were both insignificant (B-H FDR $q < 0.05$).

2.5. Representing the Regulatory Patterns of miRNA-mRNA Pairs. Following the two steps described above, we obtained the miRNA-mRNA pairs whose regulatory patterns varied significantly with an ER+ versus ER− status (ER MMPVs). We used the letters U and D to represent positive and negative regulations and $*$ to indicate the significance in the statistic. Thus, given that the regulatory pattern of experimentally validated miRNA-mRNA pairs downloaded from Tarbase 6.0 in the healthy population was known, we used U or D to represent the regulatory pattern of miRNA-mRNA pairs in breast cancer patients with an ER+ or ER− status and in the healthy population. For example, if the regulatory pattern of hsa-miR-1 and CA3 corresponds to negative regulation in the healthy population and to positive regulation in ER+ breast cancer patients and significant positive regulation in ER− breast cancer patients, then we can refer to this pair as D_U_U*. As another example, the regulation of has-mir-375 and FOLR1 is negative in the healthy population, while in ER+ specimens it is positive, whereas it is significantly negative in ER− specimens. Hence, the change in the pattern of regulation can be represented as D_U_D* to indicate the regulatory pattern of hsa-miR-1 and CA3 in the healthy

population and in ER+ breast cancer patients and ER− breast cancer patients.

2.6. Gene Ontology (GO) Enrichment Analysis of MMPVs. GO database was used to explore the biological function involved in MMPVs. We used Gorilla [18] to conduct GO enrichment analysis, and the P value threshold is set as $1.0E-03$. The background list comprised all of the genes for the miRNA-mRNA pairs that we obtained from Tarbase 6.0. We placed the mRNAs of each type of MMPV into a target set and obtained the results for the biological process cellular component.

3. Results and Discussion

Our method examines miRNA and mRNA gene expression data to obtain MMVPs for five breast cancer-related biopathological features. The statistical summary is shown in Table 1. The definitions of the fold change pairs and sign change pairs were given above.

First, we discuss the general unbalanced enrichment trend in the distribution of MMPVs associated with every type of biopathological feature. Second, we select (B-H FDR $q < 0.05$) MMPVs whose miRNA and mRNA are both significantly differentially expressed between the two statuses of the biopathological features (DE-MMPV), and we check the published literature to confirm their relationship with the corresponding biopathological feature. Third, we analyse MMPVs that are shared by multiple types of biopathological features, and we propose the existence of potential crosstalk between ER and HER2. Fourth, mRNAs of each type of MMPV are analysed for GO enrichment through a hypergeometric gene set enrichment analysis. Finally, we map the mRNAs of the MMPVs to Human Protein Reference Database (HPRD) protein-protein interaction networks to explore the topological features of genes of MMPVs.

Table 1: Statistical results for the MMPV responses to different features.

Feature	Number of fold change pairs	Number of sign change pairs	Total number of MMPVs
TP53	110	49	159
ER	72	30	102
Her2	516	768	1,284
Survival time	0	0	0

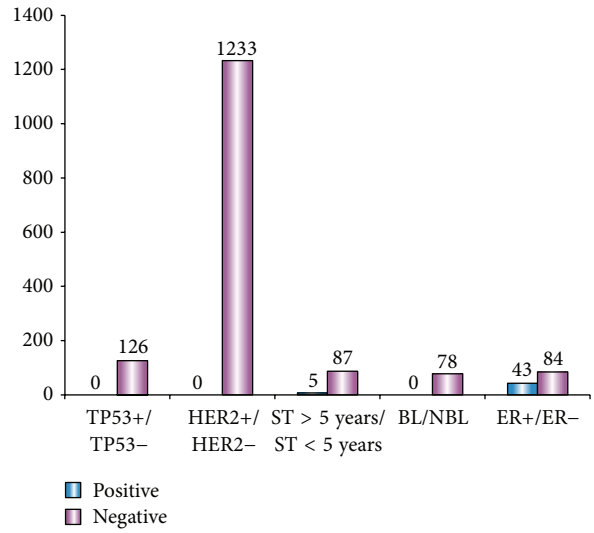


Figure 1: The distribution of MMPVs among the five types of biopathological features. The height of positive bar (in blue colour) represents the number of MMPVs whose regulatory pattern is significant for the first status of a specific biopathological feature, and the height of the negative bar (in red colour) represents the number of MPPVs whose regulatory pattern is significant for the second status of a specific biopathological feature.

3.1. The General Unbalanced Distribution of MMPVs. Because the number of miRNA-mRNA pairs whose regulatory pattern is positive in the healthy population is relatively small and, hence, cannot reach statistical significance, in this section, we examine only miRNA-mRNA pairs whose regulatory pattern is negative in the healthy population. Their distribution among the five examined types of biopathological features is shown in Figure 1.

For each of the biopathological features, we found that the regulatory pattern of MMPVs tended to be significant for one specific status and to be insignificant for the other status. To be more specific, the miRNAs of the MMPVs tended to have a significant regulatory effect on their target mRNAs for the following statuses: wild type TP53 gene, HER2−, survival time of less than five years, nonbasal-like breast cancer subtype, and ER−.

The above result showed that the overexpression of HER2 is the result of deregulation of genes, rather than gene amplification, and this discovery is consistent with the result of Menard et al. [19]. Considering this result together with our findings, we propose that the widespread decreasing regulatory effect of miRNAs on their target mRNAs contributes to HER2 expression.

Cheng et al. adopted a regulatory effect score (RE score) to evaluate the regulatory effect of miRNAs and discovered that, compared with ER+ patients, most miRNAs exhibit a higher RE score. In other words, they have a more significant regulatory effect on their target mRNAs in ER− patients [20]. This discovery is consistent with our findings.

Suzuki et al. found that three types of missense mutation in the DNA-binding domain of p53 can lead to decreased processing of pri-miRNAs by Drosha [21]. They therefore proposed that p53 mutants might reduce the interaction between pri-miRNAs and Drosha complex proteins and, hence, affect the genesis of mature miRNAs. In this context, mutation of the TP53 gene could decrease the production and activity of miRNAs and ultimately lead to the results that we obtained here: MMPVs related to TP53 tend to exhibit a significant regulatory pattern in a population with the wild TP53 gene.

Most miRNAs can be regarded as antioncogenesis miRNAs given the fact that compared to the healthy population, most miRNAs are downregulated in cancer patients [1]. Moreover, many genes encoding miRNAs are located in regions that are related to cancer, and genes in these regions frequently undergo rearrangement, amplification, and loss [22]. Specifically, genomes associated with basal-like breast cancer tend to be more unstable than those associated with other subtypes of breast cancer [23]. In addition, Blenkiron et al. compared the expression of genes that are involved in the genesis of miRNAs in several breast cancer subtypes and found that Dicer1 was significantly downregulated in basal-like, HER2+, and luminal B cases, all of which are closely associated with poorer prognostic results [24]. Given that basal-like breast cancer patients usually display poorer prognostic results, we propose that compared to nonbasal-like breast cancer, the genome of basal-like breast cancer patients is more unstable, with miRNAs more often being downregulated and gene amplification occurring more frequently, as gene loss and gene rearrangement do. Thus, the production as well as the activity of miRNAs is expected to be lower in basal-like cancer patients, which is consistent with our results.

We did not find any relevant studies that provide any clues about the unbalance in the distribution of the survival time-associated MMPVs. However, we noted that the regulatory patterns of MMPVs related to ER and HER2 status tended to be significant in ER− and HER2− patients. ER− breast cancer patients are usually resistant to Tamoxifen therapy and, thus, show poorer prognostic results [25–27]. Similarly, most HER− breast cancer patients cannot benefit from Trastuzumab therapy, which greatly increases the survival rate in the HER2+ breast cancer patients. Based on the above results, it can be observed that the regulatory pattern of MMPVs tends to be significant in association with statuses that suggest poorer prognostic results. Thus, it is reasonable to propose that the unbalance in the distribution of the survival time-associated MMPVs may result from the unbalance that remains in the distribution of ER- and HER2-related MMPVs. This result reveals the capacity of detecting biologically important regulatory events mediated by miRNAs.

3.2. Significantly Differentially Expressed Genes (DEGs) Encoding MMPVs and Their Relationship with Biopathological Features. To explore the relationship between MMPVs and biopathological features, we conducted a significant analysis of microarray (SAM) analysis to detect MMPVs whose miRNAs and mRNAs are both significantly differentially expressed between the two statuses of a given biopathological feature (DE-MMPVs). The final results are shown in Table 2.

First, we analysed the differentially expressed genes (DEGs) among ER-associated DE-MMPVs, and we found that they shared the same miRNA: hsa-miR-375. Pedro de Souza Rocha Simonini reported that hsa-miR-375 is overexpressed in breast cancer tumours with an ER+ status and that decreasing the expression of hsa-miR-375 will decrease the activity of ER accordingly [28]. This observation is consistent with our findings. Thus, we searched the relevant literature to examine five of the target mRNAs of hsa-miR-375. FLOR1 tends to show low expression in ER+ cancers [29]. Signal traducing adaptor protein (STAP2) is regarded as a potential drug target for ER− breast cancer patients because this protein can facilitate the growth of breast cancer cells by interacting with BRK and STAT3/5 [30–33]. STAP2 can also increase the activity of NF-Kb, whose expression is negatively correlated with ER activity [34]. Interestingly, the regulatory patterns of hsa-miR-375 and STAP2 in the healthy population and in ER− and ER+ breast cancer patients are all positive.

Table 2: Distribution of DE-MMPVs associated with ER, TP53, and subtype status.

Feature	miRNA	mRNA	Regulatory pattern
ER	hsa-miR-375(D)	PRKX(U)	D_D_D*
	hsa-miR-375(D)	FOLR1(U)	D_U_D*
	hsa-miR-375(D)	STAP2(U)	U_U_U*
	hsa-miR-375(D)	KIAA0232(D)	U_U_U*
	hsa-miR-375(D)	TBX19(U)	U_D_D*
TP53	hsa-miR-7(D)	ALG3(D)	D_U_U*
	hsa-miR-155(D)	VCAM1(D)	D_U_U*
	hsa-miR-155(D)	ETS1(D)	D_U_U*
	hsa-miR-155(D)	CBFB(D)	D_U_U*
	hsa-miR-155(D)	ARL5B(D)	D_U_U*
	hsa-miR-145(U)	MUC1(U)	D_U_U*
	hsa-let-7b(U)	CCND1(U)	D_U_U*
	hsa-miR-375(U)	LDHB(D)	D_U_D*
	hsa-miR-7(D)	TCOF1(D)	D_D_U*
	hsa-miR-7(D)	KCNJ14(D)	D_D_U*
	hsa-miR-145(U)	FSCN1(D)	D_D_U*
	hsa-let-7b(U)	CHMP2A(U)	D_D_U*
	hsa-miR-375(U)	PRKX(D)	D_D_D*
	hsa-miR-29c(U)	LAMC1(U)	D_D_D*
	hsa-miR-29c(U)	DNMT3B(D)	D_D_D*
	hsa-miR-29c(U)	COL3A1(U)	D_D_D*
	hsa-miR-214(U)	HSPD1(D)	D_D_D*
	hsa-miR-155(D)	ARID2(U)	D_D_D*
	hsa-miR-145(U)	CCDC43(U)	D_D_D*
	hsa-miR-107(U)	CDK6(D)	D_D_D*
	hsa-let-7b(U)	SPCS3(D)	D_D_D*
Subtype	hsa-miR-155(D)	ETS1(D)	D_U_U*
	hsa-miR-155(D)	CSF1R(D)	D_U_U*
	hsa-miR-155(D)	CBFB(D)	D_U_U*
	hsa-miR-146a(D)	SAMD9L(D)	D_U_U*
	hsa-miR-146a(D)	EPSTI1(D)	D_U_U*
	hsa-miR-146a(D)	BCL2A1(D)	D_U_U*
	hsa-miR-145(U)	MUC1(U)	D_U_U*
	hsa-miR-375(U)	AKAP7(D)	D_U_D*
	hsa-miR-193b(U)	MAT2A(U)	D_D_U*
	hsa-miR-145(U)	FSCN1(D)	D_D_U*
	hsa-let-7b(U)	CHMP2A(U)	D_D_U*
	hsa-miR-29c(U)	CDK6(D)	D_D_D*

However, when the expression of hsa-miR-375 is significantly downregulated in ER− specimens, the expression of STAP2 changes, and instead of being downregulated, it is upregulated quite significantly.

Second, we analysed the DEGs among TP53-associated DE-MMPVs. Adan Valladares showed that CCND1 and LAMC1 are overexpressed in breast cancer patients [35]. We observed that, compared to patients with mutated TP53, the expression of CCND1 and LAMC1 is increased in patients with wild type TP53. This observation is of particular interest because it contradicts our expectation that because wild type TP53 inhibits the expression of oncogenes, CCND1 and LAMC1, which are both overexpressed in breast cancer patients, should be downregulated in individuals with wild type TP53.

It can be observed that the regulatory pattern of hsa-let-7b and CCND1 remains positive in breast cancer patients, regardless of the status of TP53, which is negative in the healthy population. Such D_U_U regulatory pattern variation was also found for hsa-miR-145 and MUC1. MUC1 encodes a mucoprotein that is overexpressed in many types of cancer, including breast cancer. Similar to CCND1, MUC1 is a potential biomarker for tumours, and MUC1 plays an important role in the invasion and metastasis of tumours. Specifically, MUC1 can interact with TP53 and inhibit cell apoptosis mediated by TP53, thus facilitating the proliferation of cancer cells [36]. Similar to hsa-let-7b and CCND1, the regulatory pattern of has-miR-145 and MUC1 shifts from being negative in the healthy population to positive in breast cancer patients. No relevant research has shown such an aberrant disturbance of regulatory patterns, and this disturbance is expected to be associated with the genesis and development of breast cancer.

Nguyen et al. reported that hsa-miR-29c negatively regulates DNMT3B, the overexpression of which can cause the hypermethylation of some tumour suppressor genes [37]. Our results show that the regulatory pattern of hsa-miR-29c and DNMT3B is D_D_D in association with TP53. Compared to patients with mutated TP53, the expression of hsa-miR-29c is significantly increased in individuals with wild type TP53, thus enforcing the repression effect on DNMT3B. This observation is supported by the finding of Toledo and Bardot that the wild type P53 protein can bind to the Drosha protein complex and enhance the transcription of tumour suppressor miRNAs [38]. Similar regulatory pattern variation occurs for hsa-miR-107 and CDK6, which tend to be overexpressed in aggressive tumours. We propose that the enhanced regulatory effect of has-miR-107 on CDK6 is also due to the combined action of P53 and Drosha.

Finally, we analysed the DEGs among the subtype-associated DE-MMPVs. We observed that 2 MMPVs (has-miR-145 and MUC1, hsa-miR-155 and CBFB) and one mRNA (CDK6) are shared by the TP53-associated MMPVs and subtype-related MMPVs. Specifically, compared to patients with mutated TP53, the levels of MUC1, CBFB, and CDK6 expression are significantly decreased, significantly increased, and significantly decreased, respectively, in wild type TP53 patients. Interestingly, compared to basal-like breast cancer patients, the levels of MUC1, CBFB, and CDK6 expression are significantly decreased, significantly increased, and significantly decreased in nonbasal-like breast cancer patients, respectively. All of these findings indicate that these three genes could play similar roles in TP53 pathways and biological pathways that are related to basal-like breast cancer. Moreover, we found that, compared to basal-like breast cancer patients, the expression of ETS is significantly decreased in nonbasal-like cancer patients. This finding is supported by the work of Charafe-Jauffret et al., who reported that the expression of ETS1 is higher in basal-like breast cancer than in other breast cancer subtypes [39]. Furthermore, compared to basal-like breast cancer, the levels of CSF1R and CBFB expression significantly decrease in nonbasal-like breast cancer. Furthermore, this finding is reasonable because CSF1R is overexpressed in invasive breast cancer and is strongly associated with a shorter survival time [40], and CBFB is regarded as a potential oncogene [1].

Table 3: Distribution of MPPVs that are associated with multiple pathological features.

Feature 1	Feature 2	Overlap
ER	Survival	3
Subtype	Survival	3
ER	Subtype	4
TP53	Subtype	6
TP53	Survival	6
ER	TP53	11
HER2	Subtype	12
HER2	Survival	14
HER2	TP53	31
HER2	ER	40

3.3. Analysis of MMPVs Associated with Multiple Biopathological Features. We found that many MMPVs are associated with multiple biopathological features. The distribution of these MMPVs is shown in Table 3.

To find genes that are closely associated with two different biopathological features, we further filtered the data that appear in Table 3. For example, to find genes that are related to both ER and TP53 status, we first selected all of the mRNAs in the 31 MMPVs that are associated with both TP53 and ER. Then, we selected the miRNAs present in TP53- and ER-associated MMPVs as well as in MMPVs associated with other biopathological features. Finally, we uncovered two genes (MYBL and M6PRBP1), which are expected to be closely related to ER and HER2.

There is a substantial amount of research examining crosstalk between ER and HER2. Isabel Pinhel claimed that the levels of HER2 and ER expression are positively correlated in non-HER2-overexpressing breast cancer tumours and are negatively correlated in HER2-overexpressing breast cancer tumours [41]. HER2 overexpression can repress the antiproliferation effect of TGF-β1 and, hence, enhance the growth of cancer cells [42]. Moreover, TGF-β1 can repress the expression of MYB, and ER+ status enhances the expression of MYB [43]. Importantly, MYB and MYBL are expected to display similar functions given that these two proteins

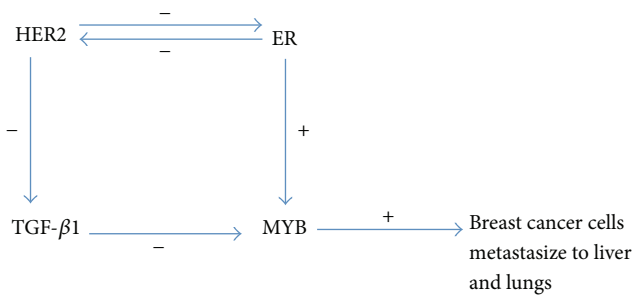


Figure 2: Proposed crosstalk between ER and HER2. Gene ontology (GO) enrichment analysis of MMPVs.

belong to the same transcription factor family, and they are homologous. It has been reported that MYB is relevant to hematopoietic function [43]. An experiment conducted by Mucenski and colleagues in MYB knockout mice showed that MYB is closely related to hematopoietic function, especially hematopoietic cells in the liver, as all MYB knockout mice ultimately die as a result of hypoxia, and their livers are anaemic and relatively small compared to the livers of mice in the control group [44]. Furthermore, the cancer cells of cancer patients exhibiting HER2 overexpression are more likely to metastasise to the lungs and liver [45]. Considering our findings together with the supporting results from the literature noted above, we propose the existence of potential crosstalk between MYB, overexpressed HER2, and ER as shown in Figure 2.

This proposed potential crosstalk between MYB, overexpressed HER2, and ER not only will contribute to further studies addressing the molecular mechanisms underlying breast cancer but also serves as an important reference for potential joint therapy with tamoxifen and trastuzumab.

We conducted a GO enrichment analysis of the mRNA components of MMPVs. The results of the biological process (BP) enrichment analysis and cellular component (CC) enrichment analysis are shown in Supplementary Table 1 in Supplementary Material available online at http://dx.doi.org/10.1155/2014/291280.

It can be observed that the GO terms related to TP53 mainly include molecules that are involved with cell adhesion to the extracellular matrix. There have been many studies concentrating on the relationship between TP53 and tumour metastasis, which is closely associated with cell adhesion and the extracellular matrix. Specifically, Abramson et al. reported that compared to individuals with wild type TP53, the strength of cell adhesion is greatly increased in the population exhibiting mutant TP53 [46]. Anaganti et al. claimed that wild type TP53 can repress the expression of focal adhesion kinase (FAK), which is a critical regulator of adhesion and motility whose overexpression is strongly associated with enhanced metastatic potential. Additionally, FAK is frequently overexpressed in populations with mutant TP53 [47].

The GO terms related to the ER are mainly associated with DNA synthesis and the cell cycle. S F Doisneau-Sixou claimed that oestrogen independently regulates the expression and function of c-Myc and cyclin D1. Antioestrogen treatment of MCF-7 cells can greatly decrease the expression of c-Myc and cyclin D1, resulting in the arrest of the cell cycle and inhibition of DNA synthesis.

Table 4: Comparison of the average degree of the different types of MPPVs with that in the HPRD-PPI network.

Feature	Average degree	P value
ER	11.76	0.03
TP53	11.23	**0.01**
HER2	11.74	**6.38E − 08**
Survival	12.81	**5E − 03**
Subtype	10.60	0.095
HPRD-PPI network	7.80	

3.4. Topological Features of Genes Encoding MMPVs. We explored the topological features of the genes encoding MMPVs by mapping the mRNAs of each type of MMPV to the HPRD protein-protein-interaction (PPI) network [48]. Specifically, we employed Student's t-test to compare the average degree of the MMPV genes with those of PPI network. In fact, there are biases which existed in current PPI databases. Here we use the commonly used database, HPRD. The results are shown in Table 4. Except for subtype-associated MMPVs, the average degree of the mRNAs related to all of the other biopathological features is significantly greater than the average degree in the HPRD-PPI network.

P values were calculated using a one-tailed t-test. The P value shown in bold indicates that the average degree of the corresponding biopathological feature is significantly greater than that in the HPRD-PPI network.

4. Conclusions

In this study, we discovered that the regulatory pattern of miRNA-mRNA pairs can vary with different statuses of biopathological features. To further explore the molecular mechanisms underlying breast cancer, we studied five biopathological features (the ER, HER2 and TP53 genes, cancer subtype, and survival time) that are closely related to breast cancer. We observed a general unbalance in the distribution of MMPVs. Moreover, the differentially expressed MMPV genes suggest that there is a potential effect of these biopathological features on the development of breast cancer at the molecular level. Furthermore, we examined the topological features of genes encoding MMPVs in the HPRD PPI network, and we propose the existence of potential crosstalk between ER and HER2. The method developed in this paper can help detecting biologically important regulatory events mediated by miRNAs.

Conflict of Interests

The authors declare that no conflict of interests exists.

Acknowledgments

This work was supported by Department of Health of Heilongjiang Province Foundation of China, Grant no. 2011-059, and Department of Education of Heilongjiang Province Foundation of China, Grant no. 11541149.

References

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

[2] D. P. Bartel, "MicroRNAs: genomics, biogenesis, mechanism, and function," *Cell*, vol. 116, no. 2, pp. 281–297, 2004.

[3] M. J. Bueno, I. P. de Castro, and M. Malumbres, "Control of cell proliferation pathways by microRNAs," *Cell Cycle*, vol. 7, no. 20, pp. 3143–3148, 2008.

[4] D. P. Bartel, "MicroRNAs: target recognition and regulatory functions," *Cell*, vol. 136, no. 2, pp. 215–233, 2009.

[5] B. Kusenda, M. Mraz, J. Mayer, and S. Pospisilova, "MicroRNA biogenesis, functionality and cancer relevance," *Biomedical Papers of the Medical Faculty of the University Palacký*, vol. 150, no. 2, pp. 205–215, 2006.

[6] R. Zhang and B. Su, "Small but influential: the role of microRNAs on gene regulatory network and 3′UTR evolution," *Journal of Genetics and Genomics*, vol. 36, no. 1, pp. 1–6, 2009.

[7] S. Vasudevan, Y. Tong, and J. A. Steitz, "Switching from repression to activation: microRNAs can up-regulate translation," *Science*, vol. 318, no. 5858, pp. 1931–1934, 2007.

[8] 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.

[9] J. Yu, F. Liu, P. Yin et al., "Integrating miRNA and mRNA expression profiles in response to heat stress-induced injury in rat small intestine," *Functional & Integrative Genomics*, vol. 11, no. 2, pp. 203–213, 2011.

[10] 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.

[11] J. A. Nielsen, P. Lau, D. Maric, J. L. Barker, and L. D. Hudson, "Integrating microRNA and mRNA expression profiles of neuronal progenitors to identify regulatory networks underlying the onset of cortical neurogenesis," *BMC Neuroscience*, vol. 10, article 98, 2009.

[12] G. Maire, J. W. Martin, M. Yoshimoto, S. Chilton-MacNeill, M. Zielenska, and J. A. Squire, "Analysis of miRNA-gene expression-genomic profiles reveals complex mechanisms of microRNA deregulation in osteosarcoma," *Cancer Genetics*, vol. 204, no. 3, pp. 138–146, 2011.

[13] C. Wang, Z. Su, N. Sanai et al., "microRNA expression profile and differentially-expressed genes in prolactinomas following bromocriptine treatment," *Oncology Reports*, vol. 27, no. 5, pp. 1312–1320, 2012.

[14] K. W. Tsai, Y. L. Liao, C. W. Wu et al., "Aberrant hypermethylation of miR-9 genes in gastric cancer," *Epigenetics*, vol. 6, no. 10, pp. 1189–1197, 2011.

[15] E. Dudziec, A. Gogol-Döring, V. Cookson, W. Chen, and J. Catto, "Integrated epigenome profiling of repressive histone modifications, DNA methylation and gene expression in normal and malignant urothelial cells," *PLoS ONE*, vol. 7, no. 3, Article ID e32750, 2012.

[16] E. Enerly, I. Steinfeld, K. Kleivi et al., "miRNA-mRNA integrated analysis reveals roles for miRNAs in primary breast tumors," *PLoS ONE*, vol. 6, no. 2, Article ID e16915, 2011.

[17] T. Vergoulis, I. S. Vlachos, and Panagiotis Alexiou, "TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support," *Nucleic Acids Research*, vol. 40, database issue, pp. D222–D229, 2012.

[18] 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, article 48, 2009.

[19] S. Menard, E. Tagliabue, M. Campiglio, and S. M. Pupa, "Role of HER2 gene overexpression in breast carcinoma," *Journal of Cellular Physiology*, vol. 182, no. 2, pp. 150–162, 2000.

[20] C. Cheng, X. Fu, P. Alves, and M. Gerstein, "MRNA expression profiles show differential regulatory effects of microRNAs between estrogen receptor-positive and estrogen receptor-negative breast cancer," *Genome Biology*, vol. 10, no. 9, article R90, 2009.

[21] H. I. Suzuki, K. Yamagata, K. Sugimoto, T. Iwamoto, S. Kato, and K. Miyazono, "Modulation of microRNA processing by p53," *Nature*, vol. 460, no. 7254, pp. 529–533, 2009.

[22] G. A. Calin, C. Sevignani, C. D. Dumitru et al., "Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 9, pp. 2999–3004, 2004.

[23] E. A. Rakha, J. S. Reis-Filho, and I. O. Ellis, "Basal-like breast cancer: a critical review," *Journal of Clinical Oncology*, vol. 26, no. 15, pp. 2568–2581, 2008.

[24] C. Blenkiron, L. D. Goldstein, N. P. Thorne et al., "MicroRNA expression profiling of human breast cancer identifies new markers of tumor subtype," *Genome Biology*, vol. 8, no. 10, article R214, 2007.

[25] G. M. Clark and W. L. McGuire, "Prognostic factors in primary breast cancer," *Breast Cancer Research and Treatment*, vol. 3, supplement 1, pp. S69–S72, 1983.

[26] Early Breast Cancer Trialists' Collaborative Group, "Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials," *The Lancet*, vol. 365, no. 9472, pp. 1687–1717, 2005.

[27] S. M. Thorpe, "Estrogen and progesterone receptor determinations in breast cancer: technology, biology and clinical significance," *Acta Oncologica*, vol. 27, no. 1, pp. 1–19, 1988.

[28] P. D. S. Rocha Simonini, A. Breiling, N. Gupta et al., "Epigenetically deregulated microRNA-375 is involved in a positive feedback loop with estrogen receptor α in breast cancer cells," *Cancer Research*, vol. 70, no. 22, pp. 9175–9184, 2010.

[29] K. M. M. Kelley, B. G. Rowan, and M. Ratnam, "Modulation of the folate receptor α gene by the estrogen receptor: mechanism and implications in tumor targeting," *Cancer Research*, vol. 63, no. 11, pp. 2820–2828, 2003.

[30] P. J. Mitchell, E. A. Sara, and M. R. Crompton, "A novel adaptor-like protein which is a substrate for the non-receptor tyrosine kinase, BRK," *Oncogene*, vol. 19, no. 37, pp. 4273–4282, 2000.

[31] O. Ikeda, Y. Sekine, T. Yasui et al., "STAP-2 negatively regulates both canonical and noncanonical NF-κB activation induced by Epstein-Barr virus-derived latent membrane protein," *Molecular and Cellular Biology*, vol. 28, no. 16, pp. 5027–5042, 2008.

[32] O. Ikeda, Y. Sekine, A. Mizushima et al., "Interactions of STAP-2 with Brk and STAT3 participate in cell growth of human breast

cancer cells," *Journal of Biological Chemistry*, vol. 285, no. 49, pp. 38093–38103, 2010.

[33] O. Ikeda, A. Mizushima, Y. Sekine et al., "Involvement of STAP-2 in Brk-mediated phosphorylation and activation of STAT5 in breast cancer cells," *Cancer Science*, vol. 102, no. 4, pp. 756–761, 2011.

[34] D. K. Biswas, A. P. Cruz, E. Gansberger, and A. B. Pardee, "Epidermal growth factor-induced nuclear factor κB activation: a major pathway of cell-cycle progression in estrogen-receptor negative breast cancer cells," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 15, pp. 8542–8547, 2000.

[35] A. Valladares, N. G. Hernández, F. S. Gómez et al., "Genetic expression profiles and chromosomal alterations in sporadic breast cancer in Mexican women," *Cancer Genetics and Cytogenetics*, vol. 170, no. 2, pp. 147–151, 2006.

[36] X. Wei, H. Xu, and D. Kufe, "Human mucin 1 oncoprotein represses transcription of the p53 tumor suppressor gene," *Cancer Research*, vol. 67, no. 4, pp. 1853–1858, 2007.

[37] T. Nguyen, C. Kuo, M. B. Nicholl et al., "Downregulation of microRNA-29c is associated with hypermethylation of tumor-related genes and disease outcome in cutaneous melanoma," *Epigenetics*, vol. 6, no. 3, pp. 388–394, 2011.

[38] F. Toledo and B. Bardot, "Cancer: three birds with one stone," *Nature*, vol. 460, no. 7254, pp. 466–467, 2009.

[39] E. Charafe-Jauffret, C. Ginestier, F. Monville et al., "Gene expression profiling of breast cell lines identifies potential new basal markers," *Oncogene*, vol. 25, no. 15, pp. 2273–2284, 2006.

[40] H. M. Kluger, M. Dolled-Filhart, S. Rodov, B. M. Kacinski, R. L. Camp, and D. L. Rimm, "Macrophage colony-stimulating factor-1 receptor expression is associated with poor outcome in breast cancer by large cohort tissue microarray analysis," *Clinical Cancer Research*, vol. 10, no. 1, pp. 173–177, 2004.

[41] I. Pinhel, M. Hills, S. Drury et al., "ER and HER2 expression are positively correlated in HER2 non-overexpressing breast cancer," *Breast Cancer Research*, vol. 14, no. 2, article R46, 2012.

[42] C. A. Wilson, E. E. Cajulis, J. L. Green et al., "HER-2 overexpression differentially alters transforming growth factor-beta responses in luminal versus mesenchymal human breast cancer cells," *Breast Cancer Research*, vol. 7, no. 6, pp. R1058–R1079, 2005.

[43] R. G. Ramsay and T. J. Gonda, "MYB function in normal and cancer cells," *Nature Reviews Cancer*, vol. 8, no. 7, pp. 523–534, 2008.

[44] M. L. Mucenski, K. McLain, A. B. Kier et al., "A functional c-myb gene is required for normal murine fetal hepatic hematopoiesis," *Cell*, vol. 65, no. 4, pp. 677–689, 1991.

[45] Y. M. Li, Y. Pan, Y. Wei et al., "Upregulation of CXCR4 is essential for HER2-mediated tumor metastasis," *Cancer Cell*, vol. 6, no. 5, pp. 459–469, 2004.

[46] V. G. Abramson, A. B. Troxel, M. Feldman et al., "Cyclin D1b in human breast carcinoma and coexpression with cyclin D1a is associated with poor outcome," *Anticancer Research*, vol. 30, no. 4, pp. 1279–1285, 2010.

[47] S. Anaganti, L. Fernández-Cuesta, A. Langerød, P. Hainaut, and M. Olivier, "p53-dependent repression of focal adhesion kinase in response to estradiol in breast cancer cell-lines," *Cancer Letters*, vol. 300, no. 2, pp. 215–224, 2011.

[48] Y. Yang, L. Han, Y. Yuan, J. Li, N. Hei, and H. Liang, "Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types," *Nature Communications*, vol. 5, article 3231, 2014.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 971857, 7 pages
http://dx.doi.org/10.1155/2014/971857

Research Article

O18Quant: A Semiautomatic Strategy for Quantitative Analysis of High-Resolution $^{16}O/^{18}O$ Labeled Data

Yan Guo,[1] Masaru Miyagi,[2] Rong Zeng,[3] and Quanhu Sheng[1,3]

[1] *Center for Quantitative Sciences, Vanderbilt University, Nashville, TN 37027, USA*
[2] *Center for Proteomics and Bioinformatics, School of Medicine, Case Western Reserve University, Cleveland, OH 44106, USA*
[3] *Key Laboratory of Systems Biology, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Science, Chinese Academy of Sciences, Shanghai 200031, China*

Correspondence should be addressed to Quanhu Sheng; qhsheng@sibs.ac.cn

Received 28 February 2014; Accepted 18 April 2014; Published 11 May 2014

Academic Editor: Leng Han

Proteolytic ^{18}O-labeling has been widely used in quantitative proteomics since it can uniformly label all peptides from different kinds of proteins. There have been multiple algorithms and tools developed over the last few years to analyze high-resolution proteolytic $^{16}O/^{18}O$ labeled mass spectra. We have developed a software package, O18Quant, which addresses two major issues in the previously developed algorithms. First, O18Quant uses a robust linear model (RLM) for peptide-to-protein ratio estimation. RLM can minimize the effect of outliers instead of iteratively removing them which is a common practice in other approaches. Second, the existing algorithms lack applicable implementation. We address this by implementing O18Quant using C# under Microsoft.net framework and R. O18Quant automatically calculates the peptide/protein relative ratio and provides a friendly graphical user interface (GUI) which allows the user to manually validate the quantification results at scan, peptide, and protein levels. The intuitive GUI of O18Quant can greatly enhance the user's visualization and understanding of the data analysis. O18Quant can be downloaded for free as part of the software suite ProteomicsTools.

1. Introduction

Proteomic research refers to high-throughput studies of large amount of proteins. With the rise of high-throughput sequencing, many researchers have shifted their focus to the genome using RNAseq technology. However, high-throughput sequencing technology does not help us answer proteomic questions, and the study of proteomics provides an entirely different level of genomic understanding. For example, messenger RNA abundance does not always translate into protein abundance [1], posttranslational modification is not observable through RNAseq, and protein degradation rate may play a significant role in protein content [2]. Thus, proteomics should always be a pivotal part of our quest to understand the complete human biology.

Mass spectrometry is a powerful method for quantifying proteins. It produces spectra of masses of molecules from the protein. The spectra can be used to determine the isotopic signature of the sample. Labeling is a nonoptional step in mass spectrometry. There are currently four major labeling techniques: SILAC, ICAT, ITRAQ, and ^{18}O. Compared to other labeling techniques, ^{18}O labeling requires less reagents and synthetic steps. However, ^{18}O labeling does require extra time and labels. Our software O18Quant is specially designed for ^{18}O labeled data.

Isotopic labeling has been commonly used for the quantification of peptides and proteins in biological samples [3, 4]. A natural extension of isotopic labeling is isotope dilution analysis [5]. Isotope dilution analysis is usually conducted in comparative scenarios, because it is difficult to accurately obtain absolute measurement [6]. During the comparative method, usually, labeled proteins obtained from an unstressed system are pooled together with the same amount of unlabeled protein from a second stressed system.

Then, mass spectrometry is performed on the combined pool to obtain differentially expressed proteins between stressed and unstressed systems.

Researchers have developed a more convenient isotope dilution approach taking advantage of ^{18}O, which can be easily added to peptides by the enzyme-catalyzed incorporation of oxygen in the C-terminal carboxylic acids during the digestion step [7]. A quick equilibrium can be achieved by exchanging at either or both of the C-terminal carboxyl oxygen atoms if the kinetics for complex formation is faster than the digestion time. Thus, the $^{18}O/^{16}O$ ratio can be used to estimate the relative abundance of the protein between the stressed and unstressed systems. Since the early 2000s, proteolytic ^{18}O-labeling has been commonly adopted for use in comparative proteomics because it can uniformly label all peptides from different kinds of proteins [8–10].

During the last ten years, multiple algorithms [11–15] have been developed to analyze high-resolution proteolytic $^{16}O/^{18}O$ labeled mass spectra. Unfortunately, the majority of these algorithms lack actual implementation. Few software packages are freely available for users. Thus, there is a strong interest in developing a software package for $^{16}O/^{18}O$ labeled protein ratio calculation and validation. Here we present a semiautomatic tool, O18Quant, for analysis of such data. O18Quant differs from other previously published algorithms in two major ways. First, O18Quant has been implemented using C# and R, and a useable package is available for download. Second, O18Quant uses RLM to compute protein ratios. RLM accounts for the effect of outlier peptides instead of completely removing them iteratively. RLM has also been used in the evaluation of peptide identifications [16], reducing technical variability in functional protein microarrays [17], and SILAC peptide ratio calculation [18].

O18Quant calculates the protein ratio automatically based on user-predefined parameters such as purity of ^{18}O water and resolution of mass spectrometry. Then the quantification results can be manually validated at scan, peptide, and protein levels through a user-friendly GUI. Only protein quantifications that pass quality control at all three levels are considered to be used in further analysis. O18Quant and its source code can be downloaded freely from https://github.com/shengqh/RCPA.Tools/releases/ and its detailed introduction can be viewed at https://github.com/shengqh/RCPA.Tools/wiki/.

2. Materials and Methods

2.1. Preparation of ^{18}O Labeled Test Samples. To demonstrate O18Quant's effectiveness, we excised retina samples from a three-month-old male Sprague-Dawley weanling rat (Harlan Inc., Indianapolis, IN) as described previously [19]. The excised retinas were suspended in 400 μL of 100 mM ammonium bicarbonate containing a protease inhibitor cocktail (Sigma-Aldrich, St. Louis, MO), and proteins were extracted by ultrasonication (4.5 kHz three times for 9 s with a 3 min pause on ice between the strokes) using a VirSonic 100 ultrasonic cell disrupter (SP Scientific, Gardiner, NY). The resulting protein extract was centrifuged at 15,000 g for 10 min, and the supernatant was collected. The proteins in the supernatant were then precipitated by mixing with a 4-fold excess volume of ice-cold acetone and left for 1 h at −20°C. The protein precipitate was solubilized in 400 μL of formic acid-methanol (1 : 1, v/v) and subjected to performic acid oxidation to oxidatively cleave disulfide bonds [20]. After the reaction, the reaction mixture was dried in a SpeedVac, redissolved in 200 μL of 100 mM ammonium bicarbonate containing 2 M urea, and the amount of dissolved protein was determined with a DC protein assay kit (Bio-Rad, Hercules, CA). A total of 100 μg of protein was digested by trypsin (1 : 50 substrate to protein ratio, w/w) at 25°C for 16 h. After the digestion, the digest was desalted using Vydac C18 UltraMicro Tip Column (The Nest Group, Southborough, MA) as per the manufacturer's instructions, divided equally into two tubes, and dried in a SpeedVac. Then, the digests were redissolved in 100 μL 100 mM *N*-ethylmorpholine-acetic acid buffer at pH 6 made either with $H_2\,^{16}O$ or $H_2\,^{18}O$. The peptides were then incubated with trypsin (1 : 50 substrate to protein ratio, w/w) at 25°C for 16 h to incorporate ^{16}O or ^{18}O, respectively, into the carboxyl termini of the peptides [21]. Following the oxygen labeling reaction, the reaction mixtures were dried, redissolved in 100 μL formic acid-methanol (1 : 1, v/v), and subjected to performic acid oxidation to inactivate trypsin.

2.2. LC-MS/MS Analysis. The resulting ^{16}O and ^{18}O labeled samples were dissolved in 0.1% formic acid, mixed in 1 : 2, 1 : 1, and 2 : 1 ratios, and analyzed by LC-MS/MS using an UltiMate 3000 LC system (Dionex, San Francisco, CA, USA) interfaced to an LTQ-Orbitrap XL mass spectrometer (Thermo-Finnigan, Bremen, Germany) [22]. Peptides were chromatographed on a reverse phase column (C18, 75 μm × 150 mm, 3 μm, 100 Å; Dionex) using a linear gradient of acetonitrile from 0% to 40% in aqueous 0.1% formic acid over a period of 90 minutes at 300 nL/minute. The mass spectrometer was operated in a data-dependent MS to MS/MS switching mode, with the eight most intense ions in each MS scan subjected to MS/MS analysis. MS spectra were acquired at 60,000 resolution (FWHM) in the Orbitrap detector (~1 s cycle time) and MS/MS spectra were in the ion trap by collision-induced dissociation (CID). Automatic gain control (AGC) target for MS acquisition was set to 5×10^5. Maximum ion injection times for MS1 and MS2 were 500 and 100 ms, respectively. The threshold intensity for the MS/MS trigger was set at 1,000 and normalized collision energy (NCE) at 35. The data were collected in profile mode for the full scan and in centroid mode for the MS/MS scans. The dynamic exclusion function for previously selected precursor ions was enabled during the analysis such that the following parameters were applied: repeat count of two, repeat duration of 45 s, exclusion duration of 60 s, and exclusion size list of 150. Xcalibur software (version 2.2, Thermo-Finnigan Inc.) was used for instrument control, data acquisition, and data processing.

2.3. Mass Spectrometry Data Analysis. Proteins were identified by comparing all of the experimental peptide MS/MS spectra to the Swiss-Prot database using Mascot database

search software (version 2.3.2, Matrix Science, London, UK). Oxidation of cysteine to cysteic acid and methionine to methionine sulfone was set as fixed modifications while the modification of the C-terminal carboxyl group with ^{18}O was a variable modification. The mass tolerance was set to 10 ppm for the precursor ion and to 0.8 Da for the product ion. Strict trypsin specificity was applied, allowing for one missed cleavage. Only peptides with a minimum score of 20 were considered significant. BuildSummary [23] was used to generate a confident protein list with a false discovery rate for both peptide and protein of ≤0.01. Only the proteins with at least two unique peptides were used in quantification analysis.

2.4. Peptide Abundance Estimation. For each identified peptide with observed mass-to-charge m/z, charge z, ^{18}O modification state s, and user-defined purity of $H_2{}^{18}O$ p, the abundance of the peptide from the light sample A (light) and the heavy sample A (heavy) was calculated.

The nature form, the heavy forms with one or two ^{18}O labels of the peptide as ^{16}O, $^{18}O_1$, and $^{18}O_2$, and the abundance of those three forms are A (^{16}O), A ($^{18}O_1$), and A ($^{18}O_2$), the corresponding mass-to-charge of those three forms can be theoretically predicted by formula (1a) to (1c), respectively:

$$\frac{m}{z\left(^{16}O\right)} = \begin{cases} \frac{m}{z} - 2 * \frac{d}{z} & \text{if } s = \text{modified} \\ m & \text{if } s = \text{unmodified,} \end{cases} \quad (1a)$$

$$\frac{m}{z\left(^{18}O_1\right)} = \begin{cases} \frac{m}{z} - \frac{d}{z} & \text{if } s = \text{modified} \\ \frac{m}{z} + \frac{d}{z} & \text{if } s = \text{unmodified,} \end{cases} \quad (1b)$$

$$\frac{m}{z\left(^{18}O_2\right)} = \begin{cases} \frac{m}{z} & \text{if } s = \text{modified} \\ \frac{m}{z} + 2 * \frac{d}{z} & \text{if } s = \text{unmodified,} \end{cases} \quad (1c)$$

where d = mass(^{18}O isotope) − mass(^{16}O isotope).

The potential isotope cluster of the peptide is predicted as $mp = \{m(^{16}O), m(^{16}O) + c, m(^{18}O_1), m(^{18}O_1) + c, m(^{18}O_2), m(^{18}O_2) + c\}$, where c = mass(^{13}C isotope) − mass(^{12}C isotope).

From the scan in which the peptide is identified, the ions of the potential isotope cluster in previous and next scans are extracted from the raw data until the charges of both $m/z(^{16}O)$ and $m/z(^{18}O_2)$ ions equal zero meaning there is not enough evidence to support the isotope cluster in that scan. Assuming that there are n scans containing a potential isotope cluster, an overall observed abundance vector $Y = \{y_1, y_2, y_3, y_4, y_5, y_6\}$ is calculated by formula (2) where i indicates the position of the ion in the isotope cluster and k indicates the kth scan. Consider

$$y_i = \sum_{k=1}^{n} a_{k,i}. \quad (2)$$

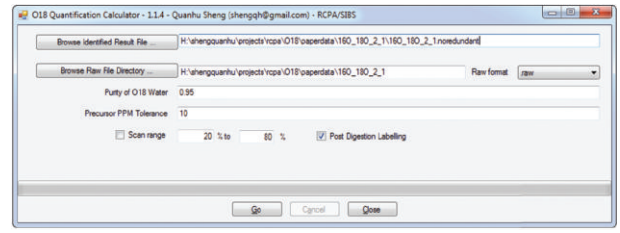


Figure 1: O18 Quantification calculator. The interface is used to calculate peptide/protein relative ratios automatically. User can control the values of various parameters and load in raw data using this interface.

A theoretical isotopic abundance vector $V = \{v_1, v_2, v_3, v_4, v_5, v_6\}$ is generated by emass algorithm [24] based on the sequence of the peptide p. Then a matrix X is constructed, where each row indicates the theoretical isotopic abundance contributed by ^{16}O, $^{18}O_1$, and $^{18}O_2$, respectively, for an ion in the isotope cluster. Consider

$$X = \begin{bmatrix} v_1 & 0 & 0 \\ v_2 & 0 & 0 \\ v_3 & v_1 & 0 \\ v_4 & v_2 & 0 \\ v_5 & v_3 & v_1 \\ v_6 & v_4 & v_2 \end{bmatrix}. \quad (3)$$

The expected abundance vector $A = \{A(^{16}O), A(^{18}O_1), A(^{18}O_2)\}$ can be estimated by solving the formula (4) using the nonnegative least square model:

$$\overline{y} = X * \overline{A}. \quad (4)$$

Then, A (light) and A (heavy) are calculated by the method described by Mason et al. [13].

2.5. Protein Quantification. For each protein, multiple peptides may be detected and quantified. Other than combining an outlier rejection scheme with other peptide-to-protein algorithms [25], a robust fitting of linear models is used in our method to estimate the protein ratio from the unlabeled and labeled abundance of each peptide. Detailed information about the algorithm is described in the R package "MASS" (http://cran.r-project.org/web/packages/MASS/index.html).

3. Results and Discussion

3.1. Implementation. The software was implemented using C# and R. R environment is required for peptide-to-protein ratio calculation. Two GUIs are built into O18Quant. The first GUI, O18 Quantification Calculator (Figure 1), is used to automatically extract ions of a potential isotope cluster from the raw file, calculate peptide abundance, estimate protein ratios, and export preliminary quantification result to a tab delimited file. The second GUI, O18 Quantification Summary Viewer (Figure 2), is used to load the preliminary quantification result, validate the result at protein, peptide,

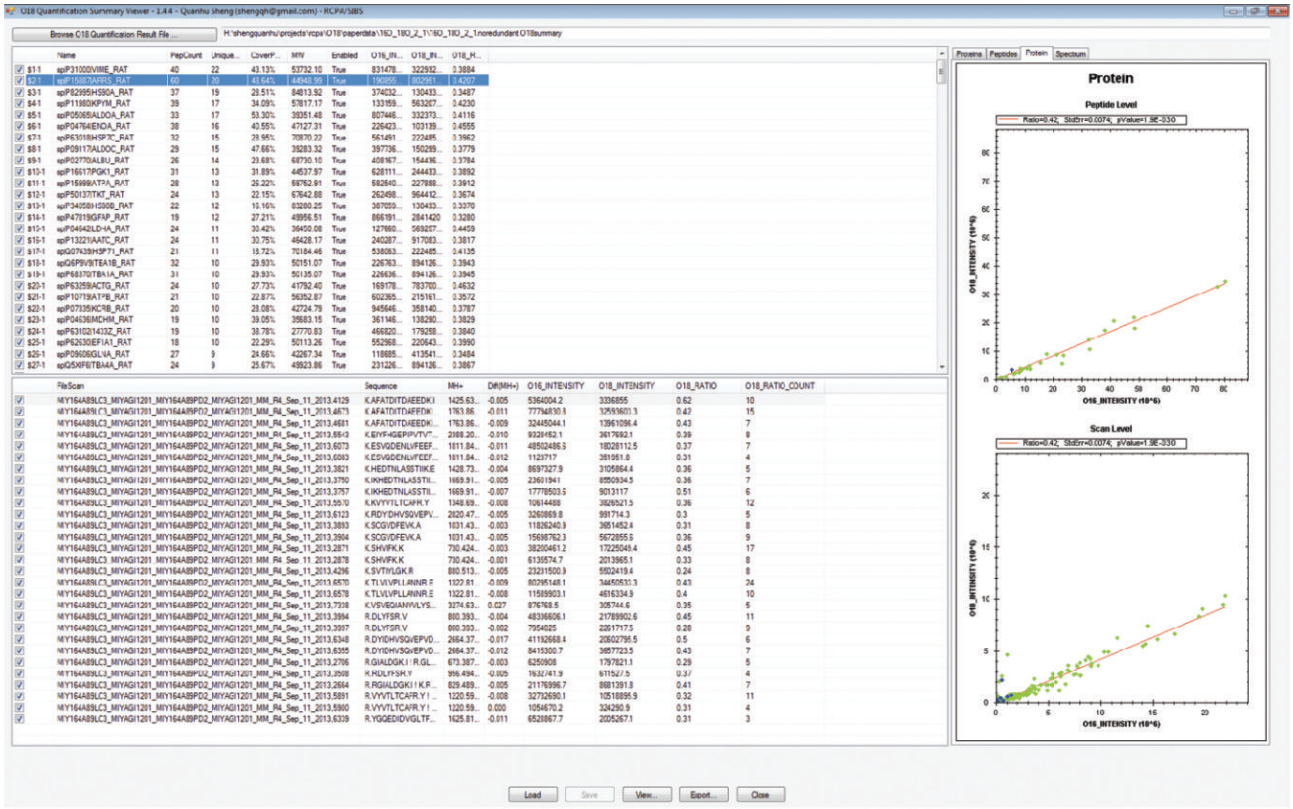


Figure 2: O18 Quantification Summary Viewer. The interface is used to validate the quantification result at protein/peptide/scan level.

and scan levels, and export the validated results. The protein and peptide information are displayed in a spread sheet (Figure 2 left). The scatter plot and RLM fitted line can be visualized within this GUI (Figure 2 right). This GUI allows users to perform visual quality control and manually exclude proteins and peptides with problematic ratios using simple point and click controls.

3.2. Visualization and Validation. Three levels of quantification information were stored and visualized for validation.

(1) *Protein Level.* Ideally, the peptides from the same protein should have similar relative ratios. From the plot of light/heavy abundances of peptides for each protein, we can easily identify outlier peptides for further validation.

(2) *Peptide Level.* For the questionable peptides, the overall scan information of each peptide can be used to validate if the LC peak boundary is properly detected.

(3) *Scan Level.* The profile of ion intensity in each scan can be used to validate the scan quality.

To demonstrate the practicality and efficiency of O18Quant's visualization functionality, we chose the protein ATP5I_RAT with the highest ratio of 65.8 in a 2 : 1 sample to validate if that ratio was correct (see Supplementary Figure 1, the first entry in top left table and the red spots in top right and bottom right graphs, in Supplementary Material available online at http://dx.doi.org/10.1155/2014/971857). Two peptides were quantified in protein ATP5I_RAT with respective ratios of 3.09 and 50 (Supplementary 1 Figure 1, the bottom left table). The one with sequence R.YSYLKPR.A and ratio 50 was highly questionable. The corresponding peptide validation page was opened by double clicking the peptide entry (Supplementary 1, Figure 2). The first seven scans contained an unusual 18O(1) ion whose abundance was larger than both 16O and 18O(2) ions. The directions of mass difference between theoretical and observed 16O/O18(1)/O18(2) ions were also different between the first seven scans and the last four scans. Both observations indicate that the detected ions in the first seven scans might belong to another peptide with very similar elution time, and the precursor m/z of that peptide was very close to the 18O(1) ion of peptide YSYLKPR. Then, the peptide abundance was calculated using only the last four scans. The ratio of the peptide became 3.36 and the ratio of the protein became 3.13, which was more similar to the designed ratio. Detailed validation procedures are described at Supplementary 1.

3.3. Quantification Result. Table 1 illustrates the identified and quantified peptides/proteins in three known-ratio samples. All proteins with at least two peptides identified were

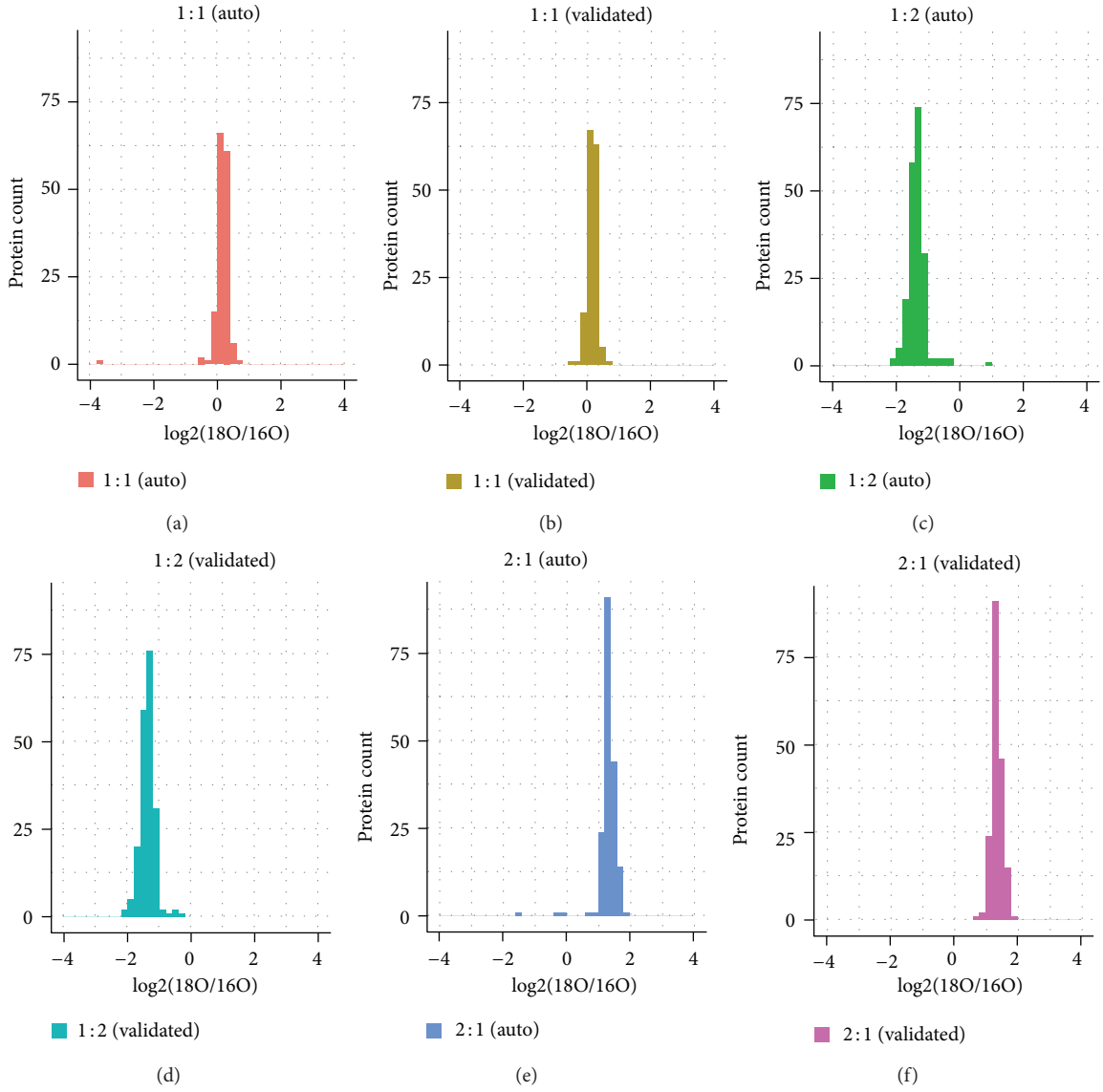


Figure 3: Histogram of log2(ratio) for the three known-ratio samples. Top/bottom three graphs were generated from the data before/after manual validation.

Table 1: Identified and quantified proteins from three known-ratio samples.

Sample	Identified peptides	Identified proteins	Identified unique 2 proteins*	Quantified peptides	Quantified proteins	Quantified unique 2 proteins*
O18/O16 = 1 : 1	752	257	138	726	251	138
O18/O16 = 1 : 2	993	325	180	961	315	180
O18/O16 = 2 : 1	813	281	162	779	272	162

*Unique 2 protein means that protein was identified with at least two unique peptides.

quantified while some proteins with only one peptide identified failed to be quantified. Here, unique peptides mean peptides with identical sequences without considering their modification states.

The quantification result before and after careful manual validation of the three samples with designed labeled/unlabeled ratio 1 : 1, 1 : 2, and 2 : 1, respectively, was illustrated as in Figure 3. Only proteins with at least two unique peptides identified were used. The mean and standard deviation of log2(ratio) before manual validation from the three samples were 0.15 ± 0.34, -1.35 ± 0.29, and 1.34 ± 0.47. After careful validation of the peptides with extreme ratios, the mean

and standard deviation of log2(ratio) from the three samples became 0.18 ± 0.14, -1.38 ± 0.23, and 1.35 ± 0.17. The standard deviations decreased significantly.

3.4. Export Protein/Peptide Summary. After manual validation, the quantification result can be exported to CSV format at protein, peptide, and scan levels for further analysis with additional customizable features. O18Quant allows the protein and peptide level quantification information to be exported into single or separated files. O18Quant is the only tool publicly available now that can export the quantification result at all three levels.

4. Conclusions

Proteomic research remains a key component in unlocking the treatment of many human diseases. Here, we present O18Quant, a software package implemented using Microsoft.net framework (C#) and R. O18Quant improves the previous $^{18}O/^{16}O$ estimation algorithms in two major areas. First, we employed the RLM model to account for the effect of outliers/extreme values rather than removing them. Second, O18Quant can automate the process of calculating the peptide/protein relative ratio with an intuitive user-friendly GUI. The GUI provides tremendous convenience for users to conduct validation of the quantification results at scan, peptide, and protein levels. O18Quant is free and it will be consistently supported in the coming years.

Conflict of Interests

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

Acknowledgment

The authors would like to thank Margot Bjoring for her editorial support.

References

[1] S. P. Gygi, Y. Rochon, B. R. Franza, and R. Aebersold, "Correlation between protein and mRNA abundance in yeast," *Molecular and Cellular Biology*, vol. 19, no. 3, pp. 1720–1730, 1999.

[2] A. Belle, A. Tanay, L. Bitincka, R. Shamir, and E. K. O'Shea, "Quantification of protein half-lives in the budding yeast proteome," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 35, pp. 13004–13009, 2006.

[3] T. P. Conrads, K. Alving, T. D. Veenstra et al., "Quantitative analysis of bacterial and mammalian proteomes using a combination of cysteine affinity tags and 15N-metabolic labeling," *Analytical Chemistry*, vol. 73, no. 9, pp. 2132–2139, 2001.

[4] M. Munchbach, M. Quadroni, G. Miotto, and P. James, "Quantitation and facilitated de novo sequencing of proteins by isotopic N-terminal labeling of peptides with a fragmentation-directing moiety," *Analytical Chemistry*, vol. 72, no. 17, pp. 4047–4057, 2000.

[5] I. I. Stewart, T. Thomson, and D. Figeys, "O labeling: a tool for proteomics," *Rapid Communications in Mass Spectrometry*, vol. 15, no. 24, pp. 2456–2465, 2001.

[6] R. D. Smith, L. Pasa-Tolic, M. S. Lipton et al., "Rapid quantitative measurements of proteomes by Fourier transform ion cyclotron resonance mass spectrometry," *Electrophoresis*, vol. 22, pp. 1652–1668, 2001.

[7] N. Sharon, V. Grisaro, and H. Neumann, "Pepsin-catalyzed exchange of oxygen atoms between water and carboxylic acids," *Archives of Biochemistry and Biophysics*, vol. 97, no. 1, pp. 219–221, 1962.

[8] J. L. Capelo, R. J. Carreira, L. Fernandes, C. Lodeiro, H. M. Santos, and J. Simal-Gandara, "Latest developments in sample treatment for ^{18}O-isotopic labeling for proteomics mass spectrometry-based approaches: a critical review," *Talanta*, vol. 80, no. 4, pp. 1476–1486, 2010.

[9] C. Fenselau and X. Yao, "$^{18}O_2$-Labeling in quantitative proteomic strategies: a status report," *Journal of Proteome Research*, vol. 8, no. 5, pp. 2140–2143, 2009.

[10] M. Miyagi and K. C. S. Rao, "Proteolytic ^{18}O-labeling strategies for quantitative proteomics," *Mass Spectrometry Reviews*, vol. 26, no. 1, pp. 121–136, 2007.

[11] J. Fernandez-De-Cossio, "Mass spectrum patterns of ^{18}O-tagged peptides labeled by enzyme-catalyzed oxygen exchange," *Analytical Chemistry*, vol. 83, no. 8, pp. 2890–2896, 2011.

[12] I. Jorge, P. Navarro, P. Martínez-Acedo et al., "Statistical model to analyze quantitative proteomics data obtained by $^{18}O/^{16}O$ labeling and linear ion trap mass spectrometry: application to the study of vascular endothelial growth factor-induced angiogenesis in endothelial cells," *Molecular and Cellular Proteomics*, vol. 8, no. 5, pp. 1130–1149, 2009.

[13] C. J. Mason, T. M. Therneau, J. E. Eckel-Passow et al., "A method for automatically interpreting mass spectra of ^{18}O-labeled isotopic clusters," *Molecular and Cellular Proteomics*, vol. 6, no. 2, pp. 305–318, 2007.

[14] A. Ramos-Fernández, D. López-Ferrer, and J. Vázquez, "Improved method for differential expression proteomics using trypsin-catalyzed ^{18}O labeling with a correction for labeling efficiency," *Molecular and Cellular Proteomics*, vol. 6, no. 7, pp. 1274–1286, 2007.

[15] Q. Zhu, D. Valkenborg, and T. Burzykowski, "Markov-chain-based heteroscedastic regression model for the analysis of high-resolution enzymatically ^{18}O-labeled mass spectra," *Journal of Proteome Research*, vol. 9, no. 5, pp. 2669–2677, 2010.

[16] H. Xu, L. Yang, and M. A. Freitas, "A robust linear regression based algorithm for automated evaluation of peptide identifications from shotgun proteomics by use of reversed-phase liquid chromatography retention time," *BMC Bioinformatics*, vol. 9, article 347, 2008.

[17] A. Sboner, A. Karpikov, G. Chen et al., "Robust-linear-model normalization to reduce technical variability in functional protein microarrays," *Journal of Proteome Research*, vol. 8, no. 12, pp. 5451–5464, 2009.

[18] X. Guan, N. Rastogi, M. R. Parthun, and M. A. Freitas, "SILAC peptide ratio calculator: a tool for SILAC quantitation of peptides and post-translational modifications," *Journal of Proteome Research*, vol. 13, no. 2, pp. 506–516, 2014.

[19] D. Hajkova, Y. Imanishi, V. Palamalai et al., "Proteomic changes in the photoreceptor outer segment upon intense light exposure," *Journal of Proteome Research*, vol. 9, no. 2, pp. 1173–1181, 2010.

[20] C. H. W. Hirs, "Performic acid oxidation," in *Methods in Enzymology*, C. H. W. Hirs, Ed., pp. 197–199, Academic Press, 1967.

[21] C. S. R. Kadiyala, S. E. Tomechko, and M. Miyagi, "Perfluorooctanoic acid for shotgun proteomics," *PLoS ONE*, vol. 5, no. 12, Article ID e15332, 2010.

[22] J. Wanner, R. Subbaiah, Y. Skomorovska-Prokvolit et al., "Proteomic profiling and functional characterization of early and late shoulder osteoarthritis," *Arthritis Research & Therapy*, vol. 15, article R180, 2013.

[23] Q. Sheng, J. Dai, Y. Wu, H. Tang, and R. Zeng, "BuildSummary: using a group-based approach to improve the sensitivity of peptide/protein identification in shotgun proteomics," *Journal of Proteome Research*, vol. 11, no. 3, pp. 1494–1502, 2012.

[24] A. L. Rockwood and P. Haimi, "Efficient calculation of accurate masses of isotopic peaks," *Journal of the American Society for Mass Spectrometry*, vol. 17, no. 3, pp. 415–419, 2006.

[25] B. Carrillo, C. Yanofsky, S. Laboissiere, R. Nadon, and R. E. Kearney, "Methods for combining peptide intensities to estimate relative protein abundance," *Bioinformatics*, vol. 26, no. 1, pp. 98–103, 2010.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 708903, 17 pages
http://dx.doi.org/10.1155/2014/708903

Research Article

Antioxidant Defense Enzyme Genes and Asthma Susceptibility: Gender-Specific Effects and Heterogeneity in Gene-Gene Interactions between Pathogenetic Variants of the Disease

Alexey V. Polonikov,[1] Vladimir P. Ivanov,[1] Alexey D. Bogomazov,[2] Maxim B. Freidin,[3] Thomas Illig,[4,5] and Maria A. Solodilova[1]

[1] *Department of Biology, Medical Genetics and Ecology, Kursk State Medical University, 3 Karl Marx Street, Kursk 305041, Russia*
[2] *Department of Pediatrics, Kursk State Medical University, 11a Koltsov Street, Kursk 305035, Russia*
[3] *Research Institute for Medical Genetics, Siberian Branch of Russian Academy of Medical Sciences, 10 Nabereznaya Ushaiki Tomsk 634050, Russia*
[4] *Research Unit of Molecular Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany*
[5] *Hanover Unified Biobank, Hanover Medical School, Carl-Neuberg-Strasse 1, 30625 Hanover, Germany*

Correspondence should be addressed to Alexey V. Polonikov; polonikov@rambler.ru

Received 23 February 2014; Revised 5 April 2014; Accepted 7 April 2014; Published 5 May 2014

Academic Editor: Siyuan Zheng

Oxidative stress resulting from an increased amount of reactive oxygen species and an imbalance between oxidants and antioxidants plays an important role in the pathogenesis of asthma. The present study tested the hypothesis that genetic susceptibility to allergic and nonallergic variants of asthma is determined by complex interactions between genes encoding antioxidant defense enzymes (ADE). We carried out a comprehensive analysis of the associations between adult asthma and 46 single nucleotide polymorphisms of 34 ADE genes and 12 other candidate genes of asthma in Russian population using set association analysis and multifactor dimensionality reduction approaches. We found for the first time epistatic interactions between ADE genes underlying asthma susceptibility and the genetic heterogeneity between allergic and nonallergic variants of the disease. We identified *GSR* (glutathione reductase) and *PON2* (paraoxonase 2) as novel candidate genes for asthma susceptibility. We observed gender-specific effects of ADE genes on the risk of asthma. The results of the study demonstrate complexity and diversity of interactions between genes involved in oxidative stress underlying susceptibility to allergic and nonallergic asthma.

1. Introduction

Bronchial asthma (BA) is a common chronic inflammatory disease of the airways characterized by variable and recurring symptoms, reversible airflow obstruction, and bronchospasm [1]. There is a considerable body of evidence demonstrating that asthma is a multifactorial disease which results from complex interactions between susceptibility genes of small-to-modest effects and equally important environmental factors [2, 3].

In the recent years, the relationships between common genetic variants and BA risk are being reported with rapidly increasing frequency. Large-scale genome-wide association studies (GWAS) have been recently done to look for asthma susceptibility genes in ethnically diverse populations of the world [4, 5]. However, the findings obtained by GWAS were limited by the strongest associations of a few number of genetic variants that achieved genome-wide significance level. In addition, the genome-wide associations are found to be quite difficult to interpret with respect to disease pathogenesis [4–7]. Meanwhile, hundreds to thousands of genetic markers associated with a disease risk are not interpreted because they have not achieved the genome-wide significance level, thus accounting for the "missing heritability" of

complex diseases [8]. This means that GWAS approach is powerless in the detecting genes of small-to-modest effects which represent a polygenic background of multifactorial disease. Additionally, genetic diversity of human populations [9], heterogeneity of disease pathogenesis [10], and especially a complexity of gene-gene interactions [11, 12] complicate our opportunities in unraveling the molecular mechanisms underlying complex diseases including asthma.

It is widely agreed that the expression of a disease phenotype may not accurately be predicted from the knowledge of the effects of individual genes because of complex nonlinear interactions between genes, including epistatic and additive interactions [8, 13]. To address this issue, several powerful data-mining approaches have been developed to identify susceptibility genes involved in such complex interactions [14–16]. One of them, multifactor dimensionality reduction (MDR) method, was developed to reduce the dimensionality of multilocus information to improve the ability to detect genetic combinations that confer disease risk in relatively small samples [16, 17]. With set association analysis (SAA), contributions from multiple SNPs are combined by forming a sum of single-marker statistics, which results in a single genome-wide test statistic with high power [14].

An important task for a genetic epidemiologist utilizing a candidate gene approach is the selection of appropriate genes and SNPs for testing the disease association. Compared with studying individual genes, the inferences derived from a hypothesis-driven candidate pathway study are enhanced by allowing global conclusions about the involvement of entire biochemical pathway to the pathogenesis of disease [18]. Following this approach in our previous study [19], we pointed out the potential relevance of toxicogenomic mechanisms of BA in the modern world and proposed that genes for xenobiotic-metabolizing enzymes would be the most appropriate candidate genes for asthma susceptibility whose effects on the disease risk can be associated with exposure to air pollution. Due to the fact that air pollutants are the sources of reactive oxygen species (ROS), genes involved in oxidative stress can potentiate harmful effects of xenobiotics on the respiratory system [20, 21].

It is well known that oxidative stress resulting from an increased amount of ROS and an imbalance between oxidants and antioxidants plays a role in the molecular mechanisms underlying BA [22–24]. We have demonstrated that genes of antioxidant defense enzymes (ADE), such as glutamate cysteine ligase (*GCLM*) [25], glutathione peroxidase (*GPX1*) [26], catalase (*CAT*) [27], myeloperoxidase (*MPO*) [28], NADPH oxidase (*CYBA,* p22phox subunit) [29, 30], NAD(P)H: quinone oxidoreductase type 1 (*NQO1*) [31] and microsomal epoxide hydrolase (*EPHX1*) [19], are important determinants of genetic susceptibility to asthma in Russians. In the present study, we tested the hypothesis that genetic susceptibility to both allergic and nonallergic asthma is determined by complex interactions between genes involved in oxidative stress. We performed for the first time a comprehensive analysis of genomic interactions between 34 ADE genes and 12 other candidate genes in order to identify gene-gene interactions in redox homeostasis underlying polygenic mechanisms of BA.

2. Materials and Methods

2.1. Study Population. The study protocol was approved by the Ethical Review Committee of Kursk State Medical University, and written informed consent was obtained from each participant before the study. The participants comprised a total of 429 unrelated individuals (215 patients with asthma and 214 healthy controls); all are ethnically Russians from Central Russia (mainly from the Kursk region). All study subjects were recruited from the Division of Pulmonology at the Kursk Regional Clinical Hospital between 2003 and 2004. Asthma was diagnosed by qualified pulmonologists on the basis of the WHO criteria, as described previously [19, 32]. The mean age of the patients with asthma (94 men and 121 women) was 43.3 years (ranging from 16 to 67 years), and the mean age of the healthy subjects (105 men and 109 women) was 41.3 years (ranging from 17 to 84 years). Skin prick tests were conducted and total serum IgE levels were determined in all study subjects. Patients with positive skin prick tests and high level of total IgE were defined as patients with allergic asthma (64 men and 92 women). Asthmatics who showed either negative skin prick test results (wheal size: <5 mm) or a normal total IgE level (<0.35 IU) were considered to be patients with nonallergic asthma (29 men and 27 women). Data on allergic status were not available for three asthmatics. A strong positive family history of asthma was found in the case group (40.1%) in comparison with controls (6.7%).

2.2. Selection of Candidate Genes. The candidate genes for this study were selected according to the guidelines for genetic association studies proposed by Cooper and coauthors [33]. We used the following criteria to select ADE genes and their genetic polymorphisms satisfying our study's purposes: (1) enzymes should represent key players involved in the regulation of redox processes; (2) enzymes should cover all biochemical pathways of redox homeostasis entirely, including enzymes possessing antioxidant activity (*GPX1*, *SOD2*, *CAT*, *GSTM1*, etc.) and those with prooxidant activity (i.e., ROS-generating enzymes such as CYBA, MPO, and CYP2E1); (3) enzymes should be expressed in the lung and/or airways (the expression patterns of the selected ADE genes in human tissues and organs are shown in Supplementary Material available online at http://dx.doi.org/10.1155/2014/708903); (4) SNPs should be functionally significant, whenever possible; and (5) minor allele frequency should be more than 5%. Following these criteria, 34 polymorphisms of 24 ADE genes have been selected from published literature and public databases.

2.3. DNA Extraction and Genotyping. Genomic DNA of all study participants was isolated from 5–10 mL of peripheral blood samples, collected in K3-EDTA tubes by venipuncture, and maintained at −20°C until processed. Twenty-five of the selected gene polymorphisms had been genotyped in our previous studies [19, 25–31]. In the present study, another nine ADE gene polymorphisms such as *GPX2* (rs17880492), *GPX3* (rs2070593) *GPX4* (rs713041), *GSR* (rs2551715), *SOD2* (rs4880), *SOD3* (rs2536512), *PRDX1* (rs17522918), *TXNRD1*

(rs1128446), and *FMO3* (rs2266782) have been genotyped. Additionally, we genotyped 12 polymorphisms of 9 candidate genes of asthma such as *TNF* (rs1800629), *IL1B* (rs16944), *IL3* (rs40401 and rs31480), *IL5* (rs2069812), *CSF2RB* (rs131840), *IL9* (rs2069885), *SCGB1A1* (rs11549442), and *SERPINA1* (rs17580, rs143370956, and rs11568814). Majority of them have been reported to be associated with the risk of asthma and/or asthma-related phenotypes in Russians [34–37]. A complete list of 46 studied SNPs is given in Table 1. Genotyping of the selected polymorphisms was done using restriction fragment length polymorphism assays according to the published protocols (genotyping protocols are available upon request). All of the genotyping was done blindly to the case-control status and the repeatability test was conducted for the 5% of total subjects, resulting in a 100% concordance rate.

2.4. Statistical Analysis. The concordance of genotypes prevalence in patients with asthma and healthy controls with values expected under Hardy-Weinberg equilibrium was assessed by Pearson's chi-square test. The association between ADE gene polymorphisms and asthma was examined with binary logistic regression analysis with calculation of odds ratios (OR) and 95% confidence intervals (CI). The statistical calculations were done using Statistica for Windows (v8.0) software package (StatSoft; Tulsa, OK, USA). The statistical significance was established at the $P \leq 0.05$ level. Bonferroni correction for P values (P_{adj}) was applied in cases when multiple tests were performed.

Two bioinformatic approaches, SAA and MDR, were applied for the analysis of gene-gene interactions. The principle of SAA is described in detail elsewhere [14, 55] and implemented in a statistical program SUMSTAT (http://linkage.rockefeller.edu/ott/sumstat.html/). Briefly, the method combines the information derived from measurements of allelic/genotype association and departure from Hardy-Weinberg equilibrium into a single, genome-wide statistic. The markers with high Hardy-Weinberg disequilibrium (HWD) values in the control group are trimmed and are not considered for further analysis. For the remaining markers, effects of allelic/genotype association with disease and HWD values are then combined into a single *Sum* statistic [56]. P values reported by the program were calculated by permutations. The number of permutation tests was set at 10000.

MDR is a flexible nonparametric and genetic model free method for analysis of high-order nonlinear or nonadditive gene-gene interactions [15]. The method has been proposed to overcome limitation of logistic regression which deals with many factors simultaneously and fails to characterize epistatic models in the absence of main effects, due to the hierarchical model-building process leading to an increase in type II errors and decreased power [17]. The MDR method uses a constructive induction algorithm that converts two or more variables such as SNPs into a single attribute. In particular, SNPs are pooled into high and low risk group, effectively reducing the multifactor prediction from n dimension to one dimension. Best models for each locus combination are selected by repeating the analysis for up to 10 seeds after shuffling the order of individuals and applying 10-fold cross-validation each time. Average of cross-validation consistency (CVC) together with training and test accuracy is calculated for each locus combination. CVC is defined as the number of times a particular interaction model is selected across 10 cross-validation datasets. We performed statistical calculations using MDR software (http://www.multifactordimensionalityreduction.org/). Statistical significance of the best models selected for each SNP combination was determined using 1000-fold permutation testing. The significance of the final MDR model was determined empirically by 1000 permutations using the Monte-Carlo procedure implemented into the MDRpt software (http://sourceforge.net/projects/mdr/). P values for CVC were considered statistically significant at ≤0.05 levels. To visualize and interpret the results obtained from MDR, we used interaction dendrograms.

Both the SAA and MDR are limited by the identification of a few number of high penetrance interacting genes, whereas a larger portion of genes of low-to-moderate effects remain out of the analysis. To address this issue, we performed post hoc comparisons of two-locus genotype combinations (only for those SNPs which were found in gene-gene interaction models obtained by SAA and/or MDR methods) between the case and control groups to look for the genotype combinations which determine the risk of asthma. The observed associations were adjusted for multiple tests using Bonferroni procedure.

3. Results

3.1. Allele and Genotype Frequencies in Asthmatics and Controls. Allele and genotype frequencies of the studied genes are shown in Tables 2 and 3, respectively. After adjusting for multiple tests, the only statistically significant association was found between the *IL5* C-703T polymorphism and BA. The −703CC genotype was found to be associated with the risk of allergic asthma (OR = 0.44; 95% CI 0.29–0.67; $P = 0.0001$ ($P_{adj} = 0.004$). In gender-specific analysis, this association was seen in both men (OR = 0.50; 95% CI 0.27–0.94; $P = 0.03$) and women (OR = 0.40; 95% CI 0.22–0.70; $P = 0.001$) but did not reach a statistical significance after Bonferroni correction for multiple tests ($P > 0.05$). No association of this genotype was found with nonallergic asthma in both sexes.

3.2. Gene-Gene Interactions in Asthma Revealed by Set Association Analysis. Taking into account the polygenic basis of asthma, it was an important task to investigate high-order gene-gene interactions using specialized bioinformatics approach called set association analysis which captures the simultaneous effects of multiple genes and achieves a global view of gene action and interaction [14, 55]. For trimming, we considered values of departures from HWE (HWD values) exceeding the 99th percentile of chi-square ($\chi^2 \geq 6.6$, $df = 1$) in the control group [14]. There were no HWD values larger than 6.6 in the control subjects, so the trimming

Table 1: Description of the polymorphisms included in this study.

Gene symbols (HGNC)		Gene name	Polymorphism (SNP)	Location	SNP ID
1	2	3	4	5	6
1	*GPX1*	Glutathione peroxidase 1	C>T (P198L)	exon 1	rs1050450
2	*GPX2*	Glutathione peroxidase 2 (gastrointestinal)	G>A (R146C)	exon 2	rs17880492
3	*GPX3*	Glutathione peroxidase 3 (plasma)	249G>A	3′ UTR	rs2070593
4	*GPX4*	Glutathione peroxidase 4 (phospholipid hydroperoxidase)	C718T	3′ UTR	rs713041
5	*GSR*	Glutathione reductase	T>C (30546636T>C)	intron 9	rs2551715
6	*SOD2*	Superoxide dismutase 2, mitochondrial	A16V	exon 2	rs4880
7	*SOD3*	Superoxide dismutase 3, extracellular	A40T (A58T)	exon 3	rs2536512
8	*CAT*	Catalase	−21A>T (−89A>T)	5′ UTR	rs7943316
9	*CAT*	—//—	−262C>T (4760C>T)	5′ UTR	rs1001179
10	*GCLM*	Glutamate-cysteine ligase, modifier subunit	−588C>T (4704C>T)	5′ UTR	rs41303970
11	*GCLM*		−23G>T	5′UTR	rs743119
12	*NQO1*	NAD(P)H dehydrogenase, quinone 1	P187S	exon 6	rs1800566
13	*NQO1*	—//—	R139W	exon 4	rs4986998
14	*CYBA*	Cytochrome b-245, alpha polypeptide	242C>T (Y72H)	exon 4	rs4673
15	*CYBA*	—//—	640A>G (24G>A)	3′ UTR	rs1049255
16	*CYBA*	—//—	−930A>G	5′ UTR	rs9932581
17	*MPO*	Myeloperoxidase	−463G>A (4535G>A)	5′ UTR	rs2333227
18	*PRDX1*	Peroxiredoxin 1	C>A	5′ UTR	rs17522918
19	*TXNRD1*	Thioredoxin reductase 1	C>G	5′ UTR	rs1128446
20	*FMO3*	Flavin-containing monooxygenase 3	E158K	exon 4	rs2266782
21	*CYP1A1*	Cytochrome P450, family 1, subfamily A, polypeptide 1	I462V	exon 7	rs1048943
22	*CYP1A1*		T6235C	3′ UTR	rs4646903
23	*CYP2E1*	Cytochrome P450, family 2, subfamily E, polypeptide 1	−1293G>C	5′ UTR	rs3813867
24	*CYP2E1*		−1053C>T	5′ UTR	rs2031920
25	*CYP2E1*	—//—	7632T>A	intron 6	rs6413432
26	*CYP2E1*	—//—	9896C>G	intron 7	rs2070676
27	*EPHX1*	Epoxide hydrolase 1, microsomal (xenobiotic)	Y113H (337T>C)	exon 3	rs1051740
28	*EPHX1*		H139R (416A>G)	exon 4	rs2234922
29	*PON1*	Paraoxonase 1	Q192R	exon 6	rs662
30	*PON2*	Paraoxonase 2	C311S	exon 9	rs7493
31	*GSTM1*	Glutathione S-transferase mu 1	Expressor/deletion	exons 6-7	—
32	*GSTT1*	Glutathione S-transferase theta 1	Expressor/deletion	exon 4	—
33	*GSTP1*	Glutathione S-transferase pi 1	I105V	exon 5	rs1695
34	*GSTP1*	—//—	A114V	exon 6	rs1138272
35	*TNF*	Tumor necrosis factor	−308G>A	5′ UTR	rs1800629
36	*IL1B*	Interleukin 1, beta	−511C>T	5′ UTR	rs16944
37	*IL3*	Interleukin 3 (colony-stimulating factor, multiple)	S27P	exon 1	rs40401
38	*IL3*		−15C>T	5′ UTR	rs31480
39	*IL5*	Interleukin 5 (colony-stimulating factor, eosinophil)	C-703T	5′ UTR	rs2069812
40	*CSF2RB (IL5RB)*	Colony stimulating factor 2 receptor, beta, low-affinity (granulocyte-macrophage)	G1972A	exon 5	rs131840
41	*IL9*	Interleukin 9	T113M	exon 5	rs2069885
42	*IL13*	Interleukin 13	−1111C>T	5′ UTR	rs1800925
43	*SCGB1A1 (CC16)*	Secretoglobin, family 1A, member 1 (uteroglobin)	A38G	exon 1	rs11549442
44	*SERPINA1*	Serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 1	E288V	exon 3	rs17580
45	*SERPINA1*		D365N	exon 5	rs143370956
46	*SERPINA1*		1331G>A	3′ UTR	rs11568814

Table 2: Allele frequencies of genes investigated in the present study.

Gene	Polymorphism	Alleles	Allele frequency			
			Controls (n = 214)	Asthma, entire group (n = 215)	Allergic asthma (n = 156)	Nonallergic asthma (n = 56)
1	2	3	4	5	6	7
GPX2	G>A (rs17880492)	G	0.991	0.981	0.987	0.964
		A	0.009	0.019	0.013	0.036
GPX3	G>A (rs2070593)	G	0.703	0.726	0.734	0.696
		A	0.297	0.274	0.266	0.304
GPX4	C718T (rs713041)	718T	0.402	0.391	0.407	0.348
		718C	0.598	0.609	0.593	0.652
GSR	T>C (rs2551715)	T	0.442	0.398	0.362	0.491
		C	0.558	0.602	0.638*	0.509
SOD2	A16V (rs4880)	16A	0.528	0.486	0.481	0.509
		16V	0.472	0.514	0.519	0.491
SOD3	A40T (rs2536512)	40A	0.322	0.321	0.324	0.295
		40T	0.678	0.679	0.676	0.705
PRDX1	C>A (rs17522918)	C	0.923	0.937	0.942	0.920
		A	0.077	0.063	0.058	0.080
TXNRD1	C>G (rs1128446)	C	0.808	0.821	0.808	0.857
		G	0.192	0.179	0.192	0.143
FMO3	E158K (rs2266782)	158E	0.549	0.537	0.529	0.571
		158K	0.451	0.463	0.471	0.429
TNF	−308G>A (rs1800629)	−308G	0.888	0.872	0.875	0.866
		−308A	0.112	0.128	0.125	0.134
IL1B	−511C>T (rs16944)	−511C	0.710	0.664	0.670	0.652
		−511T	0.290	0.336	0.330	0.348
IL3	S27P (rs40401)	27S	0.738	0.685	0.686	0.688
		27P	0.262	0.315	0.314	0.313
IL3	−15C>T (rs31480)	−15C	0.741	0.683	0.686	0.688
		−15T	0.259	0.317	0.314	0.313
IL5	C-703T (rs2069812)	−703C	0.673	0.778	0.788	0.732
		−703T	0.327	0.222*	0.212*	0.268
IL5RB (CSF2RB)	G1972A (rs131840)	1972G	0.831	0.866	0.872	0.848
		1972A	0.169	0.134	0.128	0.152
IL9	T113M (rs2069885)	113T	0.820	0.863	0.856	0.893
		113M	0.180	0.137	0.144	0.107
IL13	−1111C>T (rs1800925)	−1111C	0.729	0.693	0.692	0.714
		−1111T	0.271	0.307	0.308	0.286
CC16 (SCGB1A1)	A38G (rs11549442)	38A	0.347	0.367	0.372	0.366
		38G	0.653	0.633	0.628	0.634
SERPINA1	E288V (rs17580)	288E	0.993	0.991	0.990	0.991
		288V	0.007	0.009	0.010	0.009
SERPINA1	D365N (rs143370956)	365D	0.991	0.995	0.997	0.991
		365N	0.009	0.005	0.003	0.009
SERPINA1	1331G>A (rs11568814)	1331G	0.937	0.933	0.926	0.946
		1331A	0.063	0.067	0.074	0.054

* *Indicates* a difference in minor allele frequency between asthmatics and controls.

Table 3: Genotype frequencies of genes investigated in the present study.

Gene	Polymorphism	Genotypes	Genotype distributions, *n* (%)							
			Controls (*n* = 214)		Asthma, entire group (*n* = 215)		Allergic asthma (*n* = 156)		Nonallergic asthma (*n* = 56)	
1	2	3	4	5	6	7	8	9	10	11
GPX2	G>A (rs17880492)	GG	210	98.1	207	96.3	152	97.4	52	92.9
		GA	4	1.9	8	3.7	4	2.6	4	7.1
		AA	0	0.0	0	0.0	0	0.0	0	0.0
GPX3	G>A (rs2070593)	GG	105	49.1	113	52.6	83	53.2	28	50.0
		GA	91	42.5	86	40.0	63	40.4	22	39.3
		AA	18	8.4	16	7.4	10	6.4	6	10.7
GPX4	C718T (rs713041)	718TT	31	14.5	33	15.3	25	16.0	8	14.3
		718TC	110	51.4	102	47.4	77	49.4	23	41.1
		718CC	73	34.1	80	37.2	54	34.6	25	44.6
GSR	T>C (rs2551715)	TT	40	18.7	32	14.9	17	10.9*	15	26.8
		TC	109	50.9	107	49.8	79	50.6	25	44.6
		CC	65	30.4	76	35.3	60	38.5	16	28.6
SOD2	A16V (rs4880)	16AA	59	27.6	49	22.8	34	21.8	15	26.8
		16AV	108	50.5	111	51.6	82	52.6	27	48.2
		16VV	47	22.0	55	25.6	40	25.6	14	25.0
SOD3	A40T (rs2536512)	40AA	21	9.8	24	11.2	19	12.2	4	7.1
		40AT	96	44.9	90	41.9	63	40.4	25	44.6
		40TT	97	45.3	101	47.0	74	47.4	27	48.2
PRDX1	C>A (rs17522918)	CC	182	85.0	188	87.4	138	88.5	47	83.9
		CA	31	14.5	27	12.6	18	11.5	9	16.1
		AA	1	0.5	0	0.0	0	0.0	0	0.0
TXNRD1	C>G (rs1128446)	CC	140	65.4	145	67.4	101	64.7	42	75.0
		CG	66	30.8	63	29.3	50	32.1	12	21.4
		GG	8	3.7	7	3.3	5	3.2	2	3.6
FMO3	E158K (rs2266782)	158EE	57	26.6	59	27.4	39	25.0	20	35.7
		158EK	121	56.5	113	52.6	87	55.8	24	42.9
		158KK	36	16.8	43	20.0	30	19.2	12	21.4
TNF	−308G>A (rs1800629)	−308GG	170	79.4	162	75.4	118	75.6	42	75.0
		−308GA	40	18.7	51	23.7	37	23.7	13	23.2
		−308AA	4	1.9	2	0.9	1	0.6	1	1.8
IL1B	−511C>T (rs16944)	−511CC	114	53.3	91	42.1	67	43.0*	23	41.1
		−511CT	76	35.5	105	48.6	75	48.1*	27	48.2
		−511TT	24	11.2	20	9.3	14	9.0	6	10.7
IL3	S27P (rs40401)	27SS	120	56.1	104	48.2	77	49.4	25	44.6
		27SP	76	35.5	88	40.7	60	38.5	27	48.2
		273P	18	8.4	24	11.1	19	12.2	4	7.1
IL3	−15C>T (rs31480)	−15CC	120	56.1	103	47.7	77	49.4	25	44.6
		−15CT	77	36.0	89	41.2	60	38.5	27	48.2
		−15TT	17	7.9	24	11.1	19	12.2	4	7.1

Table 3: Continued.

Gene	Polymorphism	Genotypes	Genotype distributions, *n* (%)							
			Controls (*n* = 214)		Asthma, entire group (*n* = 215)		Allergic asthma (*n* = 156)		Nonallergic asthma (*n* = 56)	
IL5	C-703T (rs2069812)	−703CC	90	42.1	132	61.1	97	62.2*	31	55.4
		−703CT	108	50.5	72	33.3	52	33.3*	20	35.7*
		−703TT	16	7.5	12	5.6	7	4.5	5	8.9
IL5RB (CSF2RB)	G1972A (rs131840)	1972GG	136	67.7	160	74.1	118	75.6	39	69.6
		1972GA	62	30.9	54	25.0	36	23.1	17	30.4
		1972AA	3	1.5	2	0.9	2	1.3	0	0.0
IL9	T113M (rs2069885)	113TT	146	68.2	159	73.6	113	72.4	44	78.6
		113TM	59	27.6	55	25.5	41	26.3	12	21.4
		113MM	9	4.2	2	0.9	2	1.3	0	0.0
IL13	−1111C>T (rs1800925)	−1111CC	114	53.3	101	47.0	75	48.1	26	46.4
		−1111CT	84	39.3	96	44.7	66	42.3	28	50.0
		−1111CT	16	7.5	18	8.4	15	9.6	2	3.6
CC16 (SCGB1A1)	A38G (rs11549442)	38AA	25	11.8	28	13.0	23	14.7	5	8.9
		38AG	97	45.8	102	47.4	70	44.9	31	55.4
		38GG	90	42.5	85	39.5	63	40.4	20	35.7
SERPINA1	E288V (rs17580)	288EE	211	98.6	212	98.1	153	98.1	55	98.2
		288EV	3	1.4	4	1.9	3	1.9	1	1.8
		288VV	—	—	—	—	—	—	—	—
SERPINA1	D365N (rs143370956)	365DD	210	98.1	214	99.1	153	98.1	55	98.2
		365DN	4	1.9	2	0.9	3	1.9	1	1.8
		365NN	—	—	—	—	—	—	—	—
SERPINA1	1331G>A (rs11568814)	1331GG	188	87.9	187	86.6	133	85.3	50	89.3
		1331GA	25	11.7	29	13.4	23	14.7	6	10.7
		1331AA	1	0.5	0	0.0	0	0.0	0	0.0

**Indicates* a difference in genotype frequency between asthmatics and controls.

procedure to our dataset was avoided. To calculate single-locus test statistics, we used the difference in the distribution of genotypes for the ith SNP between cases and controls (chi-square test for 2×3 tables). We tested up to $N = 46$ sums (S_n) in allergic and nonallergic asthma, separately in men and women. When we added more SNPs to the S, P values tended to increase; that is, adding additional markers introduces noise to the S. In allergic asthma, the smallest P values were obtained for a sum of 5 SNPs (the *GPX1* P198L, *CAT* −21A>T, *EPHX1* H139R, *GCLM* −588C>T, and *IL5* C-703T) for men and a sum of 3 SNPs (the *EPHX1* Y113H, *NQO1* R139W, and *IL5* C-703T) for women (Figure 1). After 1000 permutation tests, the global significance levels ($P_{\min}$) of 0.0042 for men and 0.0001 for women were obtained. In nonallergic asthma, the smallest significance levels appeared for 5 SNPs (the *GCLM* −588C>T, *GSR* T>C, *CAT* −21A>T, *CYBA* −930A>G, and *EPHX1* Y113H) in men and for 2 SNPs (the *EPHX1* Y113H and *GPX2* G>A) in women (Figure 2). The global significance levels of 0.0003 for men and 0.0001 for women were obtained.

3.3. Modeling for Gene-Gene Interactions in Asthma Using MDR Method. The MDR method was used for a purpose of modeling gene-gene interactions underlying allergic and nonallergic asthma in men and women. Firstly, we used an exhaustive search algorithm to evaluate all interactions among all possible subsets of the polymorphisms. Table 4 shows the cross-validation consistency and the prediction error for gene-gene interactions (from two- to four-locus interactions) obtained from MDR analysis in both allergic and nonallergic asthma. The only statistically significant (empirical $P = 0.001$) three-locus model involving interactions between *EPHX1* Y113H, *IL5* C-703T, and *GPX1* P198L loci was discovered. The model had a minimum prediction error of 40.9 and a maximum cross-validation consistency of 50% in allergic asthma in women ($P_{\min} = 0.001$). None of the rest n-locus models in both allergic and nonallergic asthma showed a statistical significance in the MDR analysis using an exhaustive search algorithm, thereby motivating us to apply a forced search algorithm for further MDR analyses in order to build the best n-locus models in men with allergic asthma and in both sexes with nonallergic asthma. Following this approach, we obtained one statistically significant (empirical $P = 0.001$) four-locus model comprising interactions between *CAT* −21A>T, *GPX2* G>A, *GSR T>C*, and *IL5* C-703T in men with allergic asthma. The model had a minimum prediction error of 26.8 and a maximum cross-validation consistency of 100%.

Figure 3 shows the dendrograms illustrating high-order gene-gene interactions between the ADE loci in the pathogenetic variants of asthma in men and women. According to the figure, there is a strong difference in the structure of gene-gene interactions between men and women. In particular, synergistic interaction effect was found between the *GSR T>C* and *IL5* C-703T loci in men. Moreover, the *CAT* −21A>T and *GPX2* G>A gene polymorphisms had a strong antagonistic effect on the risk of allergic asthma in men. On the contrary, the hierarchical cluster analysis of the MDR data in women showed that the *CYBA* 640A>G, *GPX4* C718T, and *PON2* S311C gene polymorphisms have a strong synergistic interaction effect on the risk of allergic asthma. The *EPHX1* Y113H and *IL5* C-703T SNPs had a moderate antagonistic effect on the allergic asthma risk in women. Also, a relatively independent effect of the *GPX1* P198L gene polymorphism on the risk of allergic asthma was seen.

On the next step, a forced search algorithm was applied to analyze all possible n-locus interactions in nonallergic asthma. The best 4-locus model including *GPX1* P198L, *GPX3* G/A, *CYBA* −930A>G, and *FMO3* E158K polymorphisms was found in men (empirical $P = 0.01$). The model had a minimum prediction error of 26.1 and a cross-validation consistency of 100%. The forced MDR analysis performed in women revealed a model including *GPX1* P198L, *GPX2* G>A, *EPHX1* Y113H, and *IL5* C-703T polymorphisms with a minimum prediction error of 28.1 and a cross-validation consistency of 100% (empirical $P = 0.001$).

3.4. Post Hoc Association Analysis of Two-Locus Genotype Combinations. Then, we performed a post hoc comparison of genotype frequencies between the case and control groups with a focus on those ADE genes which were present in gene-gene interaction models obtained using SAA and MDR methods. Ten and nine two-locus combinations were found to be associated with allergic asthma in men and women, respectively (Table 5). However, only one genotype combination *GPX4* 718TC × *CYBA* 640AG achieved statistically significant inverse association with the risk of allergic asthma in women after adjustment for multiple tests (OR = 0.37; 95% CI 0.20–0.71; $P_{adj} = 0.002$). Twelve and five two-locus genotype combinations were found to be associated with the risk of nonallergic asthma in men and women, respectively (Table 6). Four two-locus genotype combinations showed statistically significant associations with nonallergic asthma in men after Bonferroni correction for multiple comparisons: *GPX1* 198PL × *CAT* −21AA (OR = 11.45; 95% CI 2.49–52.66; $P_{adj} = 0.001$), *GSR* TT × *GCLM* −588CT (OR = 11.58; 95% CI 3.07–43.72; $P_{adj} = 0.0001$), *CAT* −21AA × *CYBA* −930GG (OR = 15.64; 95% CI 2.44–100.3; $P_{adj} = 0.001$), and *GCLM* −588CT × *CYBA* −930GG (OR = 6.71; 95% CI 2.5–17.96; $P_{adj} < 0.0001$). One genotype combination *EPHX1* 113HH × *IL5* −703CC showed a significant association with increased risk of nonallergic asthma in women (OR = 8.58; 95% CI 2.43–30.26; $P_{adj} = 0.001$).

4. Discussion

4.1. A Summary of the Study Findings. The main purpose of our study was to investigate a comprehensive contribution of ADE genes to genetic susceptibility to allergic and nonallergic variants of BA. The single-locus analysis revealed that none of the ADE genes was associated with the risk of asthma. However, using two bioinformatics approaches, we found multilocus gene-gene interactions which are associated with the risk of allergic and nonallergic asthma in men and women in a gender-specific manner. Further, post hoc analysis allowed revealing two-locus combinations of genotypes which are

Table 4: A summary of best 2-, 3-, and 4-locus models of gene-gene interactions obtained by MDR analysis in allergic and nonallergic asthma (exhaustive search algorithm).

Number of loci	Best *n*-locus (2-, 3-, and 4-locus) models of gene-gene interactions	Cross-validation consistency, %	Prediction error, %
	Allergic asthma (men)		
2	*CYP2E1* 9896C>G × *IL5* C-703T	40	52.5
3	*CAT* −21A>T × *IL5* C-703T × *GSR* T/C*	50	50.3
4	*CAT* −21A>T × *GPX1* P198L × *PON2* S311C × *IL3* S27P	30	53.6
	Allergic asthma (women)		
2	*EPHX1* Y113H × *IL5* C-703T	50	45.8
3	*EPHX1* Y113H × *IL5* C-703T × *GPX1* P198L**	50	40.9
4	*EPHX1* Y113H × *CYBA* 640A>G × *GPX4* C718T × *PON2* S311C	20	50.1
	Nonallergic asthma (men)		
2	*GPX3* G/A × *GCLM* −588C>T	30	52.3
3	*GCLM* −588C>T × *GSR* T/C × *CYBA* 242C>T	20	57.0
4	*GPX3* G/A × *FMO3* E158K × *GPX1* P198L × *CYBA* −930A>G*	30	49.2
	Nonallergic asthma (women)		
2	*EPHX1* Y113H × *IL3* S27P	30	46.7
3	*EPHX1* Y113H × *GSR* T/C × *SOD2* A16V	60	45.5
4	*EPHX1* Y113H × *GSR* T/C × *SOD2* A16V × *CYBA* 640A>G*	90	27.4

* *Indicates* best *n*-locus model of gene-gene interactions evaluated through 1000 permutation tests.
** A statistically significant (*P* value 0.001) model of gene-gene interactions.

Table 5: Associations of genotype combinations with risk of allergic asthma (stratified by gender).

Combinations of genotypes	Allergic asthma		Controls		Chi-square (*P* value[1])	OR (95% CI)
	N	%	*N*	%		
			Men			
GPX1 198PL × *GPX2* GG	34	53.1	38	36.2	4.66 (0.03)	2.00 (1.06–3.76)
GPX1 198PL × *GSR* TC	20	31.3	19	18.1	3.88 (0.05)	2.06 (1.00–4.25)
GPX1 198PL × *CAT* −21AA	7	10.9	2	1.9	4.77 (0.03)	5.40 (1.25–23.42)
GPX1 198PL × *GCLM* −588CT	12	18.8	6	5.7	5.80 (0.02)	3.64 (1.33–9.96)
GPX1 198PL × *IL5* −703CC	18	28.1	13	12.4	6.58 (0.01)	2.77 (1.25–6.14)
GPX2 GG × *CAT* −21AA	12	18.8	7	6.7	4.67 (0.03)	3.13 (1.19–8.21)
GPX2 GG × *IL5* −703CC	38	59.4	44	41.9	4.86 (0.03)	2.03 (1.08–3.81)
GSR CC × *IL5* −703CC	17	26.6	11	10.5	7.44 (0.01)	3.09 (1.34–7.13)
CAT −21AA × *IL5* −703CC	9	14.1	3	2.9	5.97 (0.01)	5.01 (1.41–17.8)
CAT −21AT × *IL5* −703CT	6	9.4	28	26.7	6.36 (0.01)	0.30 (0.12–0.76)
			Women			
NQO1 187PP × *IL5* −703CC	43	46.7	33	30.3	5.75 (0.02)	2.02 (1.13–3.60)
NQO1 187PP × *IL5* −703CT	18	19.6	39	35.8	6.46 (0.01)	0.44 (0.23–0.83)
NQO1 187PP × *IL5* −703CC	43	46.7	33	30.3	5.75 (0.02)	2.02 (1.13–3.60)
NQO1 187PP × *IL5* −703CT	18	19.6	39	35.8	6.46 (0.01)	0.44 (0.23–0.83)
EPHX1 113YY × *IL5* −703CT	9	9.8	24	22.0	4.59 (0.03)	0.40 (0.18–0.89)
EPHX1 113YY × *IL5* −703CT	9	9.8	24	22.0	4.59 (0.03)	0.40 (0.18–0.89)
GPX1 198PL × *GPX4* 718TT	11	12.0	4	3.7	3.83 (0.05)	3.31 (1.07–10.22)
GPX1 198PP × *CYBA* 640AG	13	14.1	31	28.4	5.97 (0.01)	0.41 (0.20–0.85)
GPX4 718TC × *CYBA* 640AG	18	19.6	43	39.4	9.33 (0.002)*	0.37 (0.20–0.71)

[1] *Means* unadjusted *P* value. *P* value of 0.002 (P_{adj}: adjusted for multiple tests) was set as statistically significant (* a statistically significant association).

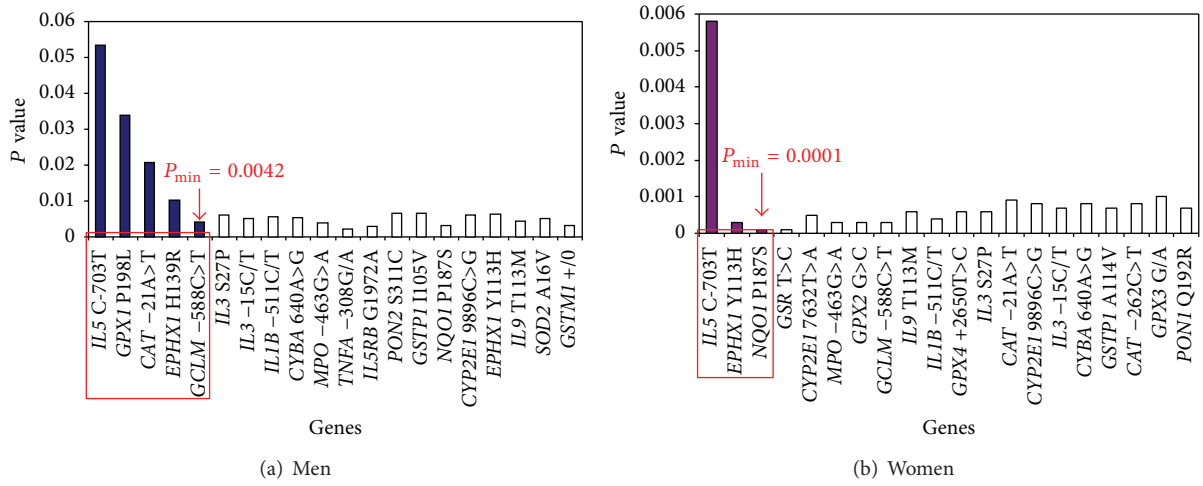


FIGURE 1: The results of statistical modeling of gene-gene interactions in allergic asthma using set association approach. Significance level of S_n statistic as a function of the number n of SNPs in different genes which are included at each step for gene-gene interactions analysis. The smallest significance level, P_{min}, occurs with 5 SNPs in males and with **3** SNPs in females. The interacting genes in the models are circled in red.

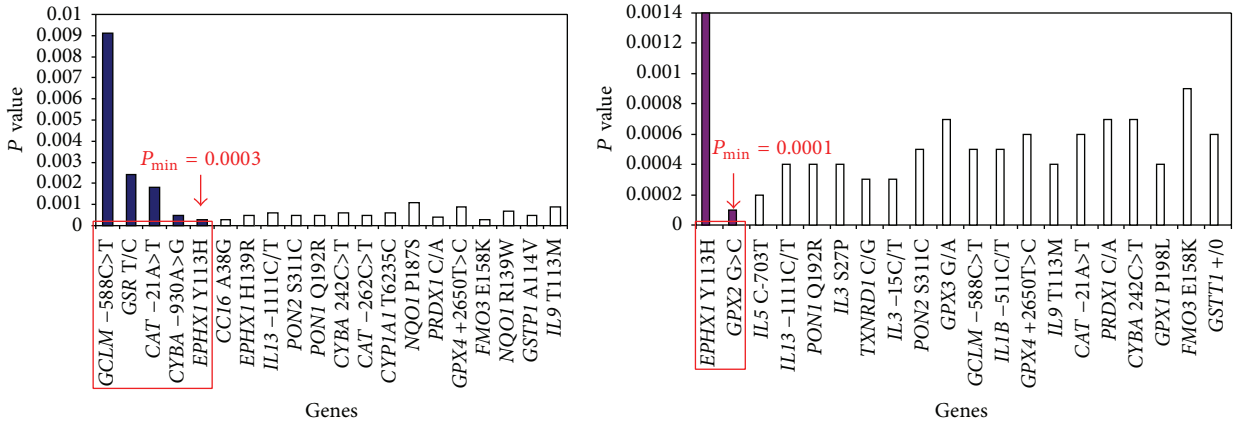


FIGURE 2: The results of statistical modeling of gene-gene interactions in nonallergic asthma using set association approach. Significance level of S_n statistic as a function of the number n of SNPs in different genes which are included at each step for gene-gene interactions analysis. The smallest significance level, P_{min}, occurs with 5 SNPs in males and with 2 SNPs in females. The interacting genes in the models are circled in red.

significantly associated with allergic and nonallergic asthma in both sexes. A majority of the susceptibility genes identified in our study represented antioxidant defense enzymes. Moreover, interactions between ADE genes varied across the pathogenetic variants of asthma and were different in men and women suggesting both genetic heterogeneity and gender-specific genetic effects in the disease susceptibility.

4.2. Genetic Heterogeneity of Asthma and Complexity of Genomic Interactions Underlying the Disease. The observed differences in gene-gene interactions between allergic and nonallergic variants of asthma demonstrate a genetic heterogeneity of the disease, a situation in which the same or similar phenotype of a complex disorder is caused by different susceptibility genes [57]. It is well known that genetic heterogeneity is the general feature of many common diseases [58] and may be explained at least partially by genetic differences between human populations [9]. Bronchial asthma is a typical example for complex multifactorial disease being characterized by genetic heterogeneity [59]. In fact, the models of gene-gene interactions in the pathogenetic variants of asthma overlap only partially, thereby reflecting, on the one hand, possible differences in the molecular mechanisms of allergic and nonallergic asthma and, on the other hand, the existence of shared genes that determine common susceptibility to the disease. In particular, three ADE genes such as *GSR*, *EPHX1,* and GPX1 showed significant interaction in both variants of asthma in both men and women (except for the *GPX1* gene in nonallergic asthma in women); therefore, they can be

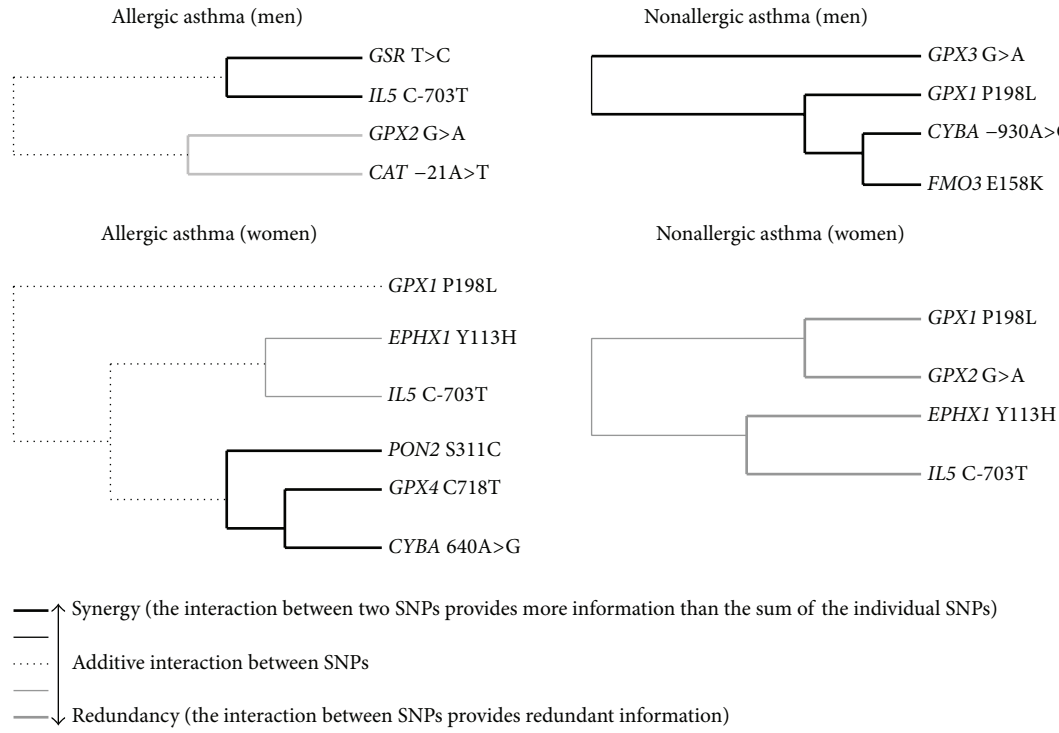


Figure 3: Dendrograms of gene-gene interactions in the pathogenetic variants of asthma (MDR method). Dendrograms show both complexity and diversity of interactions between polymorphic genes of antioxidant defense enzymes in allergic and nonallergic asthma (dendrograms are stratified by gender). Each dendrogram comprises a spectrum of lines representing a continuum from synergy (black) to redundancy (gray) of gene-gene interactions. The lines range from bold black, representing a high degree of synergy (positive information gain), thin black, representing a lesser degree, and dotted line representing the midway point between synergy and redundancy. On the redundancy end of the spectrum, the highest degree is represented by bold gray (negative information gain) with a lesser degree represented by thin gray.

Table 6: Associations of genotype combinations with risk of nonallergic asthma (stratified by gender).

Combinations of genotypes	Nonallergic asthma		Controls		Chi-square (P value[1])	OR (95% CI)
	N	%	N	%		
			Men			
GPX1 198PL × *CAT* −21AA	6	20.7	2	1.9	11.13 (0.001)*	11.45 (2.49–52.66)
GPX1 198LL × *GCLM* −588CT	4	13.8	3	2.9	3.50 (0.05)	5.17 (1.20–22.31)
GPX3 GA × *FMO3* 158KK	5	17.2	4	3.8	4.58 (0.03)	5.06 (1.37–18.99)
GPX3 GA × *GSR* TT	6	20.7	5	4.8	5.68 (0.02)	5.05 (1.49–17.14)
GPX3 GA × *CAT* −21AA	5	17.2	4	3.8	4.58 (0.03)	5.06 (1.35–18.99)
GPX3 GG × *GCLM* −588CT	11	37.9	15	14.3	8.12 (0.004)	3.67 (1.47–9.28)
GSR TT × *GCLM* −588CT	8	27.6	3	2.9	15.31 (0.0001)*	11.58 (3.07–43.72)
GSR TT × *FMO3* 158KK	4	13.8	1	1.0	7.16 (0.01)	12.29 (1.84–82.03)
CAT −21AA × *CYBA* −930GG	5	17.2	1	1.0	10.55 (0.001)*	15.64 (2.44–100.3)
CAT −21AA × *FMO3* 158EE	4	13.8	1	1.0	7.16 (0.01)	12.29 (1.84–82.03)
GCLM −588CT × *CYBA* −930GG	12	41.4	10	9.5	16.8 (0.00004)*	6.71 (2.5–17.96)
CYBA −930GG × *FMO3* 158EE	7	24.1	5	4.8	8.22 (0.004)	6.09 (1.85–20.05)
			Women			
GPX1 198PL × *GPX2* GA	2	7.4	0	0.0	3.88 (0.05)	21.47 (1.00–461.1)
GPX2 GG × *IL5* −703CT	6	22.2	56	51.4	6.29 (0.01)	0.29 (0.11–0.74)
GPX2 GA × *IL5* −703CC	2	7.4	0	0.0	3.88 (0.05)	21.5 (1.00–461.2)
EPHX1 113YH × *IL5* −703CT	1	3.7	24	22.0	3.69 (0.05)	0.20 (0.04–1.09)
EPHX1 113HH × *IL5* −703CC	7	25.9	4	3.7	11.58 (0.001)*	8.58 (2.43–30.26)

[1] *Means* unadjusted P value. P value adjusted for multiple tests (P_{adj}) is less than 0.002 in men and 0.004 in women. * A statistically significant association.

considered as common susceptibility genes to asthma. While the *IL5* and *PON2* genes showed an association only with allergic asthma, none of the studied genetic polymorphisms was found to be associated exclusively with the risk of nonallergic asthma.

The results of gene-gene interactions analysis are consistent with observations of other genetic studies which demonstrated an importance of ADE genes for asthma pathogenesis. In particular, we confirmed a potential role in the pathogenesis of asthma for *CYBA* and *CAT* genes that was associated with asthma in Czech [60] and Canadian [61] populations, respectively. In addition, the associations of asthma with the *IL5* C-703T polymorphism in Russians from the city of Tomsk [35] and *IL1B* −511C>T in the Canadian Asthma Primary Prevention Study [62] have been successfully replicated in our study. Our study is consistent with the observation that glutathione-S-transferase genes M1, T1, and P1 alone and in combination with other ADE genes do not play a substantial role in the development of BA [63]. We also found for the first time genetic polymorphisms of the *GSR* and *PON2* genes can be important determinants of susceptibility to asthma, but their associations need to be confirmed in independent populations. Further studies should also be focused on the analysis of gene-gene interactions to better understanding the role of ADE genes in asthma pathogenesis. In the study of Millstein et al. [16], interactions between the *NQO1*, *MPO*, and *CAT* genes have been identified in ethnically diverse cohorts of patients with childhood asthma, whereas marker-by-marker analysis did not reveal the associations of these genes with disease susceptibility. This means that marker-by-marker approach ignores the multigenic nature of BA and does not evaluate a complexity of interactions between susceptibility genes.

Comparing the results obtained by the three statistical approaches to the analysis of gene-gene interactions, we can say that, despite gender-specific effects of genotypes on the pathogenetic variants of the disease, each of the methods showed own uniqueness and efficacy in the detecting genes associated with asthma risk. In our point of view, the advantage of SAA method is its capacity in the identification of "gene dosage effects" of different sets of ADE genes on asthmatic phenotype. Meanwhile, MDR method, especially its cluster technique, was found to be powerful in the detecting high-order epistatic interactions between ADE genes and their synergic and antagonistic effects on the asthma risk. The variability in the structure of gene-gene interactions models across the pathogenetic variants of asthma can be partially explained by differences in bioinformatic approaches to the analysis of multiple genes. Apparently, a similarity in gene-gene interactions between the models obtained by the two different bioinformatic tools may be explained by strong effects of particular genes on the asthmatic phenotype. This means that strong phenotypic effects of the *CAT*, *GPX1*, *GSR*, *GCLM*, *EPHX1*, *CYBA,* and *IL5* genes (they showed the similarity between the models) may be considered as major gene effects. And, finally, a post hoc comparative analysis of the frequencies of genotype combinations was useful in the detection of ADE genes with low or moderate effects on asthma as well as the unique combinations of ADE genotypes strongly associated with disease susceptibility. In particular, the *PON2*, *GPX2*, *GPX3*, *GPX4*, *NQO1*, *FMO3*, *SOD2,* and *IL3* genes showed low or moderate effects on asthma risk and may represent a polygenic background of the disease susceptibility. Thus, the methods complemented each other and contributed to the understanding of the polygenic nature of asthma and complexity of gene-gene interactions underlying the asthmatic phenotypes. Due to the fact that there is no universal method for comprehensive analysis of gene-gene interactions in genomic epidemiology, it makes sense to use several methods, as it has been successfully applied in the present study.

The dendrograms obtained by MDR technique (Figure 3) clearly showed complex and hierarchic pattern of interactions between ADE genes constituting the polygenic basis of the pathogenetic variants of asthma. In particular, the *GSR* T>C and *IL5* C-703T genes in men and the *CYBA* 640A>G, *PON2* C311S, and *GPX4* C718T loci in women had the highest degree of synergy in their interactions to determine the susceptibility to allergic asthma, whereas the highest degree of synergy in gene-gene interactions in nonallergic asthma in men was found for the *CYBA* −930A>G, *FMO3* E158K, and *GPX1* P198L loci. In contrast, a different degree of redundancy (antagonism) in gene-gene interactions was observed between the *CAT* −21A>T and *GPX2* G>A loci in men and between the *EPHX1* Y113H and *IL5* loci C-703T in women with allergic asthma, as well as between the *GPX1* P198L, *GPX2* G>A, *EPHX1* Y113H, and *IL5* C-703T loci in women with nonallergic asthma. Interestingly, the *EPHX1* Y113H and *IL5* C-703T genes showed an antagonistic character of gene-gene interactions exclusively in women with both pathogenetic variants of asthma. Notably, the *EPHX1* Y113H genotypes did show the association with asthma risk in single-locus analysis performed in our previous study ($P = 0.21$ $df = 2$) [19]. A strong synergism or antagonism in the interaction between the ADE genes in determining different types of asthma may suggest that the gene-gene effect can be driven by their true interaction, rather than by the main effect from the distinct gene. These findings may indicate epistatic interactions of the ADE genes, a situation when the effect of one gene may not be disclosed if the effect of another gene is not considered [64].

A post hoc comparative analysis of the frequencies of genotype combinations in the study groups revealed two-locus combinations of the ADE genotypes which increase the risk of the development of asthma. We found relatively rare combinations of genotypes which gave the highest asthma risk estimates but were limited to small subgroups of subjects. In particular, frequencies of these genotype combinations varied from 1 to 9% among healthy controls and from 17 to 41% among patients with nonallergic asthma, whereas odds ratios for disease risk varied from 6.7 to 15.6. Moreover, there was an obvious excess of combinations of variant genotypes among asthma patients compared with healthy subjects, and these differences reached statistical significance after adjusting for multiple tests.

4.3. Genetic Variation in ADE Genes and Asthma Pathogenesis. Despite nonsignificant differences in the genotype

distributions, the two-locus comparison of genotype frequencies between the study groups has shown that asthma patients more often than healthy subjects carry combinations of the genotypes which are known to determine a diminished activity of ADE towards ROS. This is supported by a number biochemical studies that observed massive generation of ROS and the insufficiency in antioxidant capacity in asthma [65–68]. In this context, it is important to highlight that the studied ADE genes alone cannot account for the whole polygenic mechanisms underlying such biochemical abnormalities in asthma. The interactions between ADE genes that we have identified using different statistical methods make a mechanistic sense because these genes are collectively involved in the maintaining and regulation of redox homeostasis. Moreover, the integrated function of ADE genes in the lung and airways can promote a coordinated detoxification of xenobiotics-induced ROS, thus preventing oxidative stress which plays an important role in the pathogenesis of asthma [21, 24, 51, 69]. Importantly, ADE genes showed interactions with other asthma-related genes such as *IL5* and *IL1B* which are responsible for the immunological mechanisms of asthma and allergy.

Based on the literature data demonstrating biochemical abnormalities in redox homeostasis in asthma and the results of our study, we assumed possible relationships between these abnormalities and ADE genes showed the associations with asthma in our study (the data are shown in Table 7). Changes in the activity of antioxidant defense enzymes such as glutathione peroxidases and catalase in whole blood, plasma, platelets, and bronchoalveolar lavage fluid have been reported by a number of biochemical studies, the findings which are in accordance with our results demonstrating the relationship between the genes for these enzymes and the risk of different pathogenetic variants of asthma. Briefly, an enhanced production of ROS by blood neutrophils, monocytes, and eosinophils found in asthma can be explained by the effects of functional polymorphisms in the gene encoding p22 phox subunit (*CYBA*) of NADPH oxidase. Genetic variation in the *GSR* and *GCLM* genes may be responsible for biochemical perturbations of glutathione metabolism such as an increased level of oxidized glutathione in asthma. Polymorphisms of the *EPHX1* gene determine the increased activity of the enzyme, thus leading to the enhanced production of reactive semiquinones.

Although we did not perform biochemical investigations of antioxidant status, taking the observed association of asthma with the ADE genotypes and their functional significance into account, it is likely that an imbalance between oxidants and antioxidants detected in asthma can be directly related to genetically diminished capacity of ADE. Such an imbalance results in oxidative stress caused by an excessive production of ROS and/or by inadequate antioxidant defense leading to damage of airway epithelial cells and inflammation due to upregulation of redox-sensitive transcription factors and proinflammatory genes [22, 24]. We may also conclude that ADE genes seem to play a greater role in the development of nonallergic asthma than in allergic asthma.

4.4. Gender-Specific Effects of ADE Genes on Susceptibility to Bronchial Asthma. An important finding of our study was that polymorphisms of many ADE genes showed sex-specific associations with the development of asthma. For instance, the *CAT* −21A>T and *GCLM* −588C>T gene polymorphisms were associated with asthma susceptibility exclusively in men. In contrast, polymorphism 640A>G of the *CYBA* gene showed a relationship with asthma risk only in women. These findings demonstrate sexual dimorphism in genetic susceptibility to asthma, a phenomenon established for many complex human diseases [70, 71] including asthma [72]. It has been proposed that existing variation in regulatory elements of genes rather than differences in their structure in men and women may explain sex-specific genotype-phenotype interactions in complex traits [70, 73]. Sex-specific changes in age-related gene regulation can result in the difference in asthma susceptibility between the sexes [70]. We suggest that the mechanisms underlying gender-related specificity in the associations of ADE genes with BA found in our study are related to differential expression of redox-sensitive genes in men and women. Since estrogen was found to depress oxidative stress in mice [74], sex-steroid receptors might be an example of sex-specific *trans*-regulatory elements [75] for redox-sensitive genes which in turn may differentially respond to the inducers due to their functionally unequal polymorphic alleles. This means that ADE genes can function differently in men and women in some circumstances. This suggestion is supported by the finding of gender difference in both expression and activity of antioxidant enzymes demonstrated in animal studies [76, 77]. Therefore, it can be concluded that identification of gender-specific genetic variants of ADE, which contribute to the shift of redox homeostasis towards oxidative stress, will provide a better understanding of sex-specific regulation of ADE gene expression and differences in the molecular mechanisms of asthma in men and women.

4.5. Limitations of the Study. The study has limitations. Due to the relatively small sample sizes of the studied groups, the association analysis of two-locus genotype combinations was underpowered, especially after Bonferroni adjustment for multiple tests. Because of the limited sample size, we also cannot exclude the possibility that small effects of some ADE genes were not detected. Since BA is a multifactorial and genetically heterogeneous disease [78, 79], further studies with larger sample sizes with genotyping of more polymorphic variants of ADE genes are required for better understanding of the roles of these genes in asthma pathogenesis. Because we did not analyze expression profiles of the genes and biochemical parameters of redox homeostasis, both functional genomics and metabolomics studies are required to clarify the molecular mechanisms by which polymorphisms of ADE genes contribute to the development of BA. Since the risk of BA is determined by a complex interplay between genetic and environmental factors, further genetic studies should take into account environmental factors that may play a significant role in the etiology of the disease.

Table 7: Common biochemical abnormalities in redox homeostasis found in asthma and their possible relationship with genes for antioxidant defense enzymes which have been associated with risk of the disease in the present study.

Biochemical abnormalities in asthmatics [references]	ADE gene related with the abnormality	Allergic asthma		Nonallergic asthma	
		Men	Women	Men	Women
Diminished capacity of glutathione peroxidases and catalase in detoxification of hydrogen peroxide [22, 38–43].	*GPX1*	+ + +	+ +	+ +	+ +
	GPX2	+ +	−		+ + +
	GPX3	−	−	+ +	−
	GPX4	−	+	−	−
	CAT	+ + +	−	+ +	−
An enhanced production of ROS/hydrogen peroxide/superoxide anion radicals [22, 44–50].	*CYBA* (640A>G)	−	+	−	−
	CYBA (−930A>G)	−	−	+ + +	−
Perturbations in glutathione (GSH) homeostasis [41, 48, 50–52].	*GSR*	+ +	−	+ + +	−
	GCLM	+ +	−	+ + +	−
Increased EPHX1 activity, increased production of xenobiotics-generated epoxides, *trans*-dihydrodiols and reactive semiquinones resulting in ROS generation [53, 54].	*EPHX1*	+ +	+ + +	+ +	+ + +

The number of pluses means a degree of the relationship between the gene and asthma risk. These measures reflect how many times a particular gene showed the link with asthma risk through the three methods used for evaluation of gene-gene interactions in the present study, namely, set association approach (SAA), multifactor dimensionality reduction (MDR) method, and post hoc association analysis of two-locus genotype combinations (AAGC): + + + means that the link was found thrice (i.e., using SAA, MDR, and AAGC methods); + + means that the link was found twice (i.e., using SAA or MDR and AAGC methods); + means that the link was found once by AAGC method. Associations are stratified by asthma type and gender.

5. Conclusions

To the best of our knowledge, this is the first study investigating the associations between BA and 34 functionally significant polymorphic variants of ADE genes and 12 other candidate genes. So far, no genetic studies have reported a comprehensive evaluation of asthma susceptibility with a number of ADE genes at once. Methodological approaches used in this study were proved fruitful in uncovering the genetic architecture of complex interactions between genes involved in the regulation of redox homeostasis. This allowed finding for the first time that antioxidant defense enzymes genes are collectively involved in the molecular mechanisms of BA and can explain genetic heterogeneity between allergic and nonallergic variants of the disease. In particular, we found for the first time that the *GSR* and *PON2* genes can be referred to as novel asthma susceptibility genes, but their associations need to be confirmed in independent populations. We also showed both complexity and diversity of gene-gene interactions in allergic and nonallergic asthma. Finally, we have discovered gender-specific effects of ADE genes for the risk of the pathogenetic variants of asthma. Altogether the study results provide strong evidence for the pathogenetic role of ADE genes in asthma. Our data on the relationship of the ADE genes and asthma are concordant with the results of a number of biochemical studies demonstrating the massive generation of ROS and the insufficiency in antioxidant capacity which have been implicated in pathogenesis of asthma.

Further studies focusing on the molecular mechanisms regulating redox homeostasis can provide more complete understanding of the role of the ADE genes in bronchial asthma and end up in the discovery of new drug targets for antioxidant treatment and prevention of the disease.

Conflict of Interests

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

Acknowledgments

The study was supported in part by the Federal Targeted Program "Scientific and Scientific-Pedagogical Personnel of the Innovative Russia in 2009–2013." The authors thank Drs. Mikhail Kozhuhov and Valery Panfilov from Kursk Regional Clinical Hospital for their invaluable help in recruiting and examining asthma patients for the study.

References

[1] National Asthma Education and Prevention Program (NAEPP), *Expert Panel Report 3: Guidelines for the Diagnosis and Management of Asthma*, National Heart Lung and Blood Institute, Bethesda, Md, USA, 2007.

[2] C. Ober and T.-C. Yao, "The genetics of asthma and allergic disease: a 21st century perspective," *Immunological Reviews*, vol. 242, no. 1, pp. 10–30, 2011.

[3] Y. Zhang, M. F. Moffatt, and W. O. C. Cookson, "Genetic and genomic approaches to asthma: new insights for the origins," *Current Opinion in Pulmonary Medicine*, vol. 18, no. 1, pp. 6–13, 2012.

[4] M. F. Moffatt, I. G. Gut, F. Demenais et al., "A large-scale, consortium-based genomewide association study of asthma," *The New England Journal of Medicine*, vol. 363, no. 13, pp. 1211–1221, 2010.

[5] D. G. Torgerson, E. J. Ampleford, G. Y. Chiu et al., "Meta-analysis of genome-wide association studies of asthma in ethnically diverse North American populations," *Nature Genetics*, vol. 43, no. 9, pp. 887–892, 2011.

[6] T. A. Pearson and T. A. Manolio, "How to interpret a genome-wide association study," *Journal of the American Medical Association*, vol. 299, no. 11, pp. 1335–1344, 2008.

[7] E. Evangelou and J. P. Ioannidis, "Meta-analysis methods for genome-wide association studies and beyond," *Nature Reviews Genetics*, vol. 14, no. 6, pp. 379–389, 2013.

[8] E. E. Eichler, J. Flint, G. Gibson et al., "Missing heritability and strategies for finding the underlying causes of complex disease," *Nature Reviews Genetics*, vol. 11, no. 6, pp. 446–450, 2010.

[9] S. A. Tishkoff and B. C. Verrelli, "Patterns of human genetic diversity: implications for human evolutionary history and disease," *Annual Review of Genomics and Human Genetics*, vol. 4, pp. 293–340, 2003.

[10] H. Y. Kim, R. H. Dekruyff, and D. T. Umetsu, "The many paths to asthma: phenotype shaped by innate and adaptive immunity," *Nature Immunology*, vol. 11, no. 7, pp. 577–584, 2010.

[11] D. T. Swarr and H. Hakonarson, "Unraveling the complex genetic underpinnings of asthma and allergic disorders," *Current Opinion in Allergy and Clinical Immunology*, vol. 10, no. 5, pp. 434–442, 2010.

[12] N. C. Battle, S. Choudhry, H.-J. Tsai et al., "Ethnicity-specific gene-gene interaction between IL-13 and IL-4Rα among African Americans with asthma," *American Journal of Respiratory and Critical Care Medicine*, vol. 175, no. 9, pp. 881–887, 2007.

[13] H. J. Cordell, "Detecting gene-gene interactions that underlie human diseases," *Nature Reviews Genetics*, vol. 10, no. 6, pp. 392–404, 2009.

[14] J. Hoh, A. Wille, and J. Ott, "Trimming, weighting, and grouping SNPs in human case-control association studies," *Genome Research*, vol. 11, no. 12, pp. 2115–2119, 2001.

[15] M. D. Ritchie, L. W. Hahn, N. Roodi et al., "Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer," *American Journal of Human Genetics*, vol. 69, no. 1, pp. 138–147, 2001.

[16] J. Millstein, D. V. Conti, F. D. Gilliland, and W. J. Gauderman, "A testing framework for identifying susceptibility genes in the presence of epistasis," *American Journal of Human Genetics*, vol. 78, no. 1, pp. 15–27, 2006.

[17] A. A. Motsinger and M. D. Ritchie, "Multifactor dimensionality reduction: an analysis strategy for modelling and detecting gene-gene interactions in human genetics and pharmacogenomics studies," *Human Genomics*, vol. 2, no. 5, pp. 318–328, 2006.

[18] T. J. Jorgensen, I. Ruczinski, B. Kessing, M. W. Smith, Y. Y. Shugart, and A. J. Alberg, "Hypothesis-driven candidate gene association studies: practical design and analytical considerations," *American Journal of Epidemiology*, vol. 170, no. 8, pp. 986–993, 2009.

[19] A. V. Polonikov, V. P. Ivanov, and M. A. Solodilova, "Genetic variation of genes for xenobiotic-metabolizing enzymes and

risk of bronchial asthma: the importance of gene-gene and gene-environment interactions for disease susceptibility," *Journal of Human Genetics*, vol. 54, no. 8, pp. 440–449, 2009.

[20] S. J. London, "Gene-air pollution interactions in asthma," *Proceedings of the American Thoracic Society*, vol. 4, no. 3, pp. 217–220, 2007.

[21] F. Castro-Giner, N. Künzli, B. Jacquemin et al., "Traffic-related air pollution, oxidative stress genes, and asthma (ECHRS)," *Environmental Health Perspectives*, vol. 117, no. 12, pp. 1919–1924, 2009.

[22] J. C. Mak and M. M. Chan-Yeung, "Reactive oxidant species in asthma," *Current Opinion in Pulmonary Medicine*, vol. 12, no. 1, pp. 7–11, 2006.

[23] H. Sugiura and M. Ichinose, "Oxidative and nitrative stress in bronchial asthma," *Antioxidants and Redox Signaling*, vol. 10, no. 4, pp. 785–797, 2008.

[24] M. A. Riedl and A. E. Nel, "Importance of oxidative stress in the pathogenesis and treatment of asthma," *Current Opinion in Allergy and Clinical Immunology*, vol. 8, no. 1, pp. 49–56, 2008.

[25] A. V. Polonikov, V. P. Ivanov, M. A. Solodilova, I. V. Khoroshaya, M. A. Kozhuhov, and V. I. Panfilov, "The relationship between polymorphisms in the glutamate cysteine ligase gene and asthma susceptibility," *Respiratory Medicine*, vol. 101, no. 11, pp. 2422–2424, 2007.

[26] M. A. Solodilova, V. P. Ivanov, A. V. Polonikov, I. V. Khoroshaya, M. A. Kozhukhov, and V. I. Panfilov, "Heterozygosity of 198LEU mutant allele in glutathione peroxidase-1 gene as a risk factor of bronchial asthma associated with smoking," *Terapevticheskii Arkhiv*, vol. 79, no. 3, pp. 33–36, 2007.

[27] A. V. Polonikov, V. P. Ivanov, M. A. Solodilova, M. A. Kozhuhov, and V. I. Panfilov, "Tobacco smoking, fruit and vegetable intake modify association between −21A > T polymorphism of catalase gene and risk of bronchial asthma," *Journal of Asthma*, vol. 46, no. 3, pp. 217–224, 2009.

[28] A. V. Polonikov, M. A. Solodilova, and V. P. Ivanov, "Genetic variation of myeloperoxidase gene contributes to atopic asthma susceptibility: a preliminary association study in russian population," *Journal of Asthma*, vol. 46, no. 5, pp. 523–528, 2009.

[29] V. P. Ivanov, M. A. Solodilova, A. V. Polonikov, I. V. Khoroshaia, M. A. Kozhukhov, and V. I. Panfilov, "Association of C242T and A640G polymorphisms in the gene for p22phox subunit of NADPH oxidase with the risk of bronchial asthma: a pilot study," *Genetika*, vol. 44, no. 5, pp. 693–701, 2008.

[30] A. V. Polonikov, V. P. Ivanov, M. A. Solodilova, M. A. Kozhukhov, V. I. Panfilov, and I. V. Bulgakova, "Polymorphism −930A > G of the cytochrome b gene is a novel genetic marker of predisposition to bronchial asthma," *Terapevticheskiĭ Arkhiv*, vol. 81, no. 3, pp. 31–35, 2009.

[31] A. V. Polonikov, M. A. Solodilova, V. P. Ivanov et al., "Association study of the role of polymorphic variants of gene of NAD(P)H: quinone oxidoreductase type 1 in predisposition to bronchial asthma in population of Russians from the Central Chernozem region," *Journal of Medical Genetics (Moscow)*, vol. 10, no. 4, pp. 23–27, 2011.

[32] A. V. Polonikov, V. P. Ivanov, M. A. Solodilova, I. V. Khoroshaya, M. A. Kozhuhov, and V. I. Panfilov, "Promoter polymorphism G-50T of a human CYP2J2 epoxygenase gene is associated with common susceptibility to asthma," *Chest*, vol. 132, no. 1, pp. 120–126, 2007.

[33] D. N. Cooper, R. L. Nussbaum, and M. Krawczak, "Proposed guidelines for papers describing DNA polymorphism-disease associations," *Human Genetics*, vol. 110, no. 3, pp. 207–208, 2002.

[34] M. B. Freidin, V. P. Puzyrev, L. M. L. Ogorodova et al., "Analysis of the association between the T113M polymorphism of the human interleukin 9 gene and bronchial asthma," *Genetika*, vol. 36, no. 4, pp. 559–561, 2000.

[35] M. B. Freidin, V. P. Puzyrev, L. M. Ogorodova, O. S. Kobvakova, and I. M. Kulmanakova, "Polymorphism of the interleukin- and interleukin receptor genes: population distribution and association with atopic asthma," *Genetika*, vol. 38, no. 12, pp. 1710–1718, 2002.

[36] M. B. Freidin, O. S. Kobyakova, L. M. Ogorodova, and V. P. Puzyrev, "Association of polymorphisms in the human IL4 and IL5 genes with atopic bronchial asthma and severity of the disease," *Comparative and Functional Genomics*, vol. 4, no. 3, pp. 346–350, 2003.

[37] A. V. Polonikov, V. P. Ivanov, D. A. Belugin, I. V. Khoroshaya, M. A. Solodilova, and V. I. Panfilov, "Studying associations of two common S27P and −15? > T polymorphisms of the interleukin-3 gene with development of allergic and non-allergic bronchial asthma in Russians from the Central-Chernozem region," *Medical Immunology (Saint Petersburg)*, vol. 8, no. 5-6, pp. 731–736, 2006.

[38] L. Hasselmark, R. Malmgren, G. Unge, and O. Zetterstrom, "Lowered platelet glutathione peroxidase activity in patients with intrinsic asthma," *Allergy*, vol. 45, no. 7, pp. 523–527, 1990.

[39] Z. Novák, I. Németh, K. Gyurkovits, S. I. Varga, and B. Matkovics, "Examination of the role of oxygen free radicals in bronchial asthma in childhood," *Clinica Chimica Acta*, vol. 201, no. 3, pp. 247–251, 1991.

[40] C. V. Powell, A. A. Nash, H. J. Powers, and R. A. Primhak, "Antioxidant status in asthma," *Pediatric pulmonology*, vol. 18, no. 1, pp. 34–38, 1994.

[41] A. Nadeem, S. K. Chhabra, A. Masood, and H. G. Raj, "Increased oxidative stress and altered levels of antioxidants in asthma," *Journal of Allergy and Clinical Immunology*, vol. 111, no. 1, pp. 72–78, 2003.

[42] B. Varshavskiĭ, G. V. Trubnikov, L. P. Galaktionova, N. A. Korenyak, I. L. Kolodeznaya, and A. N. Oberemok, "The oxidative-antioxidative status in patients with bronchial asthma in inhaled and oral glucocorticoid therapy," *Terapevticheskii Arkhiv*, vol. 75, no. 3, pp. 21–24, 2003.

[43] L. Bentur, Y. Mansour, R. Brik, Y. Eizenberg, and R. M. Nagler, "Salivary oxidative stress in children during acute asthmatic attack and during remission," *Respiratory Medicine*, vol. 100, no. 7, pp. 1195–1201, 2006.

[44] P. Chanez, G. Dent, T. Yukawa, P. J. Barnes, and K. F. Chung, "Generation of oxygen free radicals from blood eosinophils from asthma patients after stimulation with PAF or phorbol ester," *European Respiratory Journal*, vol. 3, no. 9, pp. 1002–1007, 1990.

[45] I. Vachier, M. Damon, C. le Doucen et al., "Increased oxygen species generation in blood monocytes of asthmatic patients," *American Review of Respiratory Disease*, vol. 146, no. 5, pp. 1161–1166, 1992.

[46] I. Rahman, D. Morrison, K. Donaldson, and W. Macnee, "Systemic oxidative stress in asthma, COPD, and smokers," *American Journal of Respiratory and Critical Care Medicine*, vol. 154, no. 4, pp. 1055–1060, 1996.

[47] K. R. Shanmugasundaram, S. S. Kumar, and S. Rajajee, "Excessive free radical generation in the blood of children suffering from asthma," *Clinica Chimica Acta*, vol. 305, no. 1-2, pp. 107–114, 2001.

[48] J. Beier, K. M. Beeh, D. Semmler, N. Beike, and R. Buhl, "Increased concentrations of glutathione in induced sputum of patients with mild or moderate allergic asthma," *Annals of Allergy, Asthma and Immunology*, vol. 92, no. 4, pp. 459–463, 2004.

[49] T. Fujisawa, "Role of oxygen radicals on bronchial asthma," *Current Drug Targets: Inflammation and Allergy*, vol. 4, no. 4, pp. 505–509, 2005.

[50] N. L. Reynaert, "Glutathione biochemistry in asthma," *Biochimica et Biophysica Acta*, vol. 1810, no. 11, pp. 1045–1051, 2011.

[51] F. J. Kelly and T. Sandström, "Air pollution, oxidative stress, and allergic response," *The Lancet*, vol. 363, no. 9403, pp. 95–96, 2004.

[52] S. A. A. Comhair, P. R. Bhathena, R. A. Dweik, M. Kavuru, and S. C. Erzurum, "Rapid loss of superoxide dismutase activity during antigen-induced asthmatic response," *The Lancet*, vol. 355, no. 9204, article 624, 2000.

[53] M. T. Salam, P.-C. Lin, E. L. Avol, W. J. Gauderman, and F. D. Gilliland, "Microsomal epoxide hydrolase, glutathione S-transferase P1, traffic and childhood asthma," *Thorax*, vol. 62, no. 12, pp. 1050–1057, 2007.

[54] K.-Y. Tung, C.-H. Tsai, and Y. L. Lee, "Microsomal epoxide hydroxylase genotypes/diplotypes, traffic air pollution, and childhood asthma," *Chest*, vol. 139, no. 4, pp. 839–848, 2011.

[55] J. Hoh and J. Ott, "Mathematical multi-locus approaches to localizing complex human trait genes," *Nature Reviews Genetics*, vol. 4, no. 9, pp. 701–709, 2003.

[56] D. M. Nielsen, M. G. Ehm, and B. S. Weir, "Detecting marker-disease association by testing for Hardy-Weinberg disequilibrium at a marker locus," *American Journal of Human Genetics*, vol. 63, no. 5, pp. 1531–1540, 1998.

[57] J. L. Davies, Y. Kawaguchl, S. T. Bennett et al., "A genome-wide search for human type 1 diabetes susceptibility genes," *Nature*, vol. 371, no. 6493, pp. 130–136, 1994.

[58] J. McClellan and M.-C. King, "Genetic heterogeneity in human disease," *Cell*, vol. 141, no. 2, pp. 210–217, 2010.

[59] F. D. Martinez, "Complexities of the genetics of asthma," *American Journal of Respiratory and Critical Care Medicine*, vol. 156, no. 4, part 2, pp. S117–S122, 1997.

[60] L. I. Holla, K. Kaňková, and V. Znojil, "Haplotype analysis of the NADPH oxidase p22 gene in patients with bronchial asthma," *International Archives of Allergy and Immunology*, vol. 148, no. 1, pp. 73–80, 2009.

[61] A. Morin, J. R. Brook, C. Duchaine, and C. Laprise, "Association study of genes associated to asthma in a specific environment, in an asthma familial collection located in a rural area influenced by different industries," *International Journal of Environmental Research and Public Health*, vol. 9, no. 8, pp. 2620–2635, 2012.

[62] D. Daley, M. Lemire, L. Akhabir et al., "Analyses of associations with asthma in four asthma population samples from Canada and Australia," *Human Genetics*, vol. 125, no. 4, pp. 445–459, 2009.

[63] C. Minelli, R. Granell, R. Newson et al., "Glutathione-S-transferase genes and asthma phenotypes: a Human Genome Epidemiology (HuGE) systematic review and meta-analysis including unpublished data," *International Journal of Epidemiology*, vol. 39, no. 2, Article ID dyp337, pp. 539–562, 2010.

[64] J. H. Moore, J. C. Gilbert, C.-T. Tsai et al., "A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility," *Journal of Theoretical Biology*, vol. 241, no. 2, pp. 252–261, 2006.

[65] I. Rahman, S. K. Biswas, and A. Kode, "Oxidant and antioxidant balance in the airways and airway diseases," *European Journal of Pharmacology*, vol. 533, no. 1–3, pp. 222–239, 2006.

[66] A. Nadeem, A. Masood, and N. Siddiqui, "Oxidant—antioxidant imbalance in asthma: scientific evidence, epidemiological data and possible therapeutic options," *Therapeutic Advances in Respiratory Disease*, vol. 2, no. 4, pp. 215–235, 2008.

[67] B. Gaston, "The biochemistry of asthma," *Biochimica et Biophysica Acta*, vol. 1810, no. 11, pp. 1017–1024, 2011.

[68] E. Babusikova, J. Jurecekova, A. Evinova, M. Jesenak, and D. Dobrota, "Oxidative damage and bronchial asthma in respiratory diseases," in *Respiratory Diseases*, M. Ghanei, Ed., InTech, Rijeka, Croatia, 2012.

[69] N. Li, M. Hao, R. F. Phalen, W. C. Hinds, and A. E. Nel, "Particulate air pollutants and asthma: a paradigm for the role of oxidative stress in PM-induced adverse health effects," *Clinical Immunology*, vol. 109, no. 3, pp. 250–265, 2003.

[70] C. Ober, D. A. Loisel, and Y. Gilad, "Sex-specific genetic architecture of human disease," *Nature Reviews Genetics*, vol. 9, no. 12, pp. 911–922, 2008.

[71] K. Mittelstrass, J. S. Ried, Z. Yu et al., "Discovery of sexual dimorphisms in metabolic and genetic biomarkers," *PLoS Genetics*, vol. 7, no. 8, Article ID e1002215, 2011.

[72] D. S. Postma, "Gender differences in asthma development and progression," *Gender Medicine*, vol. 4, supplement 2, pp. S133–S146, 2007.

[73] L. A. Weiss, L. Pan, M. Abney, and C. Ober, "The sex-specific genetic architecture of quantitative traits in humans," *Nature Genetics*, vol. 38, no. 2, pp. 218–222, 2006.

[74] K. M. Egan, J. A. Lawson, S. Fries et al., "COX-2-derived prostacyclin confers atheroprotection on female mice," *Science*, vol. 306, no. 5703, pp. 1954–1957, 2004.

[75] R. Angelopoulou, G. Lavranos, and P. Manolakou, "Establishing sexual dimorphism in humans," *Collegium Antropologicum*, vol. 30, no. 3, pp. 653–658, 2006.

[76] C. Borrás, J. Sastre, D. García-Sala, A. Lloret, F. V. Pallardó, and J. Viña, "Mitochondria from females exhibit higher antioxidant gene expression and lower oxidative damage than males," *Free Radical Biology and Medicine*, vol. 34, no. 5, pp. 546–552, 2003.

[77] J. Vina, C. Borras, M.-C. Gomez-Cabrera, and W. C. Orr, "Part of the series: from dietary antioxidants to regulators in cellular signalling and gene expression. Role of reactive oxygen species and (phyto)oestrogens in the modulation of adaptive response to stress," *Free Radical Research*, vol. 40, no. 2, pp. 111–119, 2006.

[78] T. L. Bonfield and K. R. Ross, "Asthma heterogeneity and therapeutic options from the clinic to the bench," *Current Opinion in Allergy and Clinical Immunology*, vol. 12, no. 1, pp. 60–67, 2012.

[79] R. A. Mathias, "Introduction to genetics and genomics in asthma: genetics of asthma," in *Heterogeneity in Asthma*, pp. 125–155, Springer, New York, NY, USA, 2014.