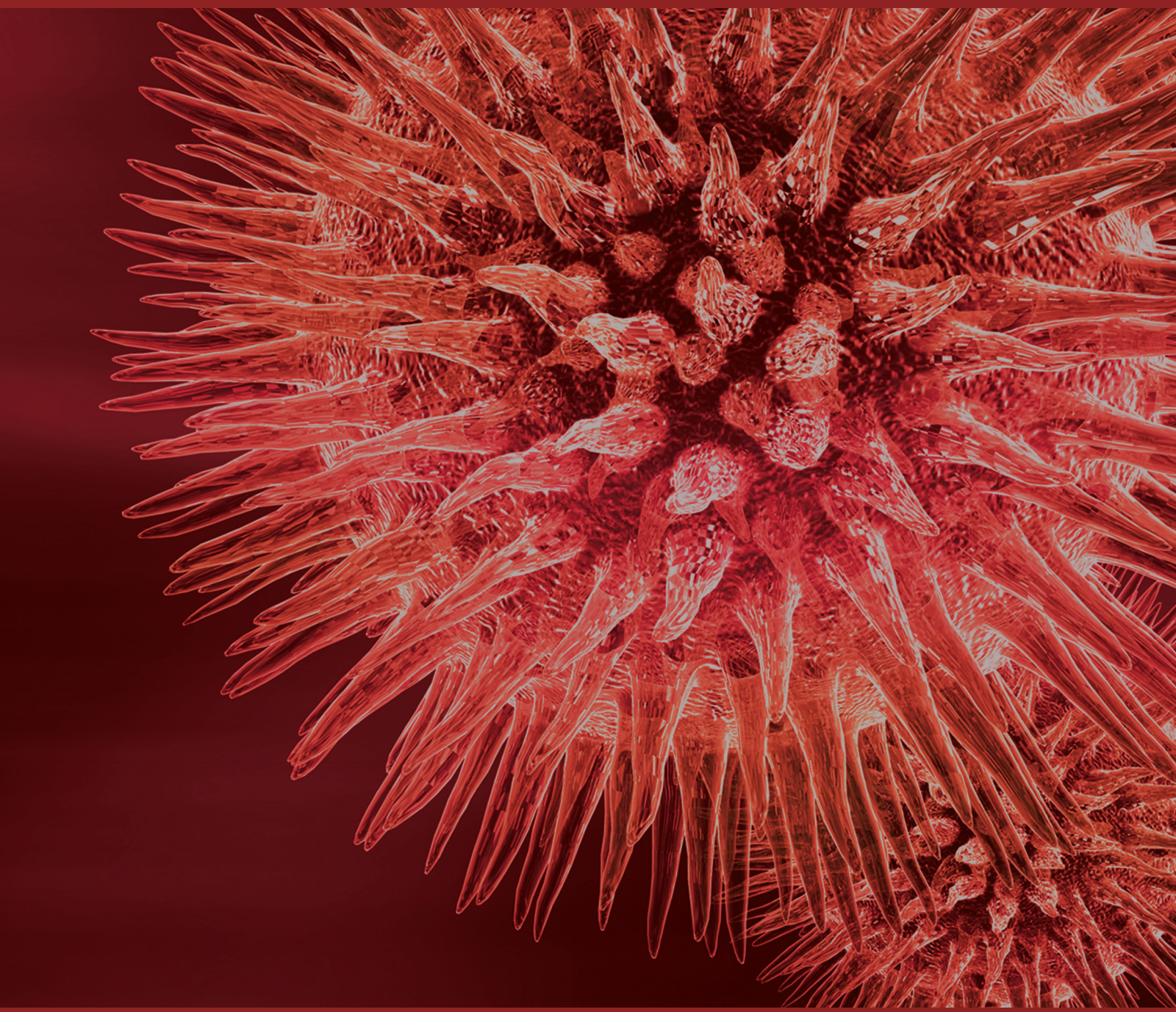
BioMed Research International

Functional Genomics, Genetics, and Bioinformatics 2016

Guest Editors: Youping Deng, Hongwei Wang, Ryuji Hamamoto, Shiwei Duan, Mehdi Pirooznia, and Yongsheng Bai



Functional Genomics, Genetics, and Bioinformatics 2016

BioMed Research International

Functional Genomics, Genetics, and Bioinformatics 2016

Guest Editors: Youping Deng, Hongwei Wang, Ryuji Hamamoto, Shiwei Duan, Mehdi Pirooznia, and Yongsheng Bai

Contents

Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 2625831, 3 pages
http://dx.doi.org/10.1155/2016/2625831

Editorial

Functional Genomics, Genetics, and Bioinformatics 2016

Youping Deng,[1] Hongwei Wang,[2] Ryuji Hamamoto,[3] Shiwei Duan,[4] Mehdi Pirooznia,[5] and Yongsheng Bai[6]

[1]*Bioinformatics Core, Office of Biostatistics & Quantitative Health Sciences, Department of Tropical Medicine, Medical Microbiology, and Pharmacology, University of Hawaii John A. Burns School of Medicine, Honolulu, HI 96813, USA*
[2]*Department of Medicine, University of Chicago, Chicago, IL 60637, USA*
[3]*Division of Molecular Modification and Cancer Biology, National Cancer Center Research Institute, Tokyo, Japan*
[4]*School of Medicine, Ningbo University, Ningbo, Zhejiang 315211, China*
[5]*Bioinformatics and Computational Biology Core Facility, National Heart, Lung, and Blood Institute, Office of the Scientific Director, National Institutes of Health, Bethesda, MD 20814, USA*
[6]*Department of Biology, Indiana State University, Terre Haute, IN 47809, USA*

Correspondence should be addressed to Youping Deng; dengy@hawaii.edu

Received 17 October 2016; Accepted 17 October 2016

During the postgenomics era, more and more "omics" data are being produced due to rapid development of cutting-age technologies such as next generation sequencing. These omics data include genomics [1–3], transcriptomics [4–7], proteomics [8–10], metabolomics [11–13], and epigenomcis [14–16]. Bioinformatics plays a central role in analyzing and managing this huge amount of "omics" data, further understanding the function of biological molecules in different levels.

Several bioinformatics tools and methods have been developed in the special issue. C.-H. Tung et al. developed a new method called QuaBinggo, a prediction system for protein quaternary structure attributes using block composition. The method is 23% of Matthews Correlation Coefficient (MCC) higher than the existing prediction systems. *Exiguobacterium antarcticum* B7 is extremophile Gram-positive bacteria able to survive in cold environments. A key factor to understanding cold adaptation processes is related to the modification of fatty acids composing the cell membranes of psychrotrophic bacteria. R. Kawasaki et al. reconstructed the fatty acid biosynthesis pathway of *E. antarcticum* B7 based on both genomic and bibliomic data using bioinformatics methods, which is a great resource for the research of *Exiguobacterium antarcticum* B7. L. Hua and C. Quan built a novel method for protein-protein interaction (PPI) extraction using a shortest dependency path based CNN (sdpCNN) model. The proposed method only takes the sdp and word embedding as input and could avoid bias from feature selection by using CNN. The new approach outperformed traditional state-of-the-art kernel based methods. Clustered regularly interspaced short palindromic repeat (CRISPR) is a genetic element with active regulation roles for foreign invasive genes in the prokaryotic genomes and has been engineered to work with the CRISPR-associated sequence (Cas) gene Cas9 as one of the modern genome editing technologies. G. Mai et al. provided a valuable comprehensive curation resource to show the dynamic evolutionary patterns of prokaryotic CRISPRs based on computational evolutional analysis of 8 completely sequenced species in the genus *Thermoanaerobacter*.

Two papers are focused on transcriptomics data analyses. Q. Sun et al. tried to understand the gene function of thga1 in *Trichoderma harzianum* Th-33, important biocontrol filamentous fungi, which are widely used for their adaptability, broad antimicrobial spectrum, and various antagonistic mechanisms. Illumina RNA-seq technology (RNA-seq) was used to determine transcriptomic differences between the wild-type strain andthga1 mutant. A total of 888 genes were identified as differentially expressed genes (DEGs), including 427 upregulated and 461 downregulated genes. According to the functional annotation of these DEGs, they found the most abundant group was "secondary metabolite

biosynthesis, transport, and catabolism." Hepatitis E virus-(HEV-) mediated hepatitis has become a global public health problem. K. Xu et al. investigated the function of ORF3 from the swine form of HEV (SHEV); high-throughput RNA-Seq-based screening was conducted to identify the differentially expressed genes in ORF3-expressing HepG2 cells. The results indicated that, in the established ORF3-expressing HepG2 cells, the mRNA levels of CLDN6, YLPM1, APOC3, NLRP1, SCARA3, FGA, FGG, FGB, and FREM1 were upregulated, whereas the mRNA levels of SLC2A3, DKK1, BPIFB2, and PTGR1 were downregulated.

Several studies are concentrated on functional genomics data analyses. Using Web service SNP_TATA_Comparator presented in their previous paper, P. Ponomarenko et al. analyzed immediate surroundings of known SNP markers of diseases and identified 53 candidate SNP markers that can significantly change the affinity of TATA-binding protein for human gene promoters, with circadian consequences. These candidate SNP markers could be potentially useful for physicians (to select optimal treatment for each patient) and for the general population (to choose a lifestyle preventing possible circadian complications of diseases). M. Abu Saleh et al. conducted a comprehensive computational analysis on the functional and structural impacts of single nucleotide polymorphisms (SNPs) of the human ADIPOR1 at protein level. Their analyses suggested that the aforementioned variants, especially H341Y, could directly or indirectly destabilize the amino acid interactions and hydrogen bonding networks of ADIPOR1. Z.-X. Chen et al. used BLAST to call SNPs for non-model organisms based on 16 mixed functional gene's sequence data of polyploidy wheat. They demonstrated that mixed samples' NGS sequences and then analysis by BLAST were an effective, low-cost, and accurate way to mine SNPs for nonmodel species. Assembled reads and polynomial fitting threshold were recommended for more accurate SNPs targets.

One article deals with proteomics data analysis. S. Wan et al. indemnified important differential proteins between patients of slow transit constipation and normal controls using two-dimensional electrophoresis followed by laser desorption ionization tandem time-of-flight mass spectrometry (MALDI-TOF-MS). One paper is related to epigenomics data analysis. MiR-23a-27a-24-2 cluster has various functions and aberrant expression of the cluster is a common event in many cancers. Y. Wang et al. found a CG-rich region spanning two SP1 sites in the cluster promoter region. The SP1 sites in the cluster were demethylated and methylated in Hep2 cells and HEK293 cells, respectively. The demethylated SP1 sites in miR-23a-27a-24-2 cluster upregulate the cluster expression, leading to proliferation promotion and early apoptosis inhibition in laryngeal cancer cells.

Interaction of gene and environmental factors plays an important role in human diseases. Y. Wang et al. have found many important risk factors that affect chronic prostatitis/chronic pelvic pain syndrome (CP/CPPS), including biological, social, and psychological factors. They also discussed the potential interaction between genes and these risk factors.

In summary, this special issue presents a broad range of topics from functional genomics, transcriptomics, proteomics, epigenomics, and bioinformatics. It covers a variety of diseases such as cancer, hepatitis, chronic prostatitis, and infectious diseases. The study organisms include human, plant, and microorganisms. We hope that the readers will find interesting knowledge and methods in the issue.

Acknowledgments

The work edition was supported by the Grant NIH 5P30GM114737 and the Grant NIH P20GM103466.

Youping Deng
Hongwei Wang
Ryuji Hamamoto
Shiwei Duan
Mehdi Pirooznia
Yongsheng Bai

References

[1] M. Jia, Y. Liu, Z. Shen et al., "HDAM: a resource of human disease associated mutations from next generation sequencing studies," *BMC Medical Genomics*, vol. 6, supplement 1, p. S16, 2013.

[2] B. Bonev and G. Cavalli, "Organization and function of the 3D genome," *Nature Reviews Genetics*, vol. 17, no. 11, pp. 661–678, 2016.

[3] X. Chen, Y. Xu, W. Yang et al., "Association of six CpG-SNPs in the inflammation-related geneswith coronary heart disease," *Human Genomics*, vol. 10, supplement 2, p. 21, 2016.

[4] L. Hu, J. Ai, H. Long et al., "IntegrativemicroRNA and gene profiling data analysis reveals novel biomarkers and mechanismsfor lung cancer," *Oncotarget*, vol. 7, no. 8, pp. 8441–8454, 2016.

[5] M. Uhlén, B. M. Hallström, C. Lindskog, A. Mardinoglu, F. Pontén, and J. Nielsen, "Transcriptomics resources of human tissues and organs," *Molecular Systems Biology*, vol. 12, no. 4, p. 862, 2016.

[6] X. Wei, J. Ai, Y. Deng et al., "Identification of biomarkers that distinguish chemical contaminants based on geneexpression profiles," *BMC Genomics*, vol. 15, article 248, 2014.

[7] Y. Deng, S. A. Meyer, X. Guan et al., "Analysis of common and specific mechanisms of liver function affected by nitrotoluene compounds," *PLoS ONE*, vol. 6, no. 2, Article ID e14662, 2011.

[8] F. Zhang, Y. Deng, M. Wang, L. Cui, and R. Drabier, "Pathway-based biomarkers for breast cancer in proteomics," *Cancer Informatics*, vol. 13, supplement 5, pp. 101–108, 2014.

[9] D. Kumar, G. Bansal, A. Narang, T. Basak, T. Abbas, and D. Dash, "Integrating transcriptome and proteome profiling: strategies and applications," *Proteomics*, vol. 16, no. 19, pp. 2533–2544, 2016.

[10] J. Wang, M. Li, Y. Deng, and Y. Pan, "Recent advances in clustering methods for protein interaction networks," *BMC Genomics*, vol. 11, supplement 3, p. S10, 2010.

[11] X. Chen, H. Chen, M. Dai et al., "Plasma lipidomics profiling identified lipid biomarkers in distinguishing early-stage breast cancer from benign lesions," *Oncotarget*, vol. 7, no. 24, pp. 36622–36631, 2016.

[12] M. Marcinkiewicz-Siemion, M. Ciborowski, A. Kretowski, W. J. Musial, and K. A. Kaminski, "Metabolomics—a wide-open door to personalized treatment in chronic heartfailure?" *International Journal of Cardiology*, vol. 219, pp. 156–163, 2016.

[13] A. Cambiaghi, M. Ferrario, and M. Masseroli, "Analysis of metabolomic data: tools, current strategies and future challenges for omics data integration," *Briefings in Bioinformatics*, 2016.

[14] J. Melson, Y. Li, E. Cassinotti et al., "Commonality and differences of methylation signatures in the plasma of patients with pancreatic cancer and colorectal cancer," *International Journal of Cancer*, vol. 134, no. 11, pp. 2656–2662, 2014.

[15] Z. Sun, J. Cunningham, S. Slager, and J. P. Kocher, "Base resolution methylome profiling: considerations in platform selection, data preprocessing and analysis," *Epigenomics*, vol. 7, no. 5, pp. 813–828, 2015.

[16] J. Zierer, C. Menni, G. Kastenmüller, and T. D. Spector, "Integration of 'omics' data in aging research: from biomarkers to systems biology," *Aging Cell*, vol. 14, no. 6, pp. 933–944, 2015.

Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 8329513, 9 pages
http://dx.doi.org/10.1155/2016/8329513

Research Article

Functions of *thga1* Gene in *Trichoderma harzianum* Based on Transcriptome Analysis

Qing Sun, Xiliang Jiang, Li Pang, Lirong Wang, and Mei Li

State Key Laboratory for Biology of Plant Diseases and Insect Pests, Institute of Plant Protection, Chinese Academy of Agricultural Sciences, Beijing 100081, China

Correspondence should be addressed to Xiliang Jiang; jiangxiliang@caas.cn and Mei Li; limei@caas.cn

Received 16 February 2016; Revised 22 May 2016; Accepted 19 July 2016

Academic Editor: Hongwei Wang

Trichoderma spp. are important biocontrol filamentous fungi, which are widely used for their adaptability, broad antimicrobial spectrum, and various antagonistic mechanisms. In our previous studies, we cloned *thga1* gene encoding GαI protein from *Trichoderma harzianum* Th-33. Its knockout mutant showed that the growth rate, conidial yield, cAMP level, antagonistic action, and hydrophobicity decreased. Therefore, Illumina RNA-seq technology (RNA-seq) was used to determine transcriptomic differences between the wild-type strain and *thga1* mutant. A total of 888 genes were identified as differentially expressed genes (DEGs), including 427 upregulated and 461 downregulated genes. All DEGs were assigned to KEGG pathway databases, and 318 genes were annotated in 184 individual pathways. KEGG analysis revealed that these unigenes were significantly enriched in metabolism and degradation pathways. GO analysis suggested that the majority of DEGs were associated with catalytic activities and metabolism processes that encode carbohydrate-active enzymes, secondary metabolites, secreted proteins, or transcription factors. According to the functional annotation of these DEGs by KOG, the most abundant group was "secondary metabolite biosynthesis, transport, and catabolism." Further studies for functional characterization of candidate genes and pathways reported in this paper are necessary to further define the G protein signaling system in *T. harzianum*.

1. Introduction

Trichoderma is an important fungal genus, whose species exhibit favorable properties, such as diverse mechanisms of antagonistic action, a broad spectrum of activity in plant disease prevention and control, survival under unfavorable conditions, and environmental friendliness. *Trichoderma harzianum* is one of the most commercially used biofungicides, particularly *T. harzianum* T22 and T39 [1]. Growth, conidiation, secondary metabolism, and mycoparasitism are all important processes that contribute to biofungicidal property [2]. An enhanced understanding of the signal regulatory mechanism in *T. harzianum* is necessary to further explore the fungi's extraordinary biocontrol potential.

G protein-mediated signal transduction system is an important transmembrane signaling system in eukaryotic cells. It plays a key role in the regulation of cellular reactions and the transmission of extracellular signals to cells. The role of G protein in the transmission of external stimuli has been studied in detail in genetic fungi models, such as *Neurospora, Penicillium* [3], and *Aspergillus* [4]. Heterotrimeric G protein is composed of α, β, and γ subunits; each subunit is encoded by independent genes. Among the G protein subunits, the Gα subunits were found more frequently and reported to regulate vegetative growth, conidiation, and the mycoparasitic responses in fungi [5]. However, current observations showed that a particular Gα subunit may have different functions in different fungal species [6]. Gα subunits are classified into three groups according to conserved motif sequences [7], and several subgroup GαI subunits from *T. atroviride* and *T. virens* have been studied [4, 7]. Gα from *T. atroviride*, *tga1*, was reported to negatively regulate conidiation but had no effect on hyphal growth [8]. *TgaA* is a homology gene of *tga1* from *T. atroviride*, which does not influence growth and conidiation in *T. virens*. The Δ*tgaA* mutants grow normally and sporulate like the wild type but possess a reduced ability to colonize *Sclerotinia sclerotiorum*, whereas they are fully pathogenic against *Rhizoctonia solani* [4].

In our previous study, a subgroup I Gα gene (*thga1*) was cloned from the *T. harzianum* Th-33 genome. Results showed that THGA1 has the same amino acid sequence as that of TGA from *T. atroviride* but with different functions. Compared with the wild-type Th-33, the Δ*thga1* mutants changed significantly in biological characteristics and physicochemical properties. In particular, the hyphal growth rate dropped by about 40%, conidiophore branches became sparse, secondary branch and phialide numbers reduced, conidiation was delayed for about 20 h, and conidia yield declined by about 300-fold. In addition, the hydrophobicity of the mutants weakened, intracellular cAMP levels decreased by about 50%, and the inhibition of the mutant against plant pathogen of *R. solani* reduced significantly. The results showed that the *thga1* gene positively affected the growth, conidial production, hydrophobicity, and antagonism of *T. harzianum* to *R. solani*.

The present study characterized the genes under- or overexpression in the mutant compared to WT associated with the Gα subunit, *thga1*, using next-generation RNA sequencing (RNA-seq) technology, to gain insight into the regulatory mechanism of Gα subunits *thga1* in Th-33. This study is the first initiative to use RNA-seq for identifying differentially expressed genes (DEGs) to clarify the function of Gα in *Trichoderma* by genome-wide transcriptional analysis of hyphal cells of the wild-type Th-33 and its Δ*thga1* mutants. The results may reveal that Gα regulated a series of biological processes as well as new targets of *thga1* function.

2. Materials and Methods

2.1. Strains and Culture Media. The wild-type *T. harzianum* Th-33 was isolated from soil samples in the Beijing region as described previously. The Δ*thga1* knockout mutant was created with hygromycin B resistance by homologous recombination and then purified by isolation of single conidia [9]. Mutant colony was identified by southern hybridization. The fungi were grown on potato dextrose agar (PDA) at 28°C for conidia production. The PDA medium was covered with cellophane for mycelia collection. Conidial suspensions were prepared by adding sterilized distilled water to the PDA plates. The conidial suspensions were inoculated on the PDA medium covered with cellophane and cultured for 32 h prior to conidia formation at the tip of the mycelia. The mycelia were then scraped off the cellophane, washed with cold distilled water, frozen in liquid nitrogen, and ground to a fine powder. Equal mixture of three biological samples was sequenced.

2.2. Preparation of the cDNA Library and Illumina Sequencing for Transcriptome Analysis. Total RNA was extracted using Trizol reagent (Invitrogen, CA, USA) according to the manufacturer's instruction. The RNA quality and quantity were determined using an Agilent 2100 Bioanalyzer. The Qubit RNA Assay Kit was used for accurate quantification of the initial total RNA. After total RNA extraction and DNase I treatment, magnetic beads with oligo (dT) were used to isolate mRNA. The mRNA was mixed with the fragmentation buffer to promote fragmentation into short segments. Subsequently, cDNA was synthesized using SuperScript II reverse transcriptase following the manufacturer's protocol. The short fragments were connected with adapters. The suitable fragments were selected as templates for PCR amplification. During the QC steps, a 2100 Bioanalyzer High Sensitivity DNA chip was used for quantification and qualification of the sample library. RNA libraries were sequenced by paired-end mode using an Illumina HiSeq2000 system.

2.3. Sequence Alignment. Clean reads were obtained after removal of low quality reads, reads with adaptors, and reads with unknown nucleotides larger than 5%. All reads were deposited in the National Center for Biotechnology Information (NCBI) database and can be found under accession number SRS823675. The reads were mapped to the reference genome [10] of *T. harzianum* Th-33 using TopHat (Version: 2.0.11) program [11].

2.4. Expression Analysis. Expression values were obtained by calculating the fragments per kilobase of transcript per million mapped reads, and differential gene expression was analyzed using the Cufflinks program [12]. Genes differentially expressed with more than twofold changes and at p values less than 0.05 were identified as DEGs. The threshold of the p value in multiple tests was determined by the value for the false discovery rate (FDR) [13]. We used "FDR $\leq$ 0.001 and the absolute value of $\log_2$ fold change ($\log_2$ FC) $\geq$ 2" as the threshold to assess the significance of gene expression differences.

2.5. Functional Annotation of T. harzianum Transcriptome and Classification of DEGs. The *T. harzianum* transcriptomes were compared against amino acid sequences available at the UniProt database using BLASTx algorithm. For each query sequence, the associated hits were searched for their respective gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) results. The highest bit score was selected, with E-value threshold of 1e201. Protein families were classified by searching the assembled transcripts against Pfam and InterProScan. Potential transmembrane domains of the G protein-coupled receptors (GPCRs) selected from InterProScan and Pfam analysis were predicted using transmembrane domain prediction tool TMHMM v2.0. GPCRs predicted to contain the seven-transmembrane domain were used for further analysis. We used the CAZy database to annotate the carbohydrate-active enzymes (http://www.cazy.org/). We annotated transcription factors (TFs) of DEGs based on protein sequence homology using the Fungal Transcription Factor Database (http://ftfd.snu.ac.kr/). The amino acid sequences of DEGs were further analyzed to predict secreted proteins. Sequences smaller than 70 amino acids were not considered for further analysis. The remaining sequences with positive SignalP prediction for signal peptide cleavage site at the N-terminal region between 10 and 40 amino acids, without any transmembrane region, were selected as the candidate secreted proteins. Secreted proteins with lengths less than 200 amino acids and cysteine content more than 4% of the protein were identified as small molecule cysteine-rich secreted proteins (SSCPs).

2.6. Validation of RNA-Seq Results by Quantitative Real-Time RT-PCR (qRT-PCR). To validate the expression profiles

Table 1: Summary of Illumina transcriptome sequencing data for the strains included in this study.

Sample	Read	Length	QV	Reads number	q20 (%)	q30 (%)	Yield
Th-33	Reads 1	101	36	14053573	98.24	94.96	1419410873
	Reads 2	101	35	14053573	96.89	93.09	1419410873
1-1	Reads 1	101	36	16049540	98.35	95.24	1621003540
	Reads 2	101	35	16049540	96.53	92.46	1621003540

of the assembled genes obtained through sequencing data analysis, qRT-PCR was performed for selected genes. Genes were randomly selected, and four reference genes were used, namely, beta-tubulin 1, actin, transcription elongation factor, and ubiquitin-conjugating enzyme (UCE). Primers used in qRT-PCR were designed using Primer Express Software v2.0 (see S2 Table in Supplementary Material available online at http://dx.doi.org/10.1155/2016/8329513). Total RNA (2 μg) from each sample was reverse transcribed into cDNA in the presence of oligo (dT) primer in a volume of 12.5 μL. The synthesized cDNA was used as a template for qRT-PCR. Reactions were performed in the ABI StepOne Plus Real-Time PCR System (Applied Biosystems). Each reaction (20 μL) contained 10 μL of 2x SYBR Green PCR mix (QIAGEN), 1 μL of forward and reverse primers (10 pM/μL each), 1 μL of cDNA template, and 7 μL of nuclease-free water. PCR cycling conditions were 2 min at 95°C (1 cycle) and 10 s at 94°C, followed by 10 s at 60°C and a melting curve of 40 s at 72°C (40 cycles). PCR cycles for negative controls without templates were carried out concurrently. All qPCRs for each gene used three biological replicates, with three technical replicates per experiment. The average threshold cycle (Ct) was calculated using the $2^{-\Delta\Delta Ct}$ method [14].

2.7. Availability of Supporting Data. Sequences have been deposited at the Sequence Read Archive (SRA) of the NCBI under BioProject number PRJNA272748. Raw sequence reads can be found in http://trace.ncbi.nlm.nih.gov/.

3. Results and Discussion

3.1. Illumina Sequencing. We got 17 transformants with stable genetic characteristics. The homologous recombination events have been confirmed by southern hybridization. A probe corresponding to part of the *thga1* gene coding region gave no signal on a southern blot of the deletion mutants, while the hygromycin resistance cassette was detected for the mutants but not for the wild type (data not shown). We selected one mutant with obvious phenotypic differences to be studied. Two cDNA libraries were constructed and subjected to Illumina deep sequencing. Approximately 2,838,821,746 and 3,242,007,080 bp clean reads were generated for the wild type and the mutant 1-1, respectively (Table 1). The reads were then compared against the genome of Th-33, which was sequenced and submitted to SRA (http://trace.ncbi.nlm.nih.gov/Traces/sra/). The whole genome of *Trichoderma harzianum* Th-33 was sequenced on a Hiseq2500 instrument. A total of 196 scaffolds were assembled and 10849 genes were predicted with an average length of 1776 bp (GenBank number: PRJNA272949). The unigenes were used for transcriptome analysis in this study. The overall mapping rate was about 95%, indicating that subsequent analysis can be performed.

3.2. DEGs Regulated by thga1. RNA-seq technology was used to investigate the transcriptional changes between the wild-type Th-33 and the mutant regulated by *thga1*. Comparisons between the gene expression of *thga1* mutant and wild-type Th-33 showed that 888 genes were differentially expressed, including 427 upregulated and 461 downregulated genes. Among the 20 significantly upregulated genes (S1 Table), 13 genes exhibited defined functions, including six metabolism-related genes and two genes predicted to encode enzymes. Among the significantly downregulated genes (S1 Table), seven were metabolism-related genes and four were catalytic activity genes. The changes in expression were evidently connected to metabolism. Hence, *thga1* might regulate a series of biological processes through metabolism.

3.3. Real-Time qRT-PCR Analyses. To verify the quality of the assembly, cDNA fragments of 15 randomly selected unigenes were amplified using unigene-specific primers (S2 Table) and then sequenced. Four reference genes (beta-tubulin 1, actin, transcription elongation factor, and UCE) were used as endogenous control. Among the reference genes, UCE remained constant in all treatments and showed the best performance in qRT-PCR analysis using Best Keeper program. Expression patterns of the tested genes are shown in Figure 1. Three genes with infinite fold changes in the RNA-seq results were not shown in the figure, as they cannot be shown in the bar chart. Expression patterns determined by real-time qRT-PCR were consistent with those obtained by RNA-seq, thereby confirming the accuracy of the RNA-seq results reported in this study.

3.4. Pathway Functional Enrichment Analysis of DEGs. In our research, 318 differentially expressed unigenes were assigned to 184 KEGG pathways. A summary of the findings is presented in S3 Table. Among the pathways identified, the metabolic pathways, especially the secondary metabolic pathways, were found to be the most active. "Bisphenol degradation" (27 unigenes, 8.49% of sequences), "aminobenzoate degradation" (27 unigenes, 8.49% of sequences), "chloroalkane and chloroalkene degradation" (25 unigenes, 7.86% of sequences), and "butanoate metabolism" (22 unigenes, 6.92% of sequences) were the dominant pathways (Figure 2). Thus, *thga1* may affect growth, conidiation, and antagonism through metabolism. The DEG catalogue provided a comprehensive understanding of the gene transcription profiles of *T. harzianum* missing the *thga1* gene and offered

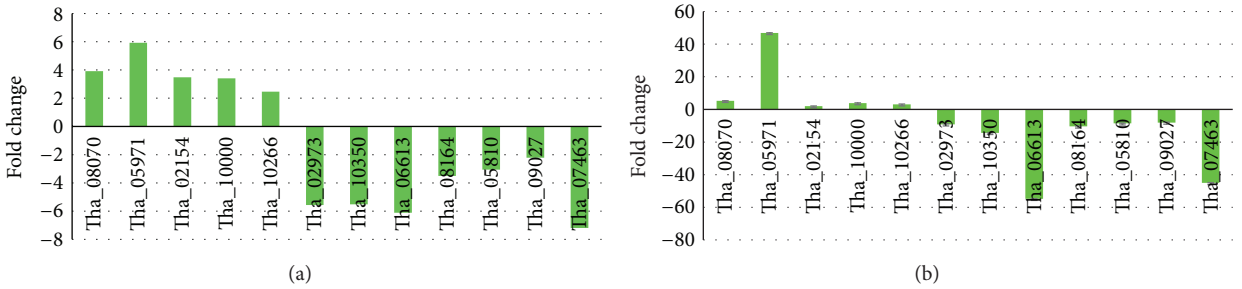


Figure 1: Transcriptome validation. (a) DEG data in transcriptome analysis. The fold changes of the genes were calculated as the $\log_2$ value of each 1-1/Th-33 comparison and are shown on the y-axis. (b) The qRT-PCR analysis of gene expression data. Expression ratios of selected genes in 1-1 compared to Th-33.

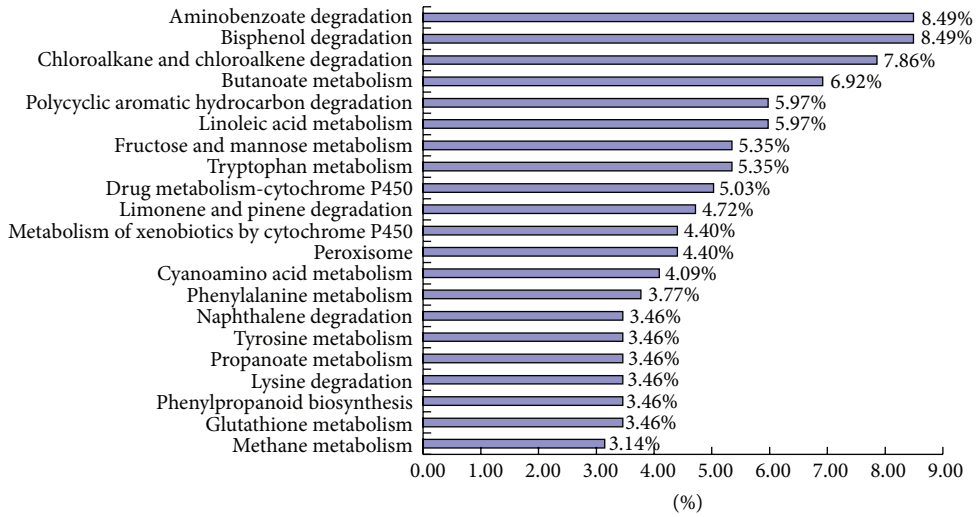


Figure 2: Pathway enrichment analysis of DEGs. The percentage of differentially expressed genes involved in KEGG pathways. Only the top 21 most abundant EC and KEGG pathways are represented.

a valuable foundation for the screening of G protein-mediated pathways.

3.5. GO Enrichment Analysis. Blast2GO [15] program was used for GO analysis. The program extracted the GO terms associated with the homologies identified by BLAST and returned a list of GO annotations, which were presented as hierarchical categories of increasing specificity. GO enrichment analyses were performed using Fisher's exact test with multiple testing corrections and an FDR of 0.05. A total of 517 DEGs were categorized into 707 functional groups in three main categories, namely, "cellular component," "molecular function," and "biological process" (S4 Table and Figure 3). Some unigenes were assigned to multiple categories of GO terms, whereas others could not be assigned to a given GO term. In the biological process category, "metabolic process" (381, 73.69%), "cellular process" (201, 38.87%), "single-organism process" (115, 22.24%), "localization" (100, 19.34%), and "establishment of localization" (99, 19.15%) were the most abundant terms. In the molecular function category, genes associated with "catalytic activity" (351, 67.89%), "binding" (191, 36.94%), and "transporter activity" (54, 10.44%) were the most abundant. In the cell component category, "membrane" (124, 23.98%) and "membrane part" (100, 19.34%) were the most abundant terms. These findings indicated that the main changes in expression between the wild-type Th-33 and the *thga1* mutant were those related to membrane parts, metabolic processes, and catalytic activities.

3.6. KOG Function Classification. DEGs were also searched against the KOG database for functional prediction and classification. Notably, out of 888 distinct DEGs, 463 could be functionally classified into 23 molecular families, which was consistent with the approximately 50% KOG annotation rate of the Frozen Gene Catalogue (S5 Table and Figure 4). The largest number of unigenes focused on "the general function of prediction (19.3%)," whereas the next largest groups were noted in "secondary metabolite biosynthesis (13.1%), transport, and catabolism (13.1%)," followed by "posttranslational modification; protein turnover; chaperones (8.1%)," "energy production and conversion (7.9%)," "carbohydrate transport and metabolism (7.0%)," "lipid metabolism (6.6%)," "amino acid transport and metabolism (6.1%)," and "signal transduction mechanisms (5.9%)." The three smallest groups were included in "chromatin structure and dynamics (0.37%)," "extracellular structures (0.37%)," and "coenzyme transport

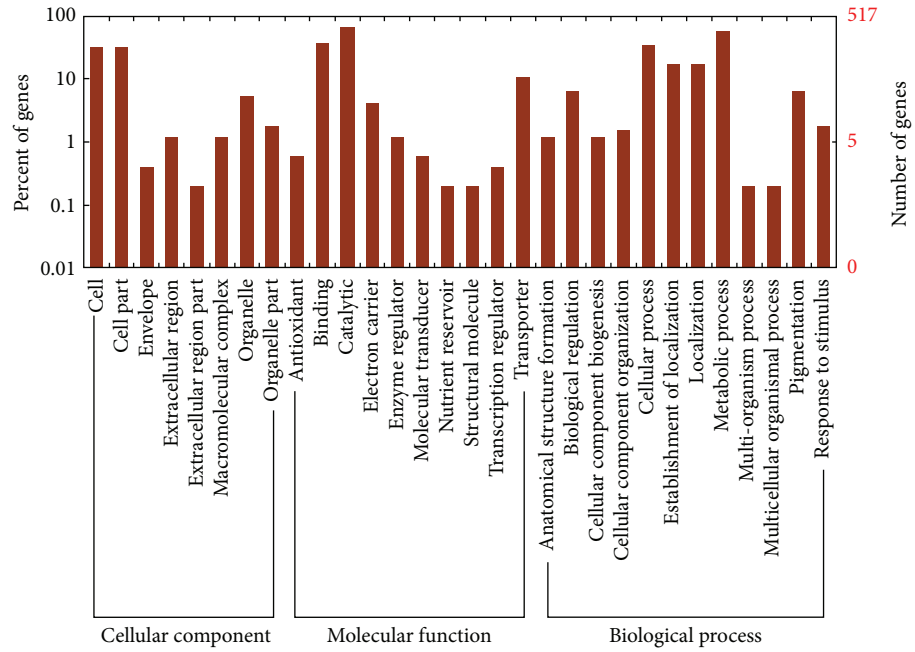


Figure 3: Gene ontology terms for differentially expressed genes. Most differentially expressed genes were grouped into three major functional categories: cellular component, molecular function, and biological process.

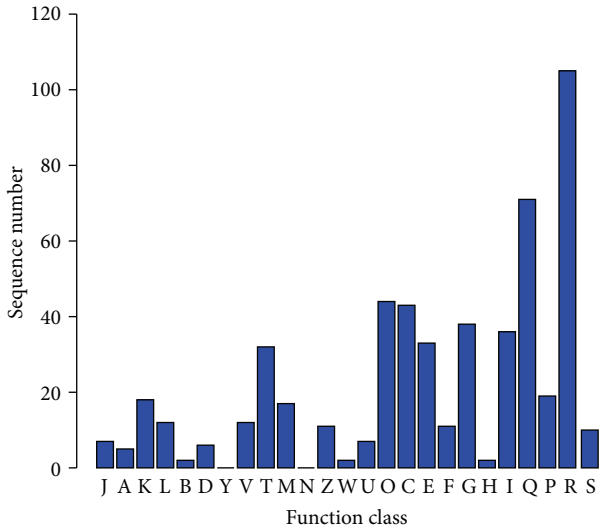


Figure 4: KOG function classification of the differentially expressed genes. J: translation; ribosomal structure; biogenesis. A: RNA processing and modification. K: transcription. L: DNA replication; recombination; repair. B: chromatin structure and dynamics. D: cell cycle control; cell division; chromosome partitioning. Y: nuclear structure. V: defense mechanisms. T: signal transduction mechanisms. M: cell wall/membrane/envelope biogenesis. N: cell motility. Z: cytoskeleton. W: extracellular structures. U: intracellular trafficking; secretion; vesicular transport. O: posttranslational modification; protein turnover; chaperones. C: energy production and conversion. E: amino acid transport and metabolism. F: nucleotide transport and metabolism. G: carbohydrate transport and metabolism. H: coenzyme transport and metabolism. I: lipid metabolism. Q: secondary metabolites biosynthesis; transport; catabolism. P: inorganic ion transport and metabolism. R: general function prediction only. S: function unknown.

and metabolism (0.37%)." Genes involved in intermediary metabolism (i.e., of carbohydrates, amino acids, and lipids) comprised a significant portion within both up- and down-regulated genes.

3.7. Genes Related to G Protein Signal Pathway. G protein transduced signals received by the heptahelical GPCRs from outside the cell influence numerous regulatory pathways via their respective effectors, which in turn affect the activities of secondary messengers [16, 17]. Three Gα-coding genes, one beta, and one gamma subunit genes were found in the transcriptomes and the genome of Th-33, which corresponded well with the data of *T. virens* [18], *Neurospora crassa* [19], and many other filamentous fungi [6]. *Thga1*, the subgroup I Gα gene, was not expressed in the Δ*thga1* mutant. Another Gα gene, a subgroup II *tgaB* homologous gene in *T. virens* [4], exhibited downregulated expression (twofold), whereas the third Gα gene β and Gγ gene were not differentially expressed. These results showed that knockout of *thga1* caused the downregulated expression of another Gα gene.

The signals transmitted through a heterotrimeric G protein signaling cascade originate from the activation of plasma membrane-localized GPCRs. The identification and characterization of GPCRs will provide insights into how *Trichoderma* communicates with G protein and downstream genes. Thirty-nine GPCRs or putative GPCR genes were found in the transcriptome of Th-33. Six of them were differentially expressed in the mutant (five of them belong to the PTH11-type; Table 2). PTH11-type GPCRs were reported to influence light responsiveness, glycoside hydrolase gene transcription, sexual development [20, 21], surface recognition, and pathogenicity [22]. PTH11 GPCR genes were

Table 2: Differentially expressed genes related to the G protein signal pathway.

Gene_id	Th-33	1-1	$\log_2$ (fold change)	p value	Annotation
Tha_09027	215.43	46.28	−2.22	0.0475	G protein alpha subunit TgaB
Tha_10266	11.00	60.75	2.47	0.0225	Cyclic AMP phosphodiesterase
Tha_08164	216.24	19.09	−3.50	0.0019	GPCR, PTH11-type
Tha_04234	107.24	11.21	−3.26	0.0053	GPCR, PTH11-type
Tha_05810	217.59	26.05	−3.06	0.0051	GPCR, PTH11-type
Tha_07855	75.77	14.01	−2.43	0.0121	G protein-coupled receptor
Tha_07047	35.10	6.85	−2.36	0.0308	GPCR, PTH11-type
Tha_06854	2.35	0	Inf	0.0078	GPCR, PTH11-type

reported to be upregulated in the mycoparasite *Coniothyrium minitans* during colonization of *S. sclerotiorum* [23]. In the current study, five PTH11-type GPCR-encoding genes were downregulated in the mutant. This observation was consistent with the phenomenon that the inhibitory effect of the mutant against *R. solani* decreased significantly compared with that of the wild-type Th-33.

cAMP acts as a secondary messenger for morphogenic signals in both prokaryotes and eukaryotes, and it is generated by adenylate cyclase and degraded by phosphodiesterase [24]. In *T. virens*, the intracellular cAMP levels can be regulated by an adenylate cyclase gene, *tac1* [25], and it has been reported to influence conidiation [26]. In *Aspergillus nidulans*, a Gα-subunit *GanB* mediates a rapid and transient activation of cAMP synthesis in response to glucose during the early period of germination [27]. In our previous study, intracellular cAMP levels in the Δ*thga1* mutant decreased by about 50% of that in the wild-type Th-33. According to the transcriptomes of the Δ*thga1* mutant and the wild-type Th-33, the adenylate cyclase-encoding gene was not expressed differentially, whereas the cAMP phosphodiesterase gene was found to be upregulated in the mutant. These findings indicated that the decrease in intracellular cAMP levels was caused by the increase in cAMP phosphodiesterase activity and not by the decrease in adenylate cyclase activity. This study showed that the G protein α subunit in *Trichoderma* may be involved in the mechanisms regulating intracellular cAMP levels.

3.8. Carbohydrate Activity Enzymes (CAZymes). CAZymes in fungi are correlated with their nutritional modes and infection mechanisms [28]. Various CAZyme genes exist in the Th-33 genome, of which 66 differentially expressed CAZyme genes were identified in the Th-33 transcriptome, including 30 genes from the glycoside hydrolase family (GH), 14 from the auxiliary activity family, 13 from the carbohydrate esterase family, 7 from the glycosyltransferase family, and 2 from the carbohydrate-binding module family (S6 Table). In addition, the number of downregulated CAZyme genes (43 genes) in the mutant was greater than that of upregulated genes (23 genes) (Figure 5). The GH family has been described to play a central role in mycoparasitism in *Trichoderma*. For example, GH18 family is highly expanded in *T. atroviride* and *T. virens* during mycoparasitism [29]. GH16 was reported to be upregulated during the mycoparasitic reaction of *T. atroviride* against *R. solani* [30]. In the present study, the inhibition of the Δ*thga1* mutant *T. harzianum* against *R. solani* and *Phytophthora capsici* was reduced significantly, and the mycelia of Δ*thga1* could not cover and parasitize the pathogenic fungi's mycelia in confrontation tests. The heterotrimeric G protein pathway is crucial in the interconnections of nutrient signaling in *Trichoderma* [21]. Results implied that the knockout of *thga1* caused the decrease in the activities of the CAZymes, which consequently influenced nutrient condition, further retarded the growth rate, and decreased the mycoparasitic ability of *Trichoderma*.

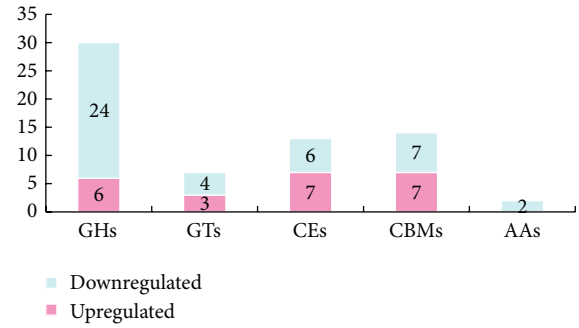


Figure 5: Numbers of carbohydrate-active enzyme modules in the differentially expressed genes.

3.9. TFs. Twenty-nine TFs were identified in the Δ*thga1* mutant and compared with that of the wild-type Th-33. Of these TFs, 15 were upregulated and 14 were downregulated TFs (S7 Table). Among those differentially expressed TFs, the Zn_2Cys_6 family (53%) was the most dominant TF family, followed by the C_2H_2 zinc finger (31%) TF family (Figure 6). The functions of most TFs in fungi appear to be highly diverse and remain rather unclear. Of the known TF families, the Zn_2Cys_6 TF family comprises TFs that are fungal-specific and dominant [31]. These TFs are involved in primary and secondary metabolism, drug resistance, and meiotic development (http://ftfd.snu.ac.kr/). BglR, a Zn_2Cys_6-type TF in *T. reesei*, can increase the expression of the β-glucosidase gene [32]. C_2H_2 zinc finger in fungi is involved in pathogenicity, carbon catabolite repression, acetamide regulation, and differentiation of ascogenous or fruiting body. GipA in *A. fumigatus* was reported to be involved in secondary

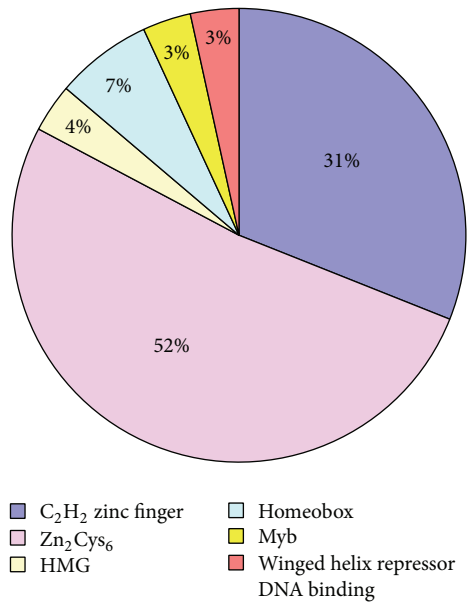


Figure 6: Percentage of different TFs in the DEGs.

metabolic processes [33]. In *A. nidulans*, brlA defined the central regulatory pathway that controls conidiation-specific gene expression, but no orthologue of brlA in *Trichoderma* conidiophores currently exists [34]. TFs can control gene transcription rates and are key mediators of cellular function [35]. In the present report, differentially expressed TFs might play key roles in cell processes in *Trichoderma* under the control of the G protein signal pathway. Understanding the regulatory mechanisms of these TFs and their functions could provide valuable insight into the signal control pathways of cell processes in *Trichoderma*.

3.10. Secretome of T. harzianum. A total of 137 genes were annotated as secreted proteins, including 51 upregulated and 86 downregulated genes among the DEGs. Secreted proteins play an important role in the process of cell signaling, cell proliferation, differentiation, apoptosis regulation, development, and other important biological processes. In *Botrytis cinerea*, *bcg1* encodes α subunits of heterotrimeric GTP-binding proteins; the *bcg1* null mutants differ in colony morphology from the wild-type strain, which do not secrete extracellular proteases [36]. SSCPs comprise one of the largest groups of proteins secreted by *Trichoderma* [29]. Hydrophobins, which are probably the most widely known SSCPs, are found on the outer surfaces of cell walls of hyphae and conidia. *Trichoderma* genomes encode an unusually large number of hydrophobins, possibly for the purpose of providing flexibility in surface properties needed for conidiation, mycoparasitism, and interaction with plant roots [2, 37]. In the present report, two SSCPs were found among the DEGs, and these genes were downregulated in the mutant strain, which might account for the phenomenon of reduced hydrophobicity in the mutant. This was in accordance with the report that G alpha subunit can mediate fungal conidiation and hydrophobin synthesis in *Cryphonectria parasitica* [38]. Furthermore, SSCPs were shown to be highly expressed in *T. atroviride* during colonization of *R. solani,* which suggested a potential role in mycoparasitism [30]. These genes may also explain the significant reduction in the inhibition of the mutant against the plant pathogen *R. solani*.

3.11. Secondary Metabolism. KOG functional classification of DEGs revealed that the number of genes involved in secondary metabolite biosynthesis, transport, and catabolism was 71, in which cytochrome P450s (CYPs) accounted for nearly 24% (S8 Table). CYPs play an important role in the physiology of fungi and are involved in the biosynthesis of secondary metabolites (SMs) and in detoxification [39]. Fungi produce a wide range of SMs that are not directly essential for growth and yet have important roles in signaling, asexual conidiation [40], and interactions with other organisms [41–43]. *Trichoderma* spp. are a rich source of SMs, such as nonribosomal peptides, polyketides, terpenoids, and pyrones. Although a clear correlation exists between conidiation of the fungus and the secretion of antifungal metabolites in the *T. atroviride* parental strain, its type GαIII gene Δ*tga3* mutants were fully impaired in the production of peptaibols despite exhibiting a hypersporulating phenotype [5]. The results in the current report implied that *thga1* regulated certain secondary metabolism processes, and analysis of SMs should be carried out for further identification.

4. Conclusion

Overall, this study identified transcripts that were possibly involved in several important molecular and cellular functions of *T. harzianum*. However, additional studies aimed at the functional characterization of the genes reported herein will aid in further defining the pathways mediated by G protein in *T. harzianum*. An enhanced understanding of the expression profiles of these genes could improve *T. harzianum* performance, either by predicting the regulation of the genes involved in sporulation or by improving their use in biotechnology processes.

Competing Interests

The authors declare that there are no competing interests regarding the publication of this paper.

Acknowledgments

The authors are grateful for the financial support from the National Natural Science Foundation of China (Grant no. 31371983).

References

[1] F. Vinale, E. L. Ghisalberti, K. Sivasithamparam et al., "Factors affecting the production of *Trichoderma harzianum* secondary metabolites during the interaction with different plant pathogens," *Letters in Applied Microbiology*, vol. 48, no. 6, pp. 705–711, 2009.

[2] P. K. Mukherjee, B. A. Horwitz, A. Herrera-Estrella, M. Schmoll, and C. M. Kenerley, "*Trichoderma* research in the genome era," *Annual Review of Phytopathology*, vol. 51, pp. 105–129, 2013.

[3] Y. Hu, G. Liu, Z. Li, Y. Qin, Y. Qu, and X. Song, "G protein-cAMP signaling pathway mediated by PGA3 plays different roles in regulating the expressions of amylases and cellulases in *Penicillium decumbens*," *Fungal Genetics and Biology*, vol. 58-59, pp. 62–70, 2013.

[4] P. K. Mukherjee, J. Latha, R. Hadar, and B. A. Horwitz, "Role of two G-protein alpha subunits, TgaA and TgaB, in the antagonism of plant pathogens by *Trichoderma virens*," *Applied and Environmental Microbiology*, vol. 70, no. 1, pp. 542–549, 2004.

[5] M. Omann and S. Zeilinger, "How a mycoparasite employs G-protein signaling: using the example of *Trichoderma*," *Journal of Signal Transduction*, vol. 2010, Article ID 123126, 8 pages, 2010.

[6] L. Li, S. J. Wright, S. Krystofova, G. Park, and K. A. Borkovich, "Heterotrimeric G protein signaling in filamentous fungi," *Annual Review of Microbiology*, vol. 61, pp. 423–452, 2007.

[7] B. A. Horwitz, A. Sharon, S.-W. Lu et al., "A G protein alpha subunit from *Cochliobolus heterostrophus* involved in mating and appressorium formation," *Fungal Genetics and Biology*, vol. 26, no. 1, pp. 19–32, 1999.

[8] B. Reithner, K. Brunner, R. Schuhmacher et al., "The G protein α subunit Tga1 of *Trichoderma atroviride* is involved in chitinase formation and differential production of antifungal metabolites," *Fungal Genetics and Biology*, vol. 42, no. 9, pp. 749–760, 2005.

[9] Z. Liu, *Cloning and functional characterization of Thga1 gene of Trichoderma harzianum Th-33 [Thesis]*, Chinese Academy of Agricultural Sciences, 2013.

[10] Q. Sun, X. Jiang, L. Pang, L. Wang, and M. Li, "The genome sequence of *Trichoderma harzianum* Th-33," *Chinese Journal of Biological Control*, vol. 32, no. 2, pp. 205–214, 2016.

[11] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg, "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions," *Genome Biology*, vol. 14, no. 4, article R36, 2013.

[12] 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.

[13] Y. Benjamini and D. Yekutieli, "The control of the false discovery rate in multiple testing under dependency," *The Annals of Statistics*, vol. 29, no. 4, pp. 1165–1188, 2001.

[14] P. Y. Muller, H. Janovjak, A. R. Miserez, and Z. Dobbie, "Processing of gene expression data generated by quantitative real-time RT-PCR," *BioTechniques*, vol. 32, no. 6, pp. 1372–1379, 2002.

[15] A. Conesa and S. Götz, "Blast2GO: a comprehensive suite for functional analysis in plant genomics," *International Journal of Plant Genomics*, vol. 2008, Article ID 619832, 12 pages, 2008.

[16] H. E. Hamm, "The many faces of G protein signaling," *The Journal of Biological Chemistry*, vol. 273, no. 2, pp. 669–672, 1998.

[17] E. J. Neer, "Heterotrimeric C proteins: organizers of transmembrane signals," *Cell*, vol. 80, no. 2, pp. 249–257, 1995.

[18] S. Gruber, M. Omann, and S. Zeilinger, "Comparative analysis of the repertoire of G protein-coupled receptors of three species of the fungal genus *Trichoderma*," *BMC Microbiology*, vol. 13, no. 1, article 108, 2013.

[19] K. A. Borkovich, L. A. Alex, O. Yarden et al., "Lessons from the genome sequence of *Neurospora crassa*: tracing the path from genomic blueprint to multicellular organism," *Microbiology and Molecular Biology Reviews*, vol. 68, no. 1, pp. 1–108, 2004.

[20] R. D. Kulkarni, M. R. Thon, H. Pan, and R. A. Dean, "Novel G-protein-coupled receptor-like proteins in the plant pathogenic fungus *Magnaporthe grisea*," *Genome Biology*, vol. 6, no. 3, p. R24, 2005.

[21] D. Tisch and M. Schmoll, "Targets of light signalling in *Trichoderma reesei*," *BMC Genomics*, vol. 14, no. 1, article 657, 2013.

[22] T. M. DeZwaan, A. M. Carroll, B. Valent, and J. A. Sweigard, "*Magnaporthe grisea* Pth11p is a novel plasma membrane protein that mediates appressorium differentiation in response to inductive substrate cues," *Plant Cell*, vol. 11, no. 10, pp. 2013–2030, 1999.

[23] S. Muthumeenakshi, S. Sreenivasaprasad, C. W. Rogers, M. P. Challen, and J. M. Whipps, "Analysis of cDNA transcripts from *Coniothyrium minitans* reveals a diverse array of genes involved in key processes during sclerotial mycoparasitism," *Fungal Genetics and Biology*, vol. 44, no. 12, pp. 1262–1284, 2007.

[24] S. M. Byrne and C. S. Huffman, "Six git genes encode a glucose-induced adenylate cyclase activation pathway in the fission yeast *Schizosaccharomyces pombe*," *Journal of Cell Science*, vol. 105, no. 4, pp. 1095–1100, 1993.

[25] M. Mukherjee, P. K. Mukherjee, and S. P. Kale, "cAMP signalling is involved in growth, germination, mycoparasitism and secondary metabolism in *Trichoderma virens*," *Microbiology*, vol. 153, no. 6, pp. 1734–1742, 2007.

[26] J. M. Steyaert, R. J. Weld, A. Mendoza-Mendoza, and A. Stewart, "Reproduction without sex: conidiation in the filamentous fungus *Trichoderma*," *Microbiology*, vol. 156, no. 10, pp. 2887–2900, 2010.

[27] A. Lafon, J.-A. Seo, K.-H. Han, J.-H. Yu, and C. D'Enfert, "The heterotrimeric G-protein GanB(α)-SfaD(β)-GpgA(γ) is a carbon source sensor involved in early cAMP-dependent germination in *Aspergillus nidulans*," *Genetics*, vol. 171, no. 1, pp. 71–80, 2005.

[28] Z. Zhao, H. Liu, C. Wang, and J. R. Xu, "Comparative analysis of fungal genomes reveals different plant cell wall degrading capacity in fungi," *BMC Genomics*, vol. 14, no. 6, article 274, 2014.

[29] C. P. Kubicek, A. Herrera-Estrella, V. Seidl-Seiboth et al., "Comparative genome sequence analysis underscores mycoparasitism as the ancestral life style of *Trichoderma*," *Genome Biology*, vol. 12, no. 4, article R40, 2011.

[30] L. Atanasova, S. L. Crom, S. Gruber et al., "Comparative transcriptomics reveals different strategies of *Trichoderma* mycoparasitism," *BMC Genomics*, vol. 14, no. 1, article 121, 2013.

[31] M. M. Babu, N. M. Luscombe, L. Aravind, M. Gerstein, and S. A. Teichmann, "Structure and evolution of transcriptional regulatory networks," *Current Opinion in Structural Biology*, vol. 14, no. 3, pp. 283–291, 2004.

[32] M. Nitta, T. Furukawa, Y. Shida et al., "A new $Zn(II)_2Cys_6$-type transcription factor BglR regulates β-glucosidase expression in *Trichoderma reesei*," *Fungal Genetics and Biology*, vol. 49, no. 5, pp. 388–397, 2012.

[33] T. J. Schoberle, C. K. Nguyen-Coleman, J. Herold et al., "A novel C2H2 transcription factor that regulates gliA expression interdependently with GliZ in *Aspergillus fumigatus*," *PLoS Genetics*, vol. 10, no. 5, Article ID e1004336, 2014.

[34] A. J. Clutterbuck, "A mutational analysis of conidial development in *Aspergillus nidulans*," *Genetics*, vol. 63, no. 2, pp. 317–327, 1969.

[35] H. Son, Y.-S. Seo, K. Min et al., "A phenome-based functional analysis of transcription factors in the cereal head blight fungus,

Fusarium graminearum," *PLoS Pathogens*, vol. 7, no. 10, Article ID e1002310, 2011.

[36] C. S. Gronover, D. Kasulke, P. Tudzynski, and B. Tudzynski, "The role of G protein alpha subunits in the infection process of the gray mold fungus *Botrytis cinerea*," *Molecular Plant-Microbe Interactions*, vol. 14, no. 11, pp. 1293–1302, 2001.

[37] D. Ribitsch, E. H. Acero, A. Przylucka et al., "Enhanced cutinase-catalyzed hydrolysis of polyethylene terephthalate by covalent fusion to hydrophobins," *Applied and Environmental Microbiology*, vol. 81, no. 11, pp. 3586–3592, 2015.

[38] G. C. Segers, J. C. Regier, and D. L. Nuss, "Evidence for a role of the regulator of G-protein signaling protein CPRGS-1 in Gα subunit CPG-1-mediated regulation of fungal virulence, conidiation, and hydrophobin synthesis in the chestnut blight fungus *Cryphonectria parasitica*," *Eukaryotic Cell*, vol. 3, no. 6, pp. 1454–1463, 2004.

[39] B. Črešnar and Š. Petrič, "Cytochrome P450 enzymes in the fungal kingdom," *Biochimica et Biophysica Acta*, vol. 1814, no. 1, pp. 29–35, 2011.

[40] P. K. Mukherjee, B. A. Horwitz, and C. M. Kenerley, "Secondary metabolism in *Trichoderma*—a genomic perspective," *Microbiology*, vol. 158, no. 1, pp. 35–45, 2012.

[41] D. Hoffmeister and N. P. Keller, "Natural products of filamentous fungi: enzymes, genes, and their regulation," *Natural Product Reports*, vol. 24, no. 2, pp. 393–416, 2007.

[42] N. P. Keller, G. Turner, and J. W. Bennett, "Fungal secondary metabolism—from biochemistry to genomics," *Nature Reviews Microbiology*, vol. 3, no. 12, pp. 937–947, 2005.

[43] A. Osbourn, "Secondary metabolic gene clusters: evolutionary toolkits for chemical innovation," *Trends in Genetics*, vol. 26, no. 10, pp. 449–457, 2010.

Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 1648030, 8 pages
http://dx.doi.org/10.1155/2016/1648030

Research Article

Transcriptome Analysis of HepG2 Cells Expressing ORF3 from Swine Hepatitis E Virus to Determine the Effects of ORF3 on Host Cells

Kailian Xu,[1] Shiyu Guo,[1] Tianjing Zhao,[1] Huapei Zhu,[1] Hanwei Jiao,[1] Qiaoyun Shi,[1] Feng Pang,[1] Yaying Li,[1] Guohua Li,[1] Dongmei Peng,[1] Xin Nie,[1] Ying Cheng,[1] Kebang Wu,[1] Li Du,[1] Ke Cui,[2] Wenguang Zhang,[3] and Fengyang Wang[1]

[1]*College of Agriculture, Hainan University, Hainan Key Lab of Tropical Animal Reproduction & Breeding and Epidemic Disease Research, Animal Genetic Engineering Key Lab of Haikou, Haikou 570228, China*
[2]*Modern Agriculture Risk Warning and Prevention and Control Center of Hainan Province, Haikou 571100, China*
[3]*College of Animal Science, Inner Mongolia Agricultural University, Inner Mongolia Key Laboratory of Animal Genetics, Breeding and Reproduction, Inner Mongolia, Hohhot 010018, China*

Correspondence should be addressed to Wenguang Zhang; atcgnmbi@hotmail.com and Fengyang Wang; fywang68@163.com

Received 2 March 2016; Revised 4 May 2016; Accepted 10 May 2016

Academic Editor: Hongwei Wang

Hepatitis E virus- (HEV-) mediated hepatitis has become a global public health problem. An important regulatory protein of HEV, ORF3, influences multiple signal pathways in host cells. In this study, to investigate the function of ORF3 from the swine form of HEV (SHEV), high-throughput RNA-Seq-based screening was performed to identify the differentially expressed genes in ORF3-expressing HepG2 cells. The results were validated with quantitative real-time PCR and gene ontology was employed to assign differentially expressed genes to functional categories. The results indicated that, in the established ORF3-expressing HepG2 cells, the mRNA levels of CLDN6, YLPM1, APOC3, NLRP1, SCARA3, FGA, FGG, FGB, and FREM1 were upregulated, whereas the mRNA levels of SLC2A3, DKK1, BPIFB2, and PTGR1 were downregulated. The deregulated expression of CLDN6 and FREM1 might contribute to changes in integral membrane protein and basement membrane protein expression, expression changes for NLRP1 might affect the apoptosis of HepG2 cells, and the altered expression of APOC3, SCARA3, and DKK1 may affect lipid metabolism in HepG2 cells. In conclusion, ORF3 plays a functional role in virus-cell interactions by affecting the expression of integral membrane protein and basement membrane proteins and by altering the process of apoptosis and lipid metabolism in host cells. These findings provide important insight into the pathogenic mechanism of HEV.

1. Introduction

Hepatitis E infection, caused by enterically transmitted hepatitis E virus (HEV), is a public health problem worldwide, particularly in developing countries such as China and India [1]. HEV infection is associated with a mortality rate of 0.2–1% in the general population, with an increased incidence and severity in pregnant women, in which mortality rates of 15–20% are observed [2–4]. As a zoonotic disease, swine infected with swine hepatitis E virus (SHEV) are the major reservoir of human HEV contamination [5, 6].

The HEV genome contains three open reading frames (ORFs), which encode ORF1, ORF2, and ORF3. ORF3 is a small molecular protein that influences multiple signal pathways in host cells [4]. In our previous study, the down-regulation of microRNAs miR-221 and miR-222 in ORF3-expressing HEK 293 cells was observed, and miR-221 and miR-222 were found to directly regulate $p27^{kip1}$. Our findings suggested that ORF3 might be involved in the proliferation of the host cells [7].

As one of the next-generation sequencing technologies, RNA-Seq can provide a complete snapshot of all of

the transcripts present at a particular moment in the cell. RNA-Seq is superior to the oligonucleotide microarray approach that analyzes a selected number of previously defined transcripts. Based on RNA-Seq transcriptome analysis results and differential expression validation with quantitative real-time PCR (qRT-PCR), the differentially expressed genes (DEGs) of Huh-7 cells transfected with the HEV replicon were obtained. These included some innate immune response associated genes and some cell survival and metabolism associated genes; however, the functional roles of ORF3 were not elucidated [8]. In our study, RNA-Seq-based screening and further qRT-PCR validation were performed to identify the DEGs in ORF3-expressing HepG2 cells, and the DEGs identified were assigned functions by gene ontology. Our findings suggested that ORF3 functions by affecting the biological processes, cellular components, and molecular functions within the host cells.

2. Materials and Methods

2.1. Cell Lines and Plasmids. HepG2 cells were purchased from the Cell Bank of the Chinese Academy of Sciences (Shanghai, China) and were grown at 37°C in Dulbecco's minimum essential medium (DMEM) (Gibco BRL, Carlsbad, CA, USA) containing 10% heat-inactivated fetal bovine serum (FBS) (Gibco BRL), supplemented with penicillin (100 U/mL; Gibco BRL) and streptomycin (100 μg/mL; Gibco BRL, USA). The recombinant plasmid, pEGFP-ORF3, which expresses EGFP-ORF3 fusion protein, was constructed in our previous study [5].

2.2. Preparation of Recombinant Lentivirus. The upstream primer (5′-GCGGCGTTAATTAAGCCACCATGGCGA-TGCCACCATGCG-3′) containing a *Pac*I site and the downstream primer (5′-ATTATTGGCGCGCCTCAGCGG-CGAAGCCCCAGCT-3′) containing an *Asc*I site were used to amplify the ORF3 fragment from pEGFP-ORF3. The obtained ORF3 fragment was ligated into the lentiviral vector, pLenti6.3-MCS-IRES-GFP, and then digested with *Pac*I and *Asc*I. The recombinant lentivirus was designated pLenti6.3-ORF3-IRES-EGFP. The recombinant lentivirus was prepared as previously described and the titers of the recombinant lentivirus were determined [7]. pLenti6.3-MCS-IRES-GFP was also used for the preparation of recombinant lentivirus, and this was used as a negative control in the experiments.

2.3. Establishment of SHEV ORF3-Expressing HepG2 Cells. As previously described, HepG2 cells were infected with the recombinant lentivirus at a multiplicity of infection (MOI) of 10 [5]. The expression of enhanced green fluorescence protein (EGFP) was observed by fluorescence microscopy (X71; Olympus, Tokyo, Japan). The stable cell lines were obtained as previously described [5].

2.4. Flow Cytometry Analysis. As previously described, FACSCalibur flow cytometer (Becton Dickinson, San Jose, CA, USA) was used to determine the percentage of fluorescent cells population and the meant fluorescent intensity of the stable cells lines [5].

2.5. Western Blot Analysis. The total protein from ORF3-expressing HepG2 cells, EGFP (only)-expressing HepG2 cells, and HepG2 cells was harvested. SDS-PAGE and western blot analysis were performed as previously described [5]. The primary antibodies used were rabbit polyclonal antibody against ORF3, which was produced as described in our previous study [5], and rabbit polyclonal anti-GAPDH antibody (Cell Signaling Technology, Beverly, MA, USA). The secondary antibody was HRP-labeled goat anti-rabbit IgG (Santa Cruz Biotechnology, Santa Cruz, CA, USA).

2.6. mRNA Library Construction and Sequencing. Following the manufacturer's procedure, TRIzol reagent (Invitrogen, Carlsbad, CA, USA) was used to extract the total RNA from ORF3-expressing HepG2 cells, EGFP (only)-expressing HepG2 cells, and HepG2 cells. The quantity and purity of the total RNA were analyzed using Bioanalyzer 2100 and the RNA 6000 Nano LabChip Kit (Agilent Technologies, Santa Clara, CA, USA).

Next, 10 μg of RNA from specific cell lines (ORF3-expressing HepG2 cells, EGFP (only)-expressing HepG2 cells, and HepG2 cells) was exposed to poly-T oligoattached magnetic beads (Invitrogen, Carlsbad, CA, USA) to isolate poly(A) mRNA. Following purification, the mRNA was fragmented into small pieces using fragmentation buffer and the cleaved RNA fragments were reverse-transcribed to construct a cDNA library using the mRNA-Seq sample preparation kit (Illumina, San Diego, CA, USA), according to the manufacturer's instructions. Then, paired-end sequencing was performed on an Illumina 2000/2500 sequence platform (LC Sciences, Houston, TX, USA), following the manufacturer's protocol.

2.7. Mapping, Normalization, and Calculation of the RPKM. Clean reads were obtained by removing the low quality reads from the raw reads. The quality of the reads was classified according to the following criteria: (1) containing sequencing adaptors, (2) ratio of N (without valid base information) above 5%; and (3) ratio of nucleotides [Q value (quality score) is lower than 10] above 20%. Then, as previously described, clean reads from specific cell lines were aligned to the genome database UCSC (http://genome.ucsc.edu/) using the Tophat package [9].

Based on the results of Tophat, the fragment per kilobase of exon per million fragments mapped (FPKM) value was used to normalize the number of fragments, as previously described. Cufflinks were used to de novo assemble the transcriptome and comerge and annotate the sequence fragments. The DEGs, their corresponding attributes, fold changes (in $\log_2$ scale), p values, and FDR (false discovery rate corrected p values) were obtained [10, 11]. The significance of the gene expression difference was determined as "yes" if the false discovery rate (Q value) was <0.05. Only the comparisons with Q value less than 0.01 and a status marked as "OK"

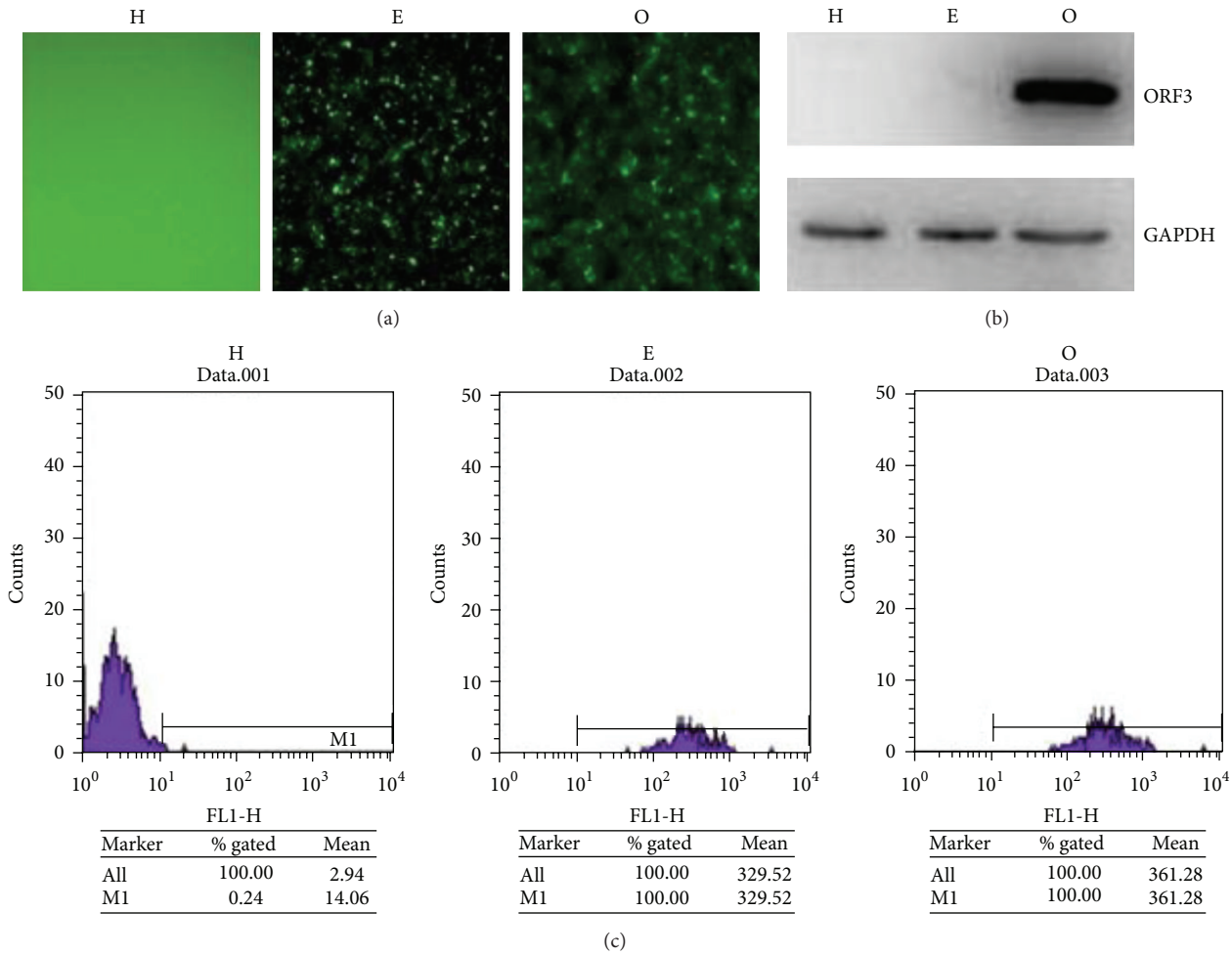


Figure 1: Characterization of ORF3-expressing HepG2 cells. (a) Fluorescence observation of H, E, and O cells. (b) Western blot results indicated that ORF3 protein was expressed in O cells. (c) Flow cytometry analysis results indicated that EGFP protein was expressed in O cells and E cells.

in the Cuffdiff output were regarded as showing differential expression [12].

2.8. Gene Ontology (GO) of DEGs. As previously described, GO was performed to analyze the DEGs [13]. GO terms with $p < 0.05$ were considered significantly enriched among the DEGs [9].

2.9. qRT-PCR for Differential Expression Validation. To validate the differential expression of genes, the specific primers were designed and qRT-PCR was performed as previously described [7].

2.10. Statistical Analysis. The statistical significance of the differences between the data from the experimental groups and the control was analyzed using Student's t-test and one-way ANOVA. $p < 0.05$ was considered to represent significant differences, and $p < 0.01$ was considered as highly different [7].

3. Results and Discussion

3.1. Identification of SHEV ORF3-Expressing HepG2 Cells. After the recombinant lentivirus vector, pLenti6.3-ORF3-IRES-EGFP, had been constructed and confirmed with DNA sequencing, recombinant lentivirus carrying ORF3 was prepared, titrated, and used to infect HepG2 cells. After blasticidin selection, ORF3-expressing HepG2 cells were obtained and designated as O, and EGFP (only)-expressing HepG2 cells were obtained and designated as E. HepG2 cells were used as the black control and designated as H. The expression of EGFP protein was observed in O and E cells (Figures 1(a) and 1(c)). Western blotting results revealed that there was a specific band at the expected molecular weight for ORF3 protein in O cells; however, no expression of ORF3 was detected in E and H cells (Figure 1(b)).

3.2. Sequencing and Mapping of the SHEV ORF3-Expressing HepG2 Cell Transcriptome. Using the Illumina paired-end RNA-Seq approach, the cDNA libraries of H, E, and

Table 1: Numbers of reads in the reference genome.

Sample	E	H	O
Valid reads	46,413,284	45,043,398	39,935,156
Mapped reads	32,068,726 (69.09%)	29,911,216 (66.41%)	27,147,943 (67.98%)
Unique mapped reads	31,520,409 (67.91%)	29,386,695 (65.24%)	26,670,680 (66.78%)
Multimapped reads	548,317 (1.18%)	524,521 (1.16%)	477,263 (1.20%)
PE mapped reads	13,909,419 (29.97%)	13,120,894 (29.13%)	11,735,308 (29.39%)
Exon	98.27%	98.05%	97.78%

O cells were sequenced. The results were uploaded into NCBI Sequence Read Archive (SRA) (accession number: SRP073936). The average insert size for the paired-end libraries was 300 bp (±50 bp). In total, 132,757,090 paired-end reads of 2 × 100 bp length were acquired. The total read length of the three samples was 16.59 gigabases (Gb). After removing the low quality reads from the raw reads, a total of 16.42 Gbp of cleaned, paired-end reads were produced, with a Q20 of over 90% (Table 1).

Alignment of the sequence reads against the reference genome yielded about 70% aligned reads across the three samples, for which the ratio of pair reads is about 30% and the ratio of unique map was about 70% and of which about 98% were located within annotated exons. Multiposition matched reads (<10%) were excluded from further analyses. The distribution of the density of the sequence was normal. These data satisfied the requirements of further gene expression level analyses.

3.3. Differential Expression Analysis. Visualization of the data in Venn diagrams indicated that the number of DEGs in O cells was 18. In O cells, the mRNA levels of claudin-6 (CLDN6), FRAS1-related extracellular matrix 1 (FREM1), scavenger receptor class A member 3 (SCARA3), fibrinogen (FGG), fibrinogen alpha (FGA), fibrinogen beta (FGB), apolipoprotein C3 (APOC3), YLP motif-containing protein 1 (YLPM1), and nucleotide-binding oligomerization domain-like receptor with pyrin domain protein 1 (NLRP1) were upregulated, while the mRNA levels of cytokeratin 19 (KRT19), BPI fold containing family B, member 2 (BPIFB2), sulforaphane (SFN), activated leukocyte cell adhesion molecule (ALCAM), solute carrier family 22 member 3 (SLC2A3), prostaglandin reductase 1 (PTGR1), Dickkopf-related protein 1 (DKK1), S100 calcium binding protein A4 (S100A4), and nuclear protein 1 (NUPR1) were downregulated (Figures 2(a) and 2(b), Table 2).

To further confirm the RNA-Seq data, specific primers were designed (Table 3), and the qRT-PCR was performed using GAPDH as an internal control. The results confirmed that the mRNA levels of CLDN6, YLPM1, APOC3, NLRP1, SCARA3, FGA, FGG, FGB, and FREM1 were upregulated and the mRNA levels of SLC2A3, DKK1, BPIFB2, and PTGR1 were downregulated (Figure 3).

3.4. Assignment of DEGs. The 13 validated DEGs in O cells were assigned into the following categories: biological process, cellular components, and molecular function ($p < 0.05$) (Figure 4; see S1 in Supplementary Material available online at http://dx.doi.org/10.1155/2016/1648030). Among them, FGG, FGA, and FGB were assigned into all 25 GO cases for category biological processes, all 18 GO cases for cellular components, and 4 GO cases for molecular function (S1).

Table 2: Genes found to be significantly differentially expressed between cell types.

Gene short name	p value	E FPKM	H FPKM	O FPKM
FGG	$2.252E-37$	205.24	190.90	465.10
FGB	$1.2802E-10$	43.87	44.37	110.70
FGA	$2.6364E-24$	147.94	145.66	328.38
APOC3	$1.71E-43$	249.10	272.37	584.34
SLC2A3	$1.3633E-31$	522.47	569.07	249.24
DKK1	$2.97E-44$	781.57	719.76	320.23
KRT19	$3.1275E-46$	415.28	205.75	83.07
BPIFB2	$5.7055E-35$	295.48	128.29	59.84
CLDN6	$5.7466E-07$	36.04	24.00	72.68
SFN	0.000023461	30.45	12.06	3.77
FREM1	0.000048529	18.29	13.07	42.37
ALCAM	0.00010227	10.00	18.79	0.62
YLPM1	0.0010757	19.50	21.74	44.22
SCARA3	0.0011159	2.53	1.08	12.00
NLRP1	0.0078786	1.51	3.31	10.90
S100A4	0.0079767	17.17	14.16	2.80
PTGR1	0.010836	30.77	31.65	12.99
NUPR1	0.022976	16.63	14.24	3.92

FPKM: fragment per kilobase of exon per million mapped fragments.

The cellular component category includes membranes, organelles, and proteins. In O cells, two validated DEGs, CLDN6 and FREM1, were assigned into the cellular component category (S1). CLDN6 encodes an integral membrane protein that is one of the entry cofactors for hepatitis C virus, which was assigned into GO: 0005886-plasma membrane [14]. FREM1 encodes a basement membrane protein, which was assigned into GO: 0044421-extracellular region part and GO: 0005576-extracellular region [15]. These results suggested that the expression of ORF3 induced the deregulated expression of two cellular components.

Biological processes include many chemical reactions and other events that result in chemical transformation including metabolism and homeostasis. In O cells, NLRP1 was assigned into the GO category of biological processes (S1). The NLRP1

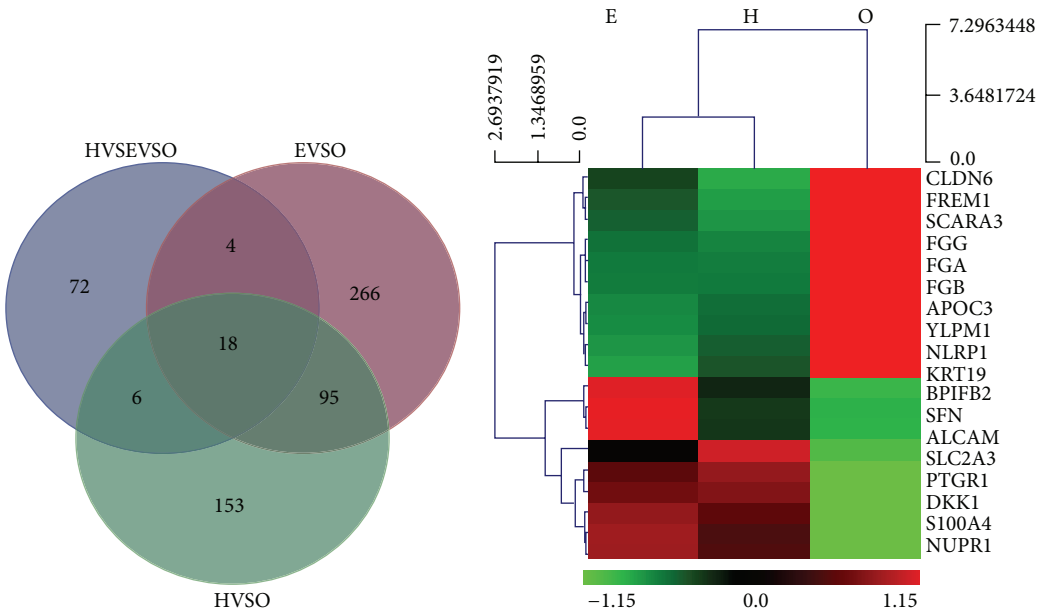


Figure 2: The DEGs obtained from H, E, and O cells by RNA-Seq. (a) The Venn diagrams indicated that the number of the significantly DEGs in O cells was 18. (b) Heat-maps indicated that, compared with that in H and E cells, in O cells, the mRNA levels of CLDN6, FREM1, SCARA3, FGG, FGA, FGB, APOC3, YLPMA1, and NLRP1 were upregulated, while KRT19, BPIFB2, SFN, ALCAM, SLC2A3, PTGR1, DKK1, S100A4, and NUPR1 were downregulated. Red indicates upregulation and green indicates downregulation.

Table 3: Primers for qRT-PCR validation.

Gene short name	Forward primer sequence 5′-3′	Reverse primer sequence 5′-3′
KRT19	GGTCATGGCCGAGCAGAA	TTCAGTCCGGCTGGTGAAC
DKK1	AGAAAAGGCTCTCATGGACTAGAAAT	CCGGCAAGACAGACCTTCTC
APOC3	TGGCTGCCTGAGACCTCAAT	AGGAGCTCGCAGGATGGATA
FGG	TGGTTGGTGGATGAACAAGTGT	TGCCACCTTGGTAATAAACTCCAT
PIFB2	CTGGATGTGGTAGTGAACTTGAGACT	ACGTGGTCCCCTGAAGCTT
SLC2A3	GGCACACGGTGCAGATAGATC	GCAGGCTCGATGCTGTTCAT
FGA	CAGATGAGGCCGGAAGTGA	GATTTAGCATGGCCTCTCTTGGT
FGB	CAGCCGGTAATGCCCTCAT	TGTGAATGGTCATGGTCCTGTT
CLDN6	CTTGGATGATGGAGCCAAAGA	TGGCTTCTAAGATGGGCATGT
SFN	GCAGGCCGAACGCTATGA	TCCACGGCGCCTTTCA
FREM1	ACCTGGGCAACCTTGTAACTGTA	TGGTCGTTCAAACCTATCCAAA
ALCAM	CAATGCCCCAAACTTTCTCATAA	TGTCCCCAATCTTCACAAGCT
YLPM1	GGAAACTGCACCTCGTCACA	GCAGCATCTTGCAGCAAAGA
SCARA3	CCCTGAGAAGTTCAACATTTATTTCTT	GGGCAGAGGCAAGGATGAAT
NLRP1	CCCTCTATCGGCGTCTATCTGT	GCTCTTACCGTCTCTTATTCAGCAT
S100A4	CGCCAGTGGGCACTTTTTT	CAGCATCAAGCACGTGTCTGA
PTGR1	AAGAAATTTGGAAGGATTGCCATA	GAAGTGGGCCGGTTCTGTTA
NUPR1	GGGTGGCAGCAACAATAAATAGA	GGATGAACACACACCCAAGCT

gene encodes a member of the Ced-4 family of apoptosis proteins. NLRP3/NLRP1 inflammasome-mediated caspase-1 activation with subsequent IL-1 secretion is essential for the subsequent bifurcation to downregulated proinflammatory cytokines and upregulated bacterial killing [16], and this gene was assigned into two GO cases within the category of biological processes (GO: 0006950 and GO: 0050896-response to stress and stimulus, resp.). These results suggested that the expression of ORF3 affects the apoptosis of HepG2 cells.

The liver is the central regulatory organ of lipid pathways. APOC3 specifically modulates the metabolism of

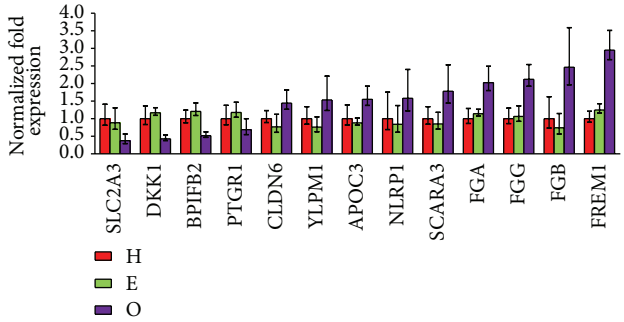


Figure 3: Validation of the RNA-Seq data by qRT-qPCR. The results confirmed the upregulation of the mRNA levels of CLDN6, YLPM1, APOC3, NLRP1, SCARA3, FGA, FGG, FGB, and FREM1 and the downregulation of the mRNA levels of SLC2A3, DKK1, BPIFB2, and PTGR1.

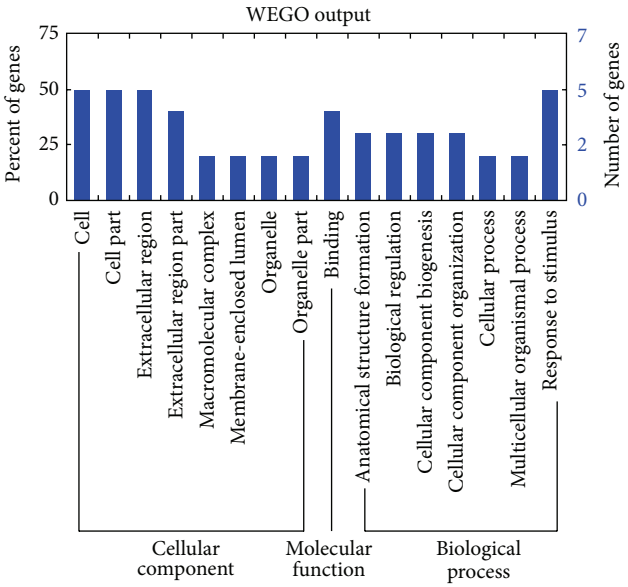


Figure 4: GO classifications of the DEGs in O cells.

triglyceride-rich lipoproteins and may contribute to the development of hyperlipidemia and other lipoprotein abnormalities in humans [17–19]. In our study, APOC3 was one of the 13 validated DEGs in O cells. APOC3 not only was assigned into three GO cases within the category "cellular components" (GO: 0044421-extracellular region part; GO: 0005576-extracellular region; GO: 0005615-extracellular space) and five GO cases of biological process (GO: 0065003-macromolecular complex assembly; GO: 0043933-macromolecular complex subunit organization; GO: 0022607-cellular component assembly; GO: 0044085-cellular component biogenesis; GO: 0065008-regulation of biological quality), but was also assigned into two GO cases within the category "molecular function" (GO: 0005102-receptor binding and GO: 0070325-lipoprotein receptor binding) (S1).

Interestingly, SCARA3 and DKK1 are also related to the lipid metabolism. As a member of the scavenger receptor family, SCARA3 protects cells by scavenging reactive oxygen species and other harmful products of oxidation [20]. SCARA3 binds to polyanionic ligands, which are an important source of fatty acids for macrophages [21]. SCARA3 was assigned into three GO cases within the category "biological processes" (GO: 0042221-response to chemical stimulus; GO: 0006950-response to stress; GO: 0050896-response to stimulus) (S1).

DKK1, a canonical Wnt/β-catenin pathway antagonist, is closely associated with adipogenesis [22]. DKK1 regulates certain aspects of placental lipid metabolism through the WNT signaling pathway [23]. DKK1 not only was assigned into two GO cases within the category "cellular components" (GO: 0005576-extracellular region; GO: 0005886-plasma membrane), but also was assigned into two GO cases within the category "molecular function" (GO: 0005102-receptor binding; GO: 0070325-lipoprotein receptor binding), as seen for APOC3 (S1).

From these data, it can be concluded that the expression of ORF3 induces the upregulation of the mRNA levels of CLDN6, YLPM1, APOC3, NLRP1, SCARA3, FGA, FGG, FGB, and FREM1 and the downregulation of the mRNA levels of SLC2A3, DKK1, BPIFB2, and PTGR1. Among these changes, the deregulated expression of CLDN6, APOC3, NLRP1, SCARA3, FGA, FGG, FGB, FREM1, SLC2A3, DKK1, and PTGR1 might contribute to the deregulation of integral membrane protein and basement membrane protein and affect the apoptosis and the lipid metabolism of HepG2 cells.

In a recent study, the RNA-Seq approach was used to explore the cellular pathway alterations during virus infection. Changes in the transcriptomes of primary bovine cells following infection with either wild type Schmallenberg virus (SBV) or SBV with a mutant lacking the nonstructural protein NSs (SBVdelNSs) were analyzed. The results suggested that nonstructural protein not only was effective in shutting down genes of the host innate immune system, but also affected a number of possible antiviral factors [24]. Because of the lack of a cell culture system and a suitable animal model, the pathogenesis of hepatitis E is poorly understood. In this study, HepG2 cells, which are a suitable in vitro model system for the study of HEV, were used as the target cells for ORF3 overexpression. RNA-Seq-based screening and further qRT-PCR validation were performed to identify the DEGs in ORF3-expressing HepG2 cells.

Correlation analysis results between the RNA-Seq and qRT-PCR data indicated that 13 of the 18 DEGs detected by RNA-Seq in O cells were validated by qRT-PCR. These results indicated that the experimental approach was effective. The five nonvalidated DEGs may have been the false-positive results generated by RNA-Seq.

As a standardized gene function classification system, GO describes the properties of genes and their products. In our study, the Database for Annotation, Visualization and Integrated Discovery (DAVID) software was used to obtain the GO ID, and Web Gene Ontology Annotation Plot (WEGO) software was used to plot the GO annotation results (http://wego.genomics.org.cn/cgi-bin/wego/index.pl).

The liver is the central regulatory organ of lipid pathways. Our findings confirmed that HEV infection causes alterations

in lipid metabolism. Among 13 validated DEGs, APOC3, SCARA3, and DKK1 played a role in lipid metabolism. In HepG2 cells, the expression of ORF3 causes the deregulation of lipid metabolism, potentially resulting in cell injury.

The expression of ORF3 also resulted in the deregulation of CLDN6, NLRP1, FGA, FGG, FGB, FREM1, SLC2A3, and PTGR1, potentially resulting in cell injury as a result of changes in biological processes, cellular components, and/or the molecular function of HepG2 cells.

To our knowledge, this is the first report of the altered expression of CLDN6, YLPM1, APOC3, NLRP1, SCARA3, FGA, FGG, FGB, FREM1, SLC2A3, DKK1, BPIFB2, and PTGR1 in HEV-infected cells. Our findings provide insight into the critical events that take place during HEV infection.

4. Conclusions

ORF3 protein is a key regulatory protein of SHEV. Here, for the first time, we report the upregulation of the mRNA levels of CLDN6, YLPM1, APOC3, NLRP1, SCARA3, FGA, FGG, FGB, and FREM1 and the downregulation of the mRNA levels of SLC2A3, DKK1, BPIFB2, and PTGR1 in the established ORF3-expressing HepG2 cells. The deregulated expression of these 13 genes may lead to changes in the deregulation of integral membrane and basement membrane proteins and may affect the processes of lipid metabolism and apoptosis in human cells. These findings provide insight into the infection processes mediated by HEV and may be valuable in the development of future therapeutic strategies.

Abbreviations

HEV: Hepatitis E virus
SHEV: Swine hepatitis E virus
ORF3: Open reading frame 3
qRT-PCR: Quantitative real-time RT-PCR
GO: Gene ontology
DEGs: Differentially expressed genes
NGS: Next-generation sequencing
DMEM: Dulbecco's minimum essential medium
FBS: Heat-inactivated fetal bovine serum
MOI: Multiplicity of infection
FCM: Flow cytometry
SDS-PAGE: Sodium dodecyl sulfate polyacrylamide gel electropheresis
EGFP: Enhanced green fluorescent protein
HRP: Horseradish peroxidase
ANOVA: Analysis of variance
CLDN6: Claudin-6
APOC3: Apolipoprotein C3
NLRP1: Nucleotide-binding oligomerization domain-like receptor with pyrin domain protein 1
FREM1: FRAS1-related extracellular matrix 1
SCARA3: Scavenger receptor class A member 3
FGG: Fibrinogen
FGA: Fibrinogen alpha
FGB: Fibrinogen beta
YLPM1: YLP motif-containing protein 1
KRT19: Cytokeratin 19
BPIFB2: BPI fold containing family B, member 2
SFN: Sulforaphane
ALCAM: Activated leukocyte cell adhesion molecule
SLC2A3: Solute carrier family 22 member 3
PTGR1: Prostaglandin reductase 1
DKK1: Dickkopf-related protein 1
S100A4: S100 calcium binding protein A4
NUPR1: Nuclear protein 1
FPKM: Fragment per kilobase of exon model per million mapped fragments
DAVID: Database for Annotation, Visualization and Integrated Discovery
WEGO: Web Gene Ontology Annotation Plot
SBV: Schmallenberg virus.

Competing Interests

All authors declare that there are no competing interests regarding the publication of this paper.

Authors' Contributions

Kailian Xu, Shiyu Guo, and Tianjing Zhao contributed equally to this work.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (no. 31360618) and the Key Science and Technology Project of Hainan (no. ZDXM2014026). The authors thank Getinet M. Tarekegn for his efforts to improve the language quality of the paper.

References

[1] I. M. Sayed, A.-S. Vercouter, S. F. Abdelwahab, K. Vercauteren, and P. Meuleman, "Is hepatitis E virus an emerging problem in industrialized countries?" *Hepatology*, vol. 62, no. 6, pp. 1883–1892, 2015.

[2] M. S. Khuroo, M. R. Teli, S. Skidmore, M. A. Sofi, and M. I. Khuroo, "Incidence and severity of viral hepatitis in pregnancy," *The American Journal of Medicine*, vol. 70, no. 2, pp. 252–255, 1981.

[3] U. Navaneethan, M. Al Mohajer, and M. T. Shata, "Hepatitis E and pregnancy: understanding the pathogenesis," *Liver International*, vol. 28, no. 9, pp. 1190–1199, 2008.

[4] V. Chandra, S. Taneja, M. Kalia, and S. Jameel, "Molecular biology and pathogenesis of hepatitis E virus," *Journal of Biosciences*, vol. 33, no. 4, pp. 451–464, 2008.

[5] T. Liu, M. Lei, H. Jiao et al., "RNA interference induces effective inhibition of mRNA accumulation and protein expression of SHEV ORF3 gene in vitro," *Current Microbiology*, vol. 62, no. 5, pp. 1355–1362, 2011.

[6] S. Rogée, M. Le Gall, P. Chafey et al., "Quantitative proteomics identifies host factors modulated during acute hepatitis E virus infection in the swine model," *Journal of Virology*, vol. 89, no. 1, pp. 129–143, 2015.

[7] Y. Cheng, L. Du, Q. Shi et al., "Identification of miR-221 and -222 as important regulators in genotype IV swine hepatitis e virus ORF3-expressing HEK 293 cells," *Virus Genes*, vol. 47, no. 1, pp. 49–55, 2013.

[8] N. Jagya, S. P. K. Varma, D. Thakral, P. Joshi, H. Durgapal, and S. K. Panda, "RNA-Seq based transcriptome analysis of hepatitis E virus (HEV) and hepatitis B virus (HBV) replicon transfected Huh-7 cells," *PLoS ONE*, vol. 9, no. 2, Article ID e87835, 2014.

[9] X. Cui, Y. Hou, S. Yang et al., "Transcriptional profiling of mammary gland in Holstein cows with extremely different milk protein and fat percentage using RNA sequencing," *BMC Genomics*, vol. 15, article 226, 2014.

[10] 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.

[11] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-Seq," *Nature Methods*, vol. 5, no. 7, pp. 621–628, 2008.

[12] F. Rapaport, R. Khanin, Y. Liang et al., "Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data," *Genome Biology*, vol. 14, no. 9, article R95, 2013.

[13] M. D. Young, M. J. Wakefield, G. K. Smyth, and A. Oshlack, "Gene ontology analysis for RNA-seq: accounting for selection bias," *Genome Biology*, vol. 11, no. 2, article r14, 2010.

[14] A. Arabzadeh, T.-C. Troy, and K. Turksen, "Role of the Cldn6 cytoplasmic tail domain in membrane targeting and epidermal differentiation in vivo," *Molecular and Cellular Biology*, vol. 26, no. 15, pp. 5876–5887, 2006.

[15] P. Petrou, E. Pavlakis, Y. Dalezios, and G. Chalepakis, "Basement membrane localization of Frem3 is independent of the Fras1/Frem1/Frem2 protein complex within the sublamina densa," *Matrix Biology*, vol. 26, no. 8, pp. 652–658, 2007.

[16] M. Hedl and C. Abraham, "NLRP1 and NLRP3 inflammasomes are essential for distinct outcomes of decreased cytokines but enhanced bacterial killing upon chronic Nod2 stimulation," *American Journal of Physiology—Gastrointestinal and Liver Physiology*, vol. 304, no. 6, pp. G583–G596, 2013.

[17] M. H. Hofker, "APOC3 null mutation affects lipoprotein profile. APOC3 deficiency: from mice to man," *European Journal of Human Genetics*, vol. 18, no. 1, pp. 1–2, 2010.

[18] T. I. Pollin, C. M. Damcott, H. Shen et al., "A null mutation in human *APOC3* confers a favorable plasma lipid profile and apparent cardioprotection," *Science*, vol. 322, no. 5908, pp. 1702–1705, 2008.

[19] M. C. Jong and L. M. Havekes, "Insights into apolipoprotein C metabolism from transgenic and gene-targeted mice," *International Journal of Tissue Reactions*, vol. 22, no. 2-3, pp. 59–66, 2000.

[20] H.-J. Han, T. Tokino, and Y. Nakamura, "CSR, a scavenger receptor-like protein with a protective role against cellular damage caused by UV irradiation and oxidative stress," *Human Molecular Genetics*, vol. 7, no. 6, pp. 1039–1046, 1998.

[21] J. E. Murphy, P. R. Tedbury, S. Homer-Vanniasinkam, J. H. Walker, and S. Ponnambalam, "Biochemistry and cell biology of mammalian scavenger receptors," *Atherosclerosis*, vol. 182, no. 1, pp. 1–15, 2005.

[22] Y. Zhang, C. Ge, L. Wang et al., "Induction of DKK1 by ox-LDL negatively regulates intracellular lipid accumulation in macrophages," *FEBS Letters*, vol. 589, no. 1, pp. 52–58, 2015.

[23] R. S. Strakovsky and Y.-X. Pan, "A decrease in DKK1, a WNT inhibitor, contributes to placental lipid accumulation in an obesity-prone rat model," *Biology of Reproduction*, vol. 86, no. 3, article 81, 2012.

[24] A.-L. Blomström, Q. Gu, G. Barry et al., "Transcriptome analysis reveals the host response to Schmallenberg virus in bovine cells and antagonistic effects of the NSs protein," *BMC Genomics*, vol. 16, no. 1, article 324, pp. 1–8, 2015.

Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 8642703, 21 pages
http://dx.doi.org/10.1155/2016/8642703

Research Article

Candidate SNP Markers of Chronopathologies Are Predicted by a Significant Change in the Affinity of TATA-Binding Protein for Human Gene Promoters

Petr Ponomarenko,[1] Dmitry Rasskazov,[2] Valentin Suslov,[2] Ekaterina Sharypova,[2] Ludmila Savinkova,[2] Olga Podkolodnaya,[2] Nikolay L. Podkolodny,[2] Natalya N. Tverdokhleb,[2,3] Irina Chadaeva,[2] Mikhail Ponomarenko,[2,3] and Nikolay Kolchanov[2,3]

[1]*Children's Hospital Los Angeles, University of Southern California, Los Angeles, CA 90027, USA*
[2]*Institute of Cytology and Genetics, Siberian Branch of Russian Academy of Sciences, Novosibirsk 630090, Russia*
[3]*Department of Natural Sciences, Novosibirsk State University, Novosibirsk 630090, Russia*

Correspondence should be addressed to Mikhail Ponomarenko; pon@bionet.nsc.ru

Received 4 March 2016; Revised 25 June 2016; Accepted 28 June 2016

Academic Editor: Rituraj Purohit

Variations in human genome (e.g., single nucleotide polymorphisms, SNPs) may be associated with hereditary diseases, their complications, comorbidities, and drug responses. Using Web service SNP_TATA_Comparator presented in our previous paper, here we analyzed immediate surroundings of known SNP markers of diseases and identified several candidate SNP markers that can significantly change the affinity of TATA-binding protein for human gene promoters, with circadian consequences. For example, rs572527200 may be related to asthma, where symptoms are circadian (worse at night), and rs367732974 may be associated with heart attacks that are characterized by a circadian preference (early morning). By the same method, we analyzed the 90 bp proximal promoter region of each protein-coding transcript of each human gene of the circadian clock core. This analysis yielded 53 candidate SNP markers, such as rs181985043 (susceptibility to acute Q fever in male patients), rs192518038 (higher risk of a heart attack in patients with diabetes), and rs374778785 (emphysema and lung cancer in smokers). If they are properly validated according to clinical standards, these candidate SNP markers may turn out to be useful for physicians (to select optimal treatment for each patient) and for the general population (to choose a lifestyle preventing possible circadian complications of diseases).

1. Introduction

Diurnal (circadian) oscillations of the expression level have been reliably identified in ~10000 genes of placental mammals [1]. The circadian clock of mammals is a system of self-sustained oscillators that function under the control of a central circadian pacemaker located in suprachiasmatic nuclei of the hypothalamus [2]. They synchronize all processes in living organisms, from gene transcription to behavior, thus ensuring their temporal adaptation to 24-hour days on Earth [1]. The minimal set of 12 genes—*CLOCK*, *ARNTL*, *ARNTL2*, *PER1*, *PER2*, *CRY1*, *CRY2*, *CSNK1E*, *CSNK1D*, *RORA*, *RORC*, and *NR1D1*—forms the core of the molecular genetic mechanism of the circadian clock, whose functioning is based on feedback relations among its components [3, 4] and on the relations of these genes with the entry points for external signals, which modulate parameters of the circadian clock in response to such external stimuli as light and food [5]. Via the retinohypothalamic tract, the central circadian oscillator imposes a rhythm on peripheral oscillators, which share their molecular genetic structure but work in each cell in accordance with their own specific rhythms of organs, tissues, and systems of tissues [1]. All these oscillators set the rhythm for a multitude of genes via expression of

tissue-specific transcription factors (short-term regulation) or chromatin remodeling (long-term regulation) [6, 7]. Indeed, transcriptomic studies have shown that genes that are subject to circadian control are characterized by overrepresentation of short GC-rich and TA-rich motifs for binding of transcription factors (e.g., TBP-binding motifs) [8, 9] in comparison with genome-wide average values of these parameters. In addition, an empirical study [10] revealed that CLOCK-ARNTL is a pioneer-like transcription factor that interacts with nucleosomes for rhythmic chromatin opening. Adjustment of the peripheral oscillation to the general circadian rhythm synchronizes the functioning of various systems of organs, whereas their desynchronosis can worsen or cause pathological changes in systems that are not interacting directly (e.g., autoimmune disorders may be caused by desynchronosis of the immune defense of the body from exotoxins and excretory/metabolizing systems dealing with analogous endotoxins [11]). Chronopharmacology is concerned with identification of circadian optima for diagnosis [12] and treatment [13, 14].

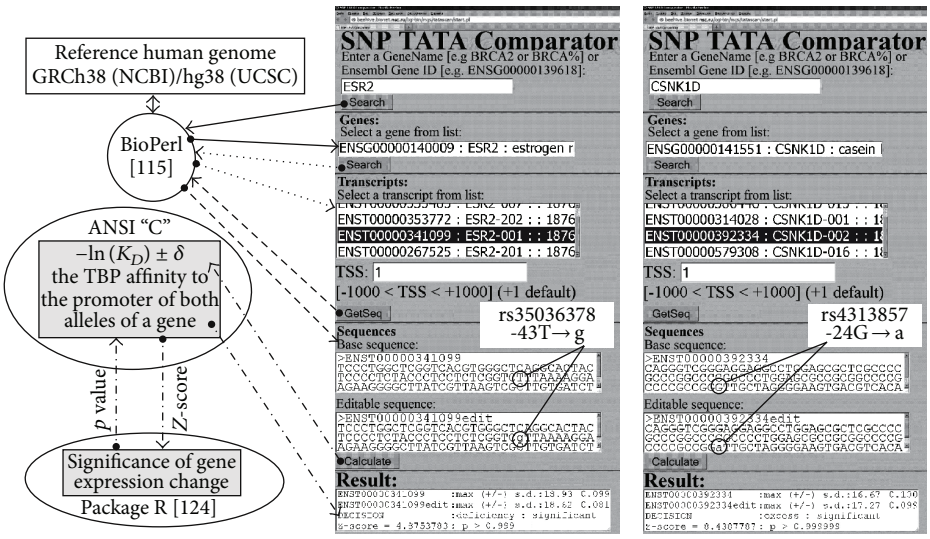
Experiments on genetic animal models have shown that, in addition to changes in parameters of the circadian clock (amplitude, a phasic response to external signals, or the period of free-flowing rhythm), the mutant animals develop such disorders as metabolic syndrome; disturbances in the system of gluconeogenesis or lipogenesis, in renal function, or in thermogenesis; and development of tumors [15, 16]. Furthermore, research in the field of genetic epidemiology uncovered associations of single nucleotide polymorphisms (SNPs) of circadian clock genes with a wide range of pathological states [17, 18]. A large number of such SNPs are located in noncoding regions of genes (these regions are responsible for regulation of expression). Functional annotation of regulatory SNPs and analysis of their manifestations at the level of gene expression are worthwhile tasks because many of such SNPs may be markers of clinical disorders.

During the "pregenomic era," association of an SNP with a disease used to be a lucky finding [19–22], whereas, now, in the "postgenomic era," identification of such associations is one of the goals of the 1000 Genomes Project [23]. Database dbSNP collects and ranks variants of each SNP by their prevalence [24]. The most frequent variant is entered into the reference human genome GRCh38 (NCBI) or hg38 (UCSC) (the terms used by the UCSC Genome Browser [25]) as an ancestral variant in the Ensembl database [26]. Minor alleles of SNPs in genes involved in a given pathological process can be found by means of the Web service *UCSC Genome Browser* [25], which visualizes a whole-genome map. Subsequent routine genotyping of these alleles in representative cohorts of patients and among healthy volunteers reveals (among minor alleles of SNPs) biomedical markers that are statistically significantly associated with the pathology in question [27]; this procedure takes up a lot of time and work. Computational (bioinformatic) analysis of many millions of unannotated SNPs from the 1000 Genomes Project may accelerate and cheapen the search for biomedical SNP markers.

Thus, the greatest success was achieved in the case of SNPs located in protein-coding regions of genes [28] because of the invariant (predictable) disruptions in the structure-function relations of the proteins encoded by these genes [29]. Moreover, advanced computer-based simulations of molecular dynamics and structures allowed researchers to predict in detail which SNPs would change the proteins. For example, molecular dynamics simulations provide deep analysis of the SNP-caused alterations in the amino acid arrangement that can affect the native three-dimensional atomic conformation of protein structure in order to estimate the most probable conformational modifications [30]. As an alternative/addition to molecular dynamics simulations for conformational sampling of proteins, so-called normal mode-based simulations guarantee multiscale modeling of protein conformational changes [31]. Besides, global minima of molecular docking for native and mutant structures can account for various substrate conformations and help identify an individual conformation with the most favorable binding energy [32]. In the case of drug resistance, computations of shape complementarity—between either widely used or promising new drugs and a binding pocket of a protein altered by an SNP—bring together the advantages of protein structure (or dynamics) simulations and the ability to dock one structure with another [33]. Finally, the alignment of multiple protein structures and/or sequences holds a key to the above calculations for comprehensive SNP analysis of protein-coding gene regions [34]. Meanwhile, the smallest progress was observed with regulatory SNPs because their manifestations may vary from cell to cell, from tissue to tissue, from patient to patient, and from subpopulation to subpopulation [26]. That is why computer-based prediction of candidate regulatory SNP markers of human diseases is a challenging problem for current functional genomics, genetics, and bioinformatics.

In our previous study [35], we described a freely available Web service, SNP_TATA_Comparator (created by us), and demonstrated its practical use on more than 40 biomedical SNP markers in the binding sites for TATA-binding protein (TBP) between positions −70 and −20 relative to the transcription start (the region where all such empirically proven sites are located [36, 37]). Recently, we showed suitability of this Web service for prediction of candidate SNP markers of complications of Mendelian diseases in obesity [38] and of autoimmune complications of these diseases [39] as well as SNP markers that can either enhance or weaken biological activity of oncogene inhibitors during cancer chemotherapy [40] (hereinafter, we use the term "Mendelian disease" according to the notation in database Online Mendelian Inheritance in Man, OMIM® [28]).

In the present work, we applied our Web service SNP_TATA_Comparator [35] to unannotated SNPs in binding sites of TBP which are located near known SNP markers of Mendelian human diseases and, for this reason, can also cause the same pathologies if these SNPs change the affinity of TBP for the same promoters of the same human genes. Furthermore, we found some data on biochemical markers of chronopathologies (where these markers have the effects on gene expression which are identical to the effects of the above SNPs) and clinical studies on the prevalence of these chronopathologies as complications of the Mendelian



(a) (b)

FIGURE 1: Examples of the predictions by SNP_TATA_Comparator [35] for statistically significant alterations in the affinity of TATA-binding protein for human gene promoters. (a) The known biomedical SNP marker rs35036378 located within a promoter of the human *ESR2* gene associated with a Mendelian disease. This SNP is now predicted (in this study) to be a candidate SNP marker of circadian comorbidities and complications of diseases that one can see in Table 1. (b) The candidate SNP marker rs4313857 identified in this study within the human *CSNK1D* gene (belongs to the circadian clock core), as shown in Table 2.

diseases caused by these SNPs. Finally, using SNP_TATA_Comparator [35], we analyzed all SNPs within 90 bp proximal promoter regions for all protein-coding transcripts of the genes of the circadian clock core. As a result, we identified 53 candidate SNP markers of human chronopathologies; validation of these markers in accordance with clinical standards may make these SNPs useful for predictive-preventive personalized medicine [41].

2. Methods

2.1. Web Service. Web service SNP_TATA_Comparator [35] is a bioinformatics application freely available on the Web (Figure 1; URL: http://beehive.bionet.nsc.ru/cgi-bin/mgs/tatascan/start.pl), which allows a user (i) to find an ancestral variant of the promoter for a transcript under study (the "Base sequence" text box) from the reference human genome (solid, dashed, dotted, and boldfaced arrows; BioPerl [115] is used), (ii) to introduce a mutation of interest (the "Editable sequence" text box), and (iii) to assess (the "Calculate" button) the values of TBP's affinity for these two promoter variants, the relative mutation-related change in transcript levels, and statistical significance according to Z-score (the "Result" text box) as described in detail in our previous study [35].

2.2. The Bioinformatics Model. For each proximal 90 bp DNA sequence $\{s_{-90} \cdots s_{-1}\}$ of a given gene promoter (where $s_i \in \{a, c, g, t\}$; s_0 is the transcription start site), our Web service SNP_TATA_Comparator [35] calculates the maximal value of $-\ln(K_D) \pm \delta$ of the estimate of TBP's binding affinity for the 26 bp window $\{s_{i-13} \cdots s_i \cdots s_{i+12}\}$ [116, 117] (where $-70 \le i \le -20$ in both DNA strands; K_D is the equilibrium dissociation constant of the TBP-DNA complex, expressed in moles per liter; M), as follows:

$$\ln(K_D) = 10.9 - 0.2\{\ln(K_1([\mathrm{TA}]; \mu)) + \ln(K_2(\mathrm{PWM}_{\mathrm{Bucher}})) + \ln(K_3([\mathrm{WR}]; [\mathrm{TV}]))\}; \quad (1)$$

$$\delta = \frac{1}{78} \cdot \sum_{\varphi \in \{a,t,g,c\}} \sum_{j=-13}^{12} \ln\left(\frac{K_D(s_{i-13} \cdots s_{i+j-1} s_{i+j} s_{i+j+1} \cdots s_{-1})}{K_D(s_{i-13} \cdots s_{i+j-1} \varphi s_{i+j+1} \cdots s_{-1})}\right), \quad (2)$$

where 10.9 (natural logarithm units) is empirical nonspecific TBP-DNA affinity, 10^{-5} M [118]; 0.2 is the stoichiometric coefficient; K_1 is an empirical estimate of the equilibrium constant of TBP sliding along DNA; the average values of TBP's affinity for double-stranded DNA were estimated using the minor groove width (μ) and the TA dinucleotide content, [TA] [119]. K_2 is an empirical estimate of the equilibrium constant of the primary corecognition between TBP and an appropriate TBP-biding site on DNA [$-\ln(K_2)$ is the maximal score of Bucher's position-weighted matrix: $\mathrm{PWM}_{\mathrm{Bucher}}$] [120]. K_3 is an empirical estimate of the equilibrium constant of stabilization of the TBP-DNA complex due to the bend of the axis of the DNA helix by an angle of 19° to 90° [121, 122] which depends on abundance of two TA-rich dinucleotides, WR $\in$ {AA, AG, TA, TG} and TV $\in$ {TA, TG, TC} [123]; δ is the standard deviation of K_D estimates for all the possible mononucleotide substitutions within the 26 bp DNA sliding window corresponding to the maximal K_D value found for the DNA sequence under study.

For two DNA sequences of the minor (mut) and ancestral (wt) alleles being compared, (1) and (2) yield $\{-\ln(K_D^{(\mathrm{mut})}) \pm$

Table 1: Candidate SNP markers of circadian complications and/or comorbidities of Mendelian diseases as predicted by a significant change in the affinity of TATA-binding protein for human gene promoters.

Gene	dbSNP [24] rel. 146					K_D, nM				Known Mendelian diseases (see [references]) *and (hypothetically) circadian complications and comorbidities whose candidate SNP markers were predicted by us in [this work] (see Figure 2)*	[References] or *[this work] (see Figure 2)*
	ID or [reference]	SNP **wt**→*mut*	5′ flank	**wt** / *mut*	3′ flank	**wt** / *mut*	Δ	Z	α		
HBB	rs397509430	**-29t***DEL*	gggctgggca	**t** / —	atacaacagt	**5** / *29*	↓	34	10^{-6}	Malaria resistance and β-thalassemia *and (hypothetically) circadian symptoms (worse at night) in restless legs syndrome caused by iron deficiency anemia cooccurring with thalassemia and for sensorineural hearing loss as a complication of deferoxamine-based therapy in thalassemia*	[24, 42], *[this work] [43–47]*
	rs33980857	**-29t**→*a, g, c*	gggctgggca	**t** / *a, g, c*	atacaacagt	**5** / *21*	↓	27	10^{-6}		
	rs34598529	**-28a**→*g*	ggctgggcat	**a** / *g*	aaagtcaggg	**5** / *18*	↓	24	10^{-6}		
	rs33931746	**-27a**→*g, c*	gctgggcata	**a** / *g, c*	aagtcagggc	**5** / *11*	↓	14	10^{-6}		
	rs33981098	**-30a**→*g, c*	agggctgggc	**a** / *g, c*	taaaagtcag	**5** / *9*	↓	10	10^{-6}		
	rs34500389	**-31c**→*a, t, g*	cagggctggg	**c** / *a, t, g*	ataaaagtca	**5** / *6*	↓	3	10^{-2}		
HBB	rs63750953	**-25aa***DEL*	ctgggcataa	**aa** / —	gtcagggcag	**5** / *8*	↓	*9*	*10^{-6}*	*(Hypothetically)malaria resistance, β-thalassemia, and circadian symptoms (worse at night) in restless legs syndrome and sensorineural hearing loss (for details, see above)*	*[This work] [42–47]*
	rs281864525	**-25a**→*c*	tgggcataaa	**a** / *c*	gtcagggcag	**5** / *7*	↓	*7*	*10^{-6}*		
HBD	rs35518301	**-31a**→*g*	caggaccagc	**a** / *g*	taaaaggcag	**4** / *8*	↓	11	10^{-6}	Malaria resistance and δ-thalassemia *and (hypothetically) circadian symptoms (worse at night) for restless legs syndrome and sensorineural hearing loss (for details, see above)*	[24, 42], *[this work] [43–47]*
HBD	rs34166473	**-30t**→*c*	aggaccagca	**t** / *c*	aaaaggcagg	**4** / *8*	↓	*18*	*10^{-6}*	*(Hypothetically) malaria resistance, δ-thalassemia, and circadian symptoms (worse at night) for restless legs syndrome and sensorineural hearing loss (for details, see above)*	*[This work] [42–47]*
MMP12	rs2276109	**-27a**→*g*	gatatcaact	**a** / *g*	tgagtcactc	**11** / *14*	↓	3	10^{-2}	Low risk of systemic sclerosis, psoriasis, and asthma whose symptoms are circadian (worse at night)	[48–51]

Table 1: Continued.

Gene	dbSNP [24] rel. 146 ID or [reference]	SNP **wt**→*mut*	5′ flank	**wt** / *mut*	3′ flank	K_D, nM **wt** / *mut*	Δ	*Z*	α	Known Mendelian diseases (see [references]) *and (hypothetically) circadian complications and comorbidities whose candidate SNP markers were predicted by us in [this work] (see Figure 2)*	[References] or *[this work] (see Figure 2)*
MMP12	rs572527200	**-30a**→*g*	gatgatatca	**a** / *g*	ctatgagtca	**11** / *14*	↓	*3*	10^{-2}	*(Hypothetically) low risk of systemic sclerosis, psoriasis, and asthma whose symptoms are circadian (worse at night)*	*[This work] [48–51]*
IL1B	rs1143627	**-31c**→*t*	ttttgaaagc	**c** / *t*	ataaaaacag	**5** / *2*	↑	15	10^{-6}	Gastric cancer in *Helicobacter pylori* infection, hepatocellular carcinoma in hepatitis C virus infection, non-small cell lung cancer, chronic gastritis and gastric ulcer in *H. pylori* infection, Graves' disease, greater body fat in older men, recurrent major depressive disorder whose diagnosis and therapy are characterized by circadian optima for use, *and (hypothetically) bipolar disorder whose diagnosis and therapy are characterized by circadian optimafor usedepending on the diet*	[24, 52–59], *[this work] [60, 61]*
IL1B	*rs549858786*	**-28a**→*t*	tgaaagccat	**a** / *t*	aaaacagcga	**5** / *6*	↓	*8*	10^{-6}	*(Hypothetically) rheumatoid arthritis that can disrupt the circadian rhythm of IL1B gene expression*	*[This work] [62, 63]*
F3	rs563763767	**-21c**→*t*	ccctttatag	**c** / *t*	gcgcggggca	**3** / *2*	↑	6	10^{-6}	Venous thromboembolism and myocardial infarction that are characterized by their circadian preference for the early morning in the elderly	[64–66]
F7	See [67]	**-33a**→*c*	ccttggaggc	**a** / *c*	gagaactttg	**53** / *62*	↓	3	10^{-2}	Moderate bleeding tendency whose symptoms are circadian: worse with chronic change of time zones and in winter	[67, 68]
F7	rs749691733	**-21c**→*t*	agaactttgc	**c** / *t*	cgtcagtccc	**53** / *66*	↓	4	10^{-3}	*(Hypothetically) moderate bleeding tendency whose symptoms are circadian: worse with chronic change of time zones and in winter*	*[This work] [67, 68]*
F7	rs367732974	**-19g**→*a*	aactttgccc	**g** / *a*	tcagtcccat	**53** / *47*	↑	*2*	*0.05*	*(Hypothetically) heart attacks that are characterized by circadian preference for the early morning in the elderly and circadian (postprandial) flare-ups of thrombogenesis*	*[This work] [69, 70]*
	rs549591993	**-13c**→*a*	gcccgtcagt	**c** / *a*	ccatggggaa	**53** / *25*	↑	*13*	10^{-6}		
	rs777947114	**-23g**→*a*	agagaacttt	**g** / *a*	cccgtcagtc	**53** / *19*	↑	*19*	10^{-6}		
	rs770113559	**-38g**→*a*	gtcacccttg	**g** / *a*	aggcagagaa	**53** / *41*	↑	*5*	10^{-6}		
	rs754814507	**-54c**→*t*	cctcccccat	**c** / *t*	cctctgtcac	**53** / *45*	↑	*3*	10^{-3}		

Table 1: Continued.

Gene	dbSNP [24] rel. 146 ID or [reference]	SNP **wt**→*mut*	5′ flank	**wt** / *mut*	3′ flank	K_D, nM **wt** / *mut*	Δ	*Z*	*α*	Known Mendelian diseases (see [references]) *and (hypothetically) circadian complications and comorbidities whose candidate SNP markers were predicted by us in [this work] (see Figure 2)*	[References] or *[this work] (see Figure 2)*
NOS2	See [71]	**-51t**→*c*	gtataaatac	**t** / *c*	tcttggctgc	**2** / *1*	↑	3	10^{-2}	Resistance to malaria and high risk of epilepsy that damages the hypothalamus and the circadian clock as a whole *and (hypothetically) remission of panic disorder whose symptoms are circadian (worse late in the evening)*	[71–73], *[this work] [74]*
DHFR	rs10168	**-26g**→*a*	ctgcacaaat	**g** / *a*	gggacgaggg	**15** / *9*	↑	9	10^{-6}	In leukemia, resistance to methotrexate therapy that is characterized by a circadian optimum for use	*[13, 75]*
DHFR	rs750793297	**-25g**→*t*	tgcacaaatg	**g** / *t*	ggacgagggg	**15** / *13*	↑	3	0.01	*(Hypothetically) in leukemia, resistance to methotrexate therapy that is characterized by a circadian optimum for use*	*[This work] [13, 75]*
DHFR	rs766799008	**-28a**→*g*	ctgcacaaat	**a** / *g*	tggggacgag	**15** / *19*	↓	3	10^{-3}	*(Hypothetically) enhancement of an apparent bioactivity of methotrexate therapy that may be characterized by a circadian optimum for use*	*[This work] [13, 40, 75]*
	rs764508464	**-a28*DEL***	ctgcacaaat	**a** / —	tggggacgag	**15** / *37*	↓	17	10^{-6}		
	rs754122321	**-31c**→*g*	ctcgcctgca	**c** / *g*	aaatggggac	**15** / *25*	↓	9	10^{-3}		
StAR	rs16887226	**-33c**→*t*	cagccttcag	**c** / *t*	gggggacatt	**10** / *10*	=	0	>0.5	Hypertension in diabetes (EMSA: an unknown TF-binding site is disrupted rather than a TBP-binding site) and *(hypothetically) weak circadian resistance to exotoxins (deficiency of the mediator between the circadian and immune systems)*	[76] *[this work] [11]*
StAR	rs544850971	**-22a**→*g*	tcagcggggg	**a** / *g*	catttaagac	**10** / *12*	↓	*5*	10^{-2}	*(Hypothetically) hypertension in diabetes and weak circadian resistance to exotoxins*	*[This work] [11, 76]*
CETP	See [77]	**-54**[**18 bp**]*DEL*	cgtgggggct	**[18]** / —	gggctccagg	**4** / *7*	↓	7	10^{-6}	Hyperalphalipoproteinemia reduces the risk of atherosclerosis that is characterized by circadian (postprandial) flare-ups	[77–79]
CETP	rs17231520	**-68g**→*a*	ggggctgggc	**g** / *a*	gacatacata	**4** / *2*	↑	*10*	10^{-6}	*(Hypothetically) risk of atherosclerosis that is characterized by circadian (postprandial) flare-ups and hypoalphalipoproteinemia causing chronopathologies of the liver*	*[This work] [80]*
	rs569033466	**-53g**→*a*	atacatatac	**g** / *a*	ggctccaggc	**4** / *3*	↑	*4*	10^{-3}		
	rs757176551	**-49c**→*g*	catatacggg	**c** / *g*	tccaggctga	**4** / *2*	↑	*10*	10^{63}		
APOA1	See [81]	**-35a**→*c*	tgcagacata	**a** / *c*	ataggccctg	**3** / *4*	↓	5	10^{-3}	Hypoalphalipoproteinemia causing chronopathologies in the liver, as well as hematuria, fatty liver, and obesity, *and (hypothetically) atherosclerosis that is characterized by circadian (postprandial) flare-ups*	[80, 81] *[this work] [77–80]*

Table 1: Continued.

Gene	dbSNP [24] rel. 146 ID or [reference]	SNP **wt**→*mut*	5′ flank	**wt** / *mut*	3′ flank	K_D, nM **wt** / *mut*	Δ	Z	α	Known Mendelian diseases (see [references]) *and (hypothetically) circadian complications and comorbidities whose candidate SNP markers were predicted by us in [this work] (see Figure 2)*	[References] or *[this work] (see Figure 2)*
CYP2B6	rs34223104	**-28t**→*c*	gatgaaattt	**t** / *c*	ataacagggt	**4** / *10*	↓	15	10^{-6}	Better metabolic activation of anticancer prodrug cyclophosphamide that is characterized by a circadian optimum for use	[14, 82]
CYP2B6	rs563558831	**-26t**→*c*	tgaaatttta	**t** / *c*	aacagggtgc	**4** / *10*	↓	*13*	10^{-6}	*(Hypothetically) better metabolic activation of cyclophosphamide that is characterized by a circadian optimum for use*	*[This work] [14, 82]*
INS	rs5505	**-9c**→*t*	agatcactgt	**c** / *t*	cttctgccat	**53** / *44*	↑	4	10^{-6}	Type 1 diabetes after neonatal diabetes mellitus and risk of hyperinsulinemia that disturbs circadian rhythms of the reproductive system, of blood pressure, and of tumor-host balance	[24, 83–85]
INS	rs563207167	**-28c**→*t*	tcagccctgc	**c** / *t*	tgtctcccag	**53** / *44*	↑	4	10^{-3}	*(Hypothetically) type 1 diabetes after neonatal diabetes mellitus and risk of hyperinsulinemia that disturbs circadian rhythms of the reproductive system, of blood pressure, and of tumor-host balance*	*[This work] [24, 83–85]*
INS	rs11557611	**-8c**→*t*	gatcactgtc	**c** / *t*	ttctgccatg	**53** / *60*	↓	2	0.05	*(Hypothetically) hypothalamic amenorrhea*	*[This work] [86]*
ESR2	rs35036378 (see Figure 1(a))	**-43t**→*g*	cctctcggtc	**t** / *g*	ttaaaaggaa	**6** / *8*	↓	5	10^{-3}	Preventive therapy for an ESR2-deficient pT1 primary tumor against its progression to breast cancer by means of tamoxifen when this treatment is characterized by a circadian optimum for use *and (hypothetically) disturbances in circadian (daytime) behavioral activity*	[87–89] *[this work] [90]*
ESR2	rs766797386	**-32g**→*t*	ttaaaaggaa	**g** / *t*	aaggggctta	**6** / *7*	↓	3	10^{-2}	*(Hypothetically) preventive therapy for an ESR2-deficient primary pT1 tumor against its progression to breast cancer by means of tamoxifen when this treatment is characterized by a circadian optimum for use; disturbances in circadian (daytime) behavioral activity*	*[This work] [87–90]*

wt, ancestral allele; mut, minor allele; K_D, an estimate [35] of the dissociation constant (K_D) of the TBP-DNA complex *in vitro* [91]; Δ, a gene expression change in comparison with the norm (=): overexpression (↑) and underexpression (↓); Z, Z-score; $\alpha = 1 - p$, significance (where p is a probability rate of acceptance of H_0 hypothesis "$K_D^{(\mathrm{mut})} \neq K_D^{(\mathrm{wt})}$"; see Figure 1); TF, transcription factor; EMSA, electrophoretic mobility shift assay.

Table 2: Candidate SNP markers within the human genes of the circadian clock core as predicted by a significant change in the affinity of TATA-binding protein for human gene promoters.

Gene	dbSNP [24] rel. 146					K_D, nM		Z	α	*Circadian complications and comorbidities whose candidate SNP markers were predicted in [this work] (clinical data, laboratory animals, or cellular model)*	*[References] found (see Figure 2)*
	ID	SNP **wt**→*mut*	5′ flank	**wt** / *mut*	3′ flank	**wt** / *mut*	Δ				
CLOCK	rs192518038	**-57g**→*t*	aggacctaag	**g** / *t*	ctagcgctct	**63** / *29*	↑	14	10^{-6}	*(Hypothetically) higher risk of heart attacks that are characterized by circadian preference for the early morning in the elderly with diabetes (CLOCK-mutant mice)*	*[This work] [92]*
	rs537333415	**-63c**→*t*	gcctccagga	**c** / *t*	ctaaggctag	**63** / *45*	↑	7	10^{-6}		
ARNTL	rs534789405	**-28g**→*a*	cggattggct	**g** / *a*	ggggcggccg	**184** / *89*	↑	12	10^{-6}	*(Hypothetically) malignant pleural mesothelioma (13 malignant pleural mesothelioma cell lines in comparison with the nontumorigenic mesothelial cell line MeT-5A)*	*[This work] [93]*
	rs758737644	**-39c**→*t*	tgcactgtta	**c** / *t*	acattctgtt	**10** / *3*	↑	4	10^{-3}		
	rs549031146	**-46t**→*a*	caaaacttat	**t** / *a*	gggtgctatg	**6** / *5*	↑	4	10^{-3}		
ARNTL2	rs776246315	**-21g**→*a, (t)*	ttccagccgc	**g** / *a, (t)*	tgagtccagg	**49** / *31*	↑	8	10^{-6}	*(Hypothetically) inhibition of type 1 diabetes (murine model)*	*[This work] [94]*
	rs140915764	**-22c**→*t*	tttccagccg	**c** / *t*	gtgagtccag	**49** / *32*	↑	7	10^{-6}		
	rs756988598	**-23g**→*a*	gtttccagcc	**g** / *a*	cgtgagtcca	**49** / *39*	↑	4	10^{-3}		
	rs753093730	**-33g**→*a, t*	ctgcccatag	**g** / *a, t*	taaagtgttg	**7** / *5*	↑	3	10^{-3}		
	rs536395877	**-46c**→*a*	ttgttgtact	**c** / *a*	tgctgcccat	**7** / *5*	↑	5	10^{-3}		
	rs769981079	**-42g**→*a, t*	ccagtgcatt	**g** / *a, t*	ctcctgtggt	**49** / *26*	↑	12	10^{-6}		
	rs111899732	**-46c**→*t*	agaaccagtg	**c** / *t*	attgctcctg	**49** / *14*	↑	18	10^{-6}		
	rs746050396	**-57g**→*a*	gttgagagag	**g** / *a*	agaaccagtg	**49** / *36*	↑	5	10^{-6}		
ARNTL2	rs369143719	**-48c**→*t*	tcttgttgta	**c** / *t*	tctgctgccc	**7** / *9*	↓	4	10^{-3}	*(Hypothetically) suppression of diabetes protection (murine model)*	*[This work] [95]*
	rs374142420	**-51g**→*c*	atgtcttgtt	**g** / *c*	tactctgctg	**7** / *10*	↓	4	10^{-3}		
	rs770635249	**-56[ttg]***DEL*	aataaatgtc	**ttg** / —	ttgtactctg	**7** / *10*	↓	4	10^{-3}		
CRY1	rs747100146	*INS-49tt*	gataggagtt	— / *tt*	aattatccta	**6** / *7*	↓	2	0.05	*(Hypothetically) susceptibility to arthritis ($CRY1^{-/-}CRY2^{-/-}$ knockout mice)*	*[This work] [96]*

Table 2: Continued.

Gene	dbSNP [24] rel. 146 ID	SNP **wt**→*mut*	5′ flank	**wt** / *mut*	3′ flank	K_D, nM **wt** / *mut*	Δ	Z	α	Circadian complications and comorbidities whose candidate SNP markers were predicted in [this work] (clinical data, laboratory animals, or cellular model)	[References] found (see Figure 2)
CRY2	rs753656899	**-16a**→*g*	gcggggacta	**a** / *g*	gggtggagtt	**27** / *56*	↓	12	10^{-6}	*(Hypothetically) susceptibility to arthritis and to mood disorders ($CRY1^{-/-}CRY2^{-/-}$ and $CRY2^{-/-}$ knockout mice, resp.)*	*[This work] [96, 97]*
CRY2	rs575588903	**-9a**→*g*	ctaagggtgg	**a** / *g*	gttgcggcgt	**27** / *25*	↑	2	0.05		
	rs757256843	**-14g**→*t*	ggggactaag	**g** / *t*	gtggagttgc	**27** / *13*	↑	10	10^{-6}	*(Hypothetically) resistance to chemotherapy and poor prognosis in colorectal cancer (colorectal cancer samples)*	*[This work] [98]*
	rs760179689	**-24g**→*a*	ccctgtgggc	**g** / *a*	gggactaagg	**27** / *22*	↑	1	10^{-3}		
	rs529410313	**-47c**→*a*	agctgtcagt	**c** / *a*	ttgcaagtca	**22** / *18*	↑	3	10^{-2}		
PER1	rs137890200	**-17c**→*t*	gccaataagg	**c** / *t*	ggagagtgtg	**21** / *14*	↑	6	10^{-6}		
	rs773740924	**-26c**→*a*	ctcgccctgg	**c** / *a*	caataaggcg	**21** / *14*	↑	7	10^{-6}	*(Hypothetically) longer survival among patients with gastric cancer*	*[This work] [99]*
	rs2518024	**-60g**→*a*	gtgctctgga	**g** / *a*	ttaaaccagc	**17** / *8*	↑	12	10^{-6}		
PER1	rs796629786	**-13t**→*g*	tcggcgcccc	**t** / *g*	aagccaataa	**56** / *118*	↓	13	10^{-6}	*(Hypothetically) prostate cancer; ethanol hepatotoxicity, defects in hippocampal development, and resulting impairment of spatial learning capacity ($PER1^{-/-}$ knockout mice)*	*[This work] [17, 100–103]*
	rs3027175	**-67c**→*t*	ccagcaggtg	**c** / *t*	tctggagtta	**17** / *19*	↓	2	0.05		
PER2	rs780846747	**-20c**→*t*	cacaccttgt	**c** / *t*	aagtagaaga	**10** / *5*	↑	12	10^{-6}	*(Hypothetically) susceptibility to acute Q fever in male patients; suppression of tumor growth (cell line S-180)*	*[This work] [104, 105]*
RORA	rs750596430	**-13c**→*t*	agccaggcag	**c** / *t*	agcggcgcgg	**106** / *60*	↑	10	10^{-6}		
	rs535050458	**-20c**→*t*	gccagcgagc	**c** / *t*	aggcagcagc	**106** / *72*	↑	7	10^{-6}		
	rs547170533	**-21c**→*a, (t)*	cgccagcgag	**c** / *a, (t)*	caggcagcag	**106** / *72*	↑	7	10^{-6}	*(Hypothetically) emphysema and its progression to lung cancer among smokers (mice, smoking machine TE-10)*	*[This work] [106]*
	rs551503425	**-50c**→*t*	tgccaatatc	**c** / *t*	aagggttgcc	**19** / *6*	↑	15	10^{-6}		
	rs374778785	**-56a**→*t*	attatccccc	**a** / *t*	tactcctccc	**34** / *28*	↑	4	10^{-3}		

Table 2: Continued.

Gene	dbSNP [24] rel. 146 ID	SNP **wt**→*mut*	5′ flank	**wt** / *mut*	3′ flank	K_D, nM **wt** / *mut*	Δ	Z	α	*Circadian complications and comorbidities whose candidate SNP markers were predicted in [this work] (clinical data, laboratory animals, or cellular model)*	*[References] found (see Figure 2)*
RORA	rs764749271	**-28[12 bp]***DEL*	ccttctcctt	**[12 bp]** / —	ttttttttt	**28** / *31*	↓	2	0.05	*(Hypothetically) severe cerebellar ataxia (RORA$^{-/-}$ knockout mice)*	*[This work] [107]*
	rs762440045	**-55a**→*c*	acatggagtc	**a** / *c*	gctccggcag	**106** / *245*	↓	15	10^{-6}		
RORC	rs568650510	**-15c**→*t*	actccttttc	**c** / *t*	ctgcctgctg	**55** / *25*	↑	14	10^{-6}	*(Hypothetically) neurological manifestations of Behçet syndrome and asthma whose symptoms are circadian (worse at night) (18 patients and 30 pediatric patients, resp.)*	*[This work] [51, 108, 109]*
CSNK1E	rs777965060	**-12c**→*t*	aggccctctg	**c** / *t*	ttgaccccca	**80** / *69*	↑	3	0.05	*(Hypothetically) higher risk of c-MYC-dependent carcinogenesis, testicular cancer, and overproduction of cerebral β-amyloid*	*[This work] [110–112]*
	rs369188273	**-16c**→*t*	ccctcccctc	**c** / *t*	gcgcccgctc	**288** / *195*	↑	7	10^{-6}		
	rs558609213	**-18c**→*t*	tcttttcttg	**c** / *t*	atccctgcag	**30** / *9*	↑	21	10^{-6}		
	rs201914168	**-20g**→*t*	cctgacagag	**g** / *t*	ccctctgctt	**80** / *53*	↑	7	10^{-6}		
	rs374953337	**-25c**→*t*	tgacccctga	**c** / *t*	agaggccctc	**80** / *27*	↑	17	10^{-6}		
	rs368019196	**-31c**→*t*	tgcctctgac	**c** / *t*	cctgacagag	**80** / *68*	↑	3	10^{-2}		
	rs775747363	**-40[gcc]***DEL*	ctggctgcct	**gcc** / —	tctgacccct	**80** / *67*	↑	4	10^{-3}		
	rs780432736	**-51g**→*a*	gccgggcgtg	**g** / *a*	ctggctgcct	**80** / *71*	↑	2	0.05		
	rs747636670	**-52g**→*a*	ggccgggcgt	**g** / *a*	gctggctgcc	**80** / *56*	↑	5	10^{-6}		

Table 2: Continued.

Gene	dbSNP [24] rel. 146					K_D, nM		Z	α	*Circadian complications and comorbidities whose candidate SNP markers were predicted in [this work] (clinical data, laboratory animals, or cellular model)*	*[References] found (see Figure 2)*
	ID	SNP **wt**→*mut*	5′ flank	**wt** / *mut*	3′ flank	**wt** / *mut*	Δ				
CSNK1E	rs775283367	**-11c**→*g*	ctcttaccta	**c** / *g*	gtcagctctt	**8** / *10*	↓	4	10^{-3}	*(Hypothetically) increased responsiveness to opioids (CSNK1E$^{-/-}$ knockout mice)*	*[This work] [113]*
	rs746761879	**-18t**→*c*	ccgactactc	**t** / *c*	tacctacgtc	**8** / *11*	↓	5	10^{-6}		
	rs777083641	**-22[ag]***DEL*	ctggctgcct	**ag** / —	tctgacccct	**80** / *93*	↓	3	10^{-2}		
	rs2899302	**-59g**→*c*	cgagaaaact	**g** / *c*	cgcgaggcct	**288** / *335*	↓	3	10^{-2}		
CSNK1D	rs4313857 (see Figure 1(b))	**-24g**→*a*	gccccgccgg	**g** / *a*	ttgctagggg	**57** / *32*	↑	8	10^{-6}	*(Hypothetically) breast cancers (27 surgically resected tumor samples)*	*[This work] [114]*
	rs571866458	**-38c**→*a*	gagccggacc	**c** / *a*	gcagtagcgg	**54** / *42*	↑	4	10^{-3}		
	rs540139460	**-56g**→*a*	gcagggtcgg	**g** / *a*	aggaggcctg	**253** / *145*	↑	10	10^{-6}		

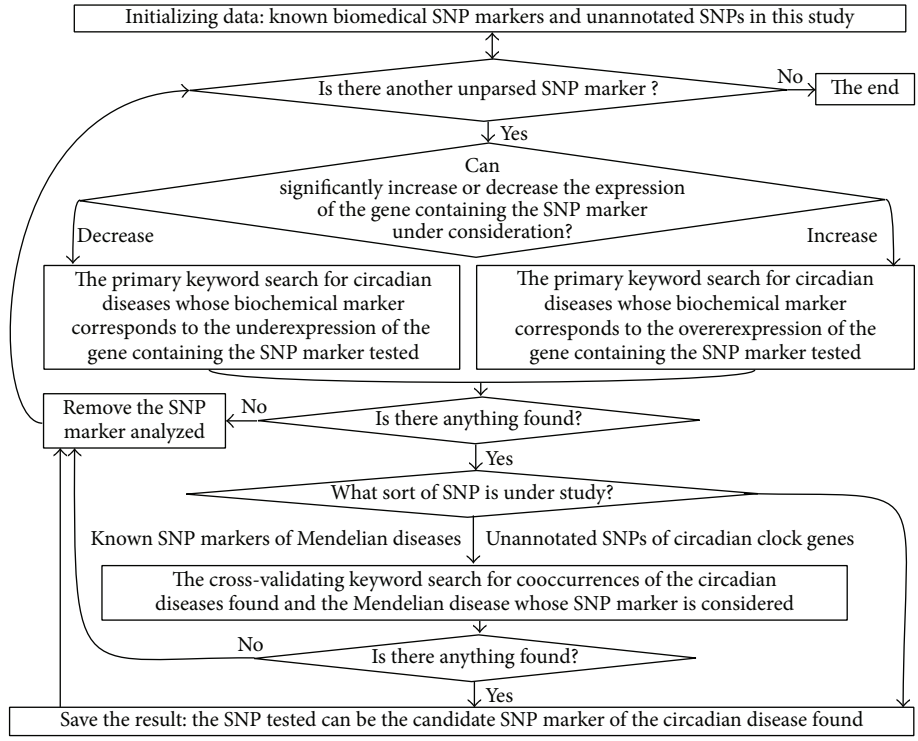


Figure 2: A flow chart showing our manual keyword search for chronopathologies (whose known biochemical markers correspond to our predictions of a significant change in the affinity of TATA-binding protein for human gene promoters).

$\delta_{(\mathrm{mut})}\}$ and $\{-\ln(K_D^{(\mathrm{wt})}) \pm \delta_{(\mathrm{wt})}\}$, respectively. Our Web service SNP_TATA_Comparator [35] compares them using Fisher's Z-score [124]:

$$Z = \frac{\left|\ln\left(K_D^{(\mathrm{mut})}/K_D^{(\mathrm{wt})}\right)\right|}{\sqrt{\delta_{(\mathrm{mut})}^2 + \delta_{(\mathrm{wt})}^2}}. \quad (3)$$

Using the standard statistical package R [124], we transform Z-score into p value of the probability of acceptance of H_0 hypothesis "$K_D^{(\mathrm{mut})} \neq K_D^{(\mathrm{wt})}$" (where $\alpha = 1 - p$ is the statistical significance). Two cases, "$K_D^{(\mathrm{mut})} < K_D^{(\mathrm{wt})}$" and "$K_D^{(\mathrm{mut})} > K_D^{(\mathrm{wt})}$," correspond, respectively, to overexpression and underexpression of the gene under study [125]. For more details, see our previous article [35].

2.3. Keyword Search. Figure 2 shows this keyword search for data on known biochemical markers of chronopathologies; these markers correspond to predictions of SNP_TATA_Comparator (Figure 1) regarding a relative mutation-induced change in gene expression. For each known or candidate SNP marker causing either significant overexpression or underexpression of the human gene containing the SNP, we performed a manual keyword search using various combinations of the terms "overexpression," "deficiency," "circadian," and many others corresponding to chronopathologies in public databases, as described in detail elsewhere [126]. In the case of genes of the circadian clock core, the obtained data are shown in Table 2 as results of this study. For SNP markers of Mendelian diseases, we conducted an additional keyword search for data on the prevalence of the uncovered chronopathologies as complications of these diseases; this procedure is some sort of cross-validation of the rough qualitative rates without statistical testing (Table 1).

Our heuristic interpretation of the keyword search results is shown in *italics* in the second rightmost column of Tables 1 and 2 and labeled with the word "*(Hypothetically)*" in front. We cite the studies (found during our manual keyword search) within the rightmost column of these tables, shown as [references] in *italics* and labeled with the phrase "*[This work]*."

3. Results and Discussion

3.1. The Results on Candidate SNP Markers of Circadian Complications of Mendelian Diseases. These results are presented in Table 1. Let us review in detail these more comprehensively studied SNP markers in order to briefly describe, in a similar fashion, the candidate SNP markers in the genes of human circadian clock core which were identified for the first time (in our study).

Genes HBB and HBD encode β- and δ-chains of hemoglobin, respectively. In the binding sites for TBP in their promoters, these two genes contain the greatest number (seven) of known SNP markers (rs35518301, rs397509430, rs33981098, rs34598529, rs33931746, rs33980857, and rs34500389) of thalassemia and resistance to malaria [24, 42], as a result of a hemoglobin deficiency (Table 1). A primary search by keywords uncovered a hemoglobin deficiency as a biochemical marker of circadian (nocturnal) aggravation of restless legs syndrome [43] and sensorineural hearing

loss [44]. A cross-validating search by keywords revealed that iron deficiency anemia substantially contributes to the pathogenesis of restless legs syndrome and cooccurs with thalassemia [45, 46], whereas sensorineural hearing loss is a complication of thalassemia in children during treatment with deferoxamine [47]. We found three additional unannotated SNPs (rs63750953, rs281864525, and rs34166473) that can also reduce expression of genes *HBB* and *HBD* and may serve as candidate SNP markers of these chronopathologies.

The MMP12 gene codes for matrix metalloproteinase 12 and, in its promoter, contains a known SNP marker (rs2276109) of a lower risk of systemic sclerosis [48], psoriasis [49], and asthma [50]. A keyword search yielded circadian (nocturnal) aggravation of asthma symptoms [51]. Here we found an unannotated SNP (rs572527200) with the same effects on the TBP-promoter affinity.

Gene IL1B encodes interleukin 1β and, in its promoter, contains one of the most widely studied SNP markers (rs1143627) of stomach ulcer, chronic gastritis, gastric cancer, hepatocellular carcinoma, non-small cell lung cancer, Graves' disease, and excess body fat in older men [52–57] as well as major depressive disorder [58] with a circadian optimum for diagnosis and treatment [59] that can be shifted by a high-fat or high-carbohydrate diet [60]. The primary search by keywords uncovered association of *IL1B* overexpression (with "-31T") with a bipolar disorder [61] that also has a circadian optimum for diagnosis and treatment depending on the diet [60]. Near this known SNP marker, we found unannotated rs549858786, which was found to lower *IL1B* expression (Table 1). The primary keyword search produced an IL1B protein deficiency as a biochemical marker of rheumatoid arthritis [62], for which an additional keyword search yielded a study showing that this disease is associated with disturbances of the circadian rhythm of *IL1B* expression [63].

The F3 gene encodes tissue thromboplastin (factor III) and, in its promoter, contains a known SNP marker (rs563763767) of an elevated risk of venous thromboembolism and myocardial infarction [64]. A keyword search produced clinical data on circadian aggravation of their symptoms (in the early morning) in the elderly [65]; these data are in agreement with basic research on a murine model of aging [66].

The F7 gene codes for serum prothrombin conversion accelerator (factor VII); in its promoter, some researchers [67] found a biomedical SNP marker: a substitution of the ancestral nucleotide A for minor nucleotide C at position -33 relative to the transcription start site (hereafter -33A→ C); this is a marker of moderate bleeding (as a result of underexpression of this gene). An additional database search revealed laboratory data on possible circadian aggravation of this disorder's symptoms during chronic changes of time zones and in the winter (data from a mouse model) [68]. Here we found an unannotated SNP (rs749691733) with the same effects on the TBP-promoter interaction. In addition, near this known SNP marker, we found five unannotated SNPs (rs367732974, rs549591993, rs777947114, rs770113559, and rs754814507) that can cause *F7* overexpression (Table 1). A keyword search produced an elevated F7 protein level as a biochemical marker of heart attacks characterized by a circadian preference for the early morning in the elderly [69] and for circadian (postprandial) development of thrombogenesis [70]. Therefore, we propose rs367732974, rs549591993, rs777947114, rs770113559, and rs754814507 as candidate SNP markers of these two chronopathologies.

Gene NOS2 encodes inducible NO synthase; in its promoter, one study [71] uncovered an SNP marker (-51T→C) of resistance to malaria [71] and of a high risk of epilepsy [72] (as a result of overexpression of this gene). A keyword search yielded a review article [73] about epilepsy-associated hypothalamic damage that can impair the circadian clock system of the body as a whole [73]. Besides, we found some data [74] suggesting that excess NO is a biochemical marker of a remission of panic disorder that is characterized by circadian (late evening) aggravation of symptoms. Thus, we propose the SNP "NOS2: -51T→C" as a candidate marker of these chronopathologies.

Gene DHFR codes for dihydrofolate reductase; its promoter contains a known SNP marker (rs10168) of methotrexate resistance [75] that is characterized by a therapeutic optimum of its use [13]. Here we found an unannotated SNP (rs750793297) with the same effects on the TBP-promoter complex. Additionally, near this known SNP marker, we found three unannotated SNPs (rs766799008, rs764508464, and rs754122321) that can cause DHFR underexpression (Table 1). According to our recent paper [40], these SNPs can elevate an apparent bioactivity of methotrexate-based antitumor chemotherapy [13, 75].

The StAR gene encodes steroidogenic acute regulatory protein and contains an SNP marker (rs16887226) of hypertension in diabetes (as a result of lowered expression of this gene because of impaired binding of its promoter with an unknown transcription factor, not TBP) [76]. A keyword search produced associations with lowered resistance to endotoxins for underexpression of the StAR protein, which is a mediator of mutual synchronicity of the immune system and circadian system [11]. Near this known SNP marker, we found the unannotated SNP rs544850971, which can lower *StAR* expression (Table 1) and therefore can be a candidate SNP marker of the above-mentioned disorders.

Gene CETP codes for cholesterol ester transfer protein; in its promoter, it contains a known biomedical SNP marker: deletion of the region G_{-72}GGCGGACATACATATAC$_{-54}$ (18 bp long) at position -54 relative to the transcription start site (hereafter: -54[18 bp]DEL); this is a marker of hyperalphalipoproteinemia that lowers the risk of atherosclerosis [77, 78]. A keyword search uncovered clinical data on circadian pathogenesis (postprandial flare-up) of this disorder in diabetes [79]. Near this known SNP marker, we found three unannotated SNPs (rs17231520, rs757176551, and rs569033466), which can increase *CETP* expression (Table 1) and thereby increase the risk of atherosclerosis [77–79] and of hypoalphalipoproteinemia which causes hepatic chronopathologies [80].

The APOA1 gene encodes apolipoprotein A1; in its promoter, some researchers [81] identified an SNP marker (-35A→C) of hematuria, hepatic steatosis, and obesity and of

hypoalphalipoproteinemia which impairs the peripheral circadian clock in the liver [80]. A keyword search yielded some data on a knockout mouse model ($APOA1^{-/-}$) regarding the risk of atherosclerosis [78] which develops in postprandial flare-ups in diabetes [79]. For this reason, we propose the SNP "APOA1: -31A→C" as a candidate marker of this chronopathology.

Gene CYP2B6 encodes cytochrome P450 2B6 and contains a known SNP marker (rs34223104) of improved bioactivation of cyclophosphamide [82] with a circadian therapeutic optimum [14]. According to empirical and computational data [82], this SNP disrupts a major variant of the TBP-binding site in the CYP2B6 promoter and in its place creates a binding site for the transcription factor (activator) C/EBP; this change shifts the TBP-binding site and transcription start by 30 bp in the $5' \rightarrow 3'$ direction and turns them into their minor alternative variants. In close proximity to this known SNP marker, we found the unannotated SNP rs563558831, which, in the same manner, lowers TBP's affinity for this promoter (Table 1) and therefore can be a candidate SNP marker of the same chronopathology.

The INS gene encodes insulin, and its promoter contains a known SNP marker (rs5505) of neonatal diabetes and hyperinsulinemia [24]. A keyword search uncovered hyperinsulinemia as a biochemical marker of aberrations in the circadian rhythms of (i) the reproductive system [83], (ii) blood pressure [84], and (iii) the tumor-host balance [85]. Near this known SNP marker, we found unannotated rs563207167, which can also cause hyperinsulinemia and therefore can be a candidate SNP marker of the same chronopathologies (Table 1). In addition, here we found unannotated rs11557611, which can cause hypoinsulinemia (Table 1). A keyword search showed that hypoinsulinemia is a biochemical marker of hypothalamic amenorrhea [86]. Consequently, rs1155761 may serve as a candidate SNP marker of this chronopathology (Table 1).

Gene ESR2 codes for estrogen receptor 2 (β) and, in its promoter, contains a known SNP marker (rs35036378) for prophylactic treatment (with tamoxifen) of an ESR2-deficient primary tumor pT1 [87] to prevent progression to breast cancer [88]; this treatment is characterized by a circadian optimum for its use [89]. A keyword search yielded basic research findings of circadian disturbances of daytime behavioral activity in ESR2-deficient female mice [90]. Near this known SNP marker, we found an unannotated SNP (rs35036378) with the same effects on the TBP-promoter affinity.

3.2. The Results on Candidate SNP Markers within the Circadian Clock Core. These results are shown in Table 2. Let us review in more detail the data in this table using the *PER1* gene as an example, which encodes a protein called period 1—a subunit of the heterodimeric PER-CRY complex—which is the main negative component of the circadian clock core: this complex inhibits the activity of transcription factor CLOCK/ARNTL [92–98].

As predicted by SNP_TATA_Comparator [35], only five of the 28 SNPs (that are known in the 90 bp proximal promoter regions for various protein-coding transcripts of this gene [24]) can affect the affinity of TBP for its promoters: rs137890200, rs773740924, and rs2518024 can enhance the TBP-promoter affinity, whereas rs796629786 and rs3027175 can reduce it. A keyword search showed that strong expression of the *PER1* gene inhibits the proliferation of tumor cells [17, 99, 100]; for example, in patients with strong expression of this gene, if they have a gastric cancer, longer survival is observed [99]. This gene is studied as a tumor suppressor; one of its mechanisms of action is the influence on the sensitivity of cells to DNA damage-induced apoptosis [17, 101]. Downregulation of PER1 was detected in human tissues of malignant tumors of the stomach and prostate [100, 101]. It should also be noted that, in studies of knockout mouse models ($PER1^{-/-}$), researchers observed impairment of spatial (3D) learning capacity and enhanced manifestations of ethanol hepatotoxicity [102, 103]. Therefore, we can hypothesize that rs137890200, rs773740924, rs2518024, rs796629786, and rs3027175 of the *PER1* gene are candidate SNP markers, as we propose in Table 2. One can see similar results for the other genes of the human circadian clock core [92–114] in this table.

Using SNP_TATA_Comparator [35], we analyzed 231 SNPs within 90 bp proximal promoter regions for the protein-coding transcripts of 12 genes of the human circadian clock core; only 52 of these SNPs (22%) were found to be capable of statistically significant changes in the affinity of TBP for promoters of these genes. As one can see in Table 2, we failed to find candidate SNP markers of chronopathologies for only one of the 12 genes, namely, *NR1D1*. This result shows that preliminary computational (bioinformatic) analysis of unannotated SNPs from the 1000 Genomes Project can indeed accelerate and cheapen the search for biomedical SNP markers because of selection (for this expensive and labor-intensive procedure) of only those candidate markers whose molecular mechanisms of pathological manifestation are easily understandable within the framework of existing clinical observations, genetic knowledge, scientific theories, hypotheses, and empirical data from animal and cellular models of human diseases.

It is also worth noting that only 13 of the 52 candidate SNP markers identified here decrease affinity of TBP for promoters of the genes of the circadian clock core, whereas the other 39 SNPs enhance it. In Table 1, however, one can see the opposite distribution of the candidate SNP markers (identified here) of circadian complications of Mendelian diseases: the majority (26 of 41, 62%) of the candidate SNP markers significantly reduce affinity of TBP for the human gene promoters, whereas the remaining 15 SNPs enhance it, as predicted by SNP_TATA_Comparator [35]. This difference is statistically significant ($p < 0.0005$) according to Fisher's exact test for 2×2 design. It is noteworthy (Table 1) that the ratio of the prevalence of candidate SNP markers of increased versus decreased affinity TBP-promoter is in agreement with independent studies by other investigators [127, 128]. Indeed, overall, in the reference human genome, the proportion of SNPs which weaken the binding sites of transcription factors is significantly greater than the share of SNPs which enhance this binding [127]. Similarly, some researchers [128] reported that SNPs of the binding sites for transcription factor NF-κB or

RNA polymerase II (significantly more often) weaken rather than enhance the binding of these proteins to the mutated DNA in comparison with the reference genome. Taken together, these findings suggest that the reduced proportion of candidate SNP markers weakening the affinity TBP-promoter may be a specific characteristic of the 12 genes of the human circadian clock core. This phenomenon may reflect the pressure of natural selection for robustness of their functioning under the conditions of incessant genetic variability of the promoter region being analyzed.

Why is the robustness of the circadian clock core so important for humans? As shown in Table 2, overall, dysregulation of these genes' expression may be a marker of a wide range of pathological conditions in humans, for example, cancer, neurodegenerative disorders, lung diseases, and cardiovascular diseases. The reason for such diversity of chronopathologies is that the circadian clock synchronizes a large number of molecular biological and biochemical processes on the whole-body level and integrates various individual signals from each cell, tissue, and organ into a united hierarchical system of circadian rhythms of the human body.

3.3. How to Use Candidate SNP Markers of Chronopathologies. In this work, we used SNP_TATA_Comparator [35] to analyze 484 SNPs within 90 bp proximal promoter regions for protein-coding transcripts of human genes. Only 53 of these SNPs (11%) were found to be candidate SNP markers of chronopathologies (Tables 1 and 2). This finding does not mean that the remaining 431 of the 484 SNPs (89%) cannot be SNP markers of some human diseases. This is because each of these SNPs may influence a specific promoter-related nucleosome [129], DNA methylation sites in promoters, binding sites for histone modifications, and binding sites for transcription factors (e.g., rs16887226 and rs34223104). At present, there is a large number and variety of freely available Web services [130–149]. Most of them rank unannotated SNPs by their generalized statistical similarity with biomedical SNP markers of human diseases; these Web services evaluate this similarity by superimposing SNPs on gene maps and on data from massively parallel high-throughput sequencing of chromatin immunoprecipitation material (ChIP-Seq) from experiments with complexes of various proteins with genomic DNA. Accuracy of such assessments is constantly increasing due to improvements in empirical formulas for whole-genome evaluation of similarity among pathological manifestations of various SNPs and due to the increasing diversity, completeness, and number of whole-genome maps for various epigenetic states of cells from various tissues and organs in health [150], during infection [151] (or other diseases [152]), or after treatment [153], as we predicted [154] on the basis of the Central Limit Theorem.

As an unexpected clever generalization of this mainstream approach, the authors of Web server GenomeRunner [155] proposed to evaluate the difference between SNPs in addition to the widely accepted notion of assessments of the similarity between them. In this active field of research, the new trend is creation of Web navigation services that help users generate their own hypotheses and ideas regarding how the SNP of interest can affect the signs and symptoms of diseases under study [156]. Another innovation that emerged here is Web service PredictSNP2 for translation from the numerical predictions to an effect of an SNP on human health which is suitable for precise computer calculations in qualitative categories that are accessible to the general population [157]. These breakthroughs mean that SNP-related predictions are becoming interesting not only to narrow specialists who treat patients with one or another disease but also to anyone who is willing to customize their lifestyle to minimize the risk of diseases.

Because statistical significance of our predicted candidate SNP markers (Tables 1 and 2) varies from high ($\alpha < 10^{-6}$) to minimally acceptable ($\alpha < 0.05$), the proposed markers should be properly validated using clinical standards before practical use. The results of this validation are dependent on climate, environmental conditions, and lifestyles and on the ethnic, social, age, and gender composition of cohorts [158]. Accordingly, we arranged the ancestral and minor alleles of each of candidate SNP markers of chronopathologies by the predicted K_D values of TBP-DNA affinity *in vitro* [91]. As shown in Tables 1 and 2, these K_D values vary from 1 to 335 nM, whereas the extent of their variation among alleles of a given SNP may be 1 nM, which is less than 0.3% of the K_D range. This level of allelic variations is too small for empirical measurement without an *a priori* known, fairly narrow range of K_D values to be measured. Thus, the predicted K_D values (Tables 1 and 2) are an integral part of each candidate SNP marker; without these data, an SNP marker cannot be validated in practice.

Finally, pathological manifestation of SNP markers of Mendelian diseases, as a rule, is limited to the consequences of changes in the expression of only those genes that contain these SNPs and can be useful only to physicians of the narrow specialties relevant to the diseases in question. Nonetheless, candidate SNP markers of chronopathologies are associated with consequences of desynchronoses either among the nervous, immune, digestive, respiratory, and other systems of the human body or between the human body and its environment (Tables 1 and 2). These data can be useful both for physicians and for the general population. For instance, the candidate SNP marker rs568650510 may be associated with an elevated risk of asthma whose symptoms are circadian (worse at night [51]; Table 2). Using this information, a physician can select the treatment timing (for asthma symptoms in a patient with minor alleles of these SNPs) that could reduce the risk of aggravation at night. By the same token, any person with the minor allele -15T of this SNP can choose a lifestyle that can reduce the systematic nocturnal influence of the environmental factor that causes the asthma symptoms. Similarly, rs367732974, rs549591993, rs192518038, and rs537333415 may help reduce the risk of a heart attack [69]; rs374778785 may be useful for lowering the risk of emphysema and lung cancer among smokers [106], whereas rs2899302 may help decide whether to use opioids [113].

4. Conclusions

Here, we predicted candidate SNP markers of chronopathologies (Tables 1 and 2); these SNPs can change affinity of

TATA-binding protein for human gene promoters. After proper validation of these candidate markers in accordance with clinical standards, these SNPs may turn out to be useful both for physicians (to select the best treatment for each patient) and for the general population (to choose a lifestyle preventing possible circadian comorbidities and complications).

Competing Interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as potential competing interests.

Acknowledgments

The authors are grateful to Shevchuk Editing (Brooklyn, NY, USA; URL: http://www.shevchuk-editing.com/) for English translation and editing. Writing of the paper was supported by project #14-04-00485 (for Ludmila Savinkova and Mikhail Ponomarenko) from the Russian Foundation for Basic Research. The software development and maintenance were supported by project #14-24-00123 (for Dmitry Rasskazov, Olga Podkolodnaya, Nikolay L. Podkolodny, Natalya N. Podkolodnaya, and Nikolay Kolchanov) from the Russian Scientific Foundation. The data compilation, processing, and analysis were supported by project #0324-2015-0003 (for Valentin Suslov and Irina Chadaeva) from the Russian Government Budget.

References

[1] R. Zhang, N. F. Lahens, H. I. Ballance, M. E. Hughes, and J. B. Hogenesch, "A circadian gene expression atlas in mammals: implications for biology and medicine," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 111, no. 45, pp. 16219–16224, 2014.

[2] S. M. Reppert and D. R. Weaver, "Molecular analysis of mammalian circadian rhythms," *Annual Review of Physiology*, vol. 63, pp. 647–676, 2001.

[3] S. A. Brown, E. Kowalska, and R. Dallmann, "(Re)inventing the circadian feedback loop," *Developmental Cell*, vol. 22, no. 3, pp. 477–487, 2012.

[4] J. K. Kim and D. B. Forger, "A mechanism for robust circadian timekeeping via stoichiometric balance," *Molecular Systems Biology*, vol. 8, no. 1, article 630, 2012.

[5] L. Chen and G. Yang, "PPARs integrate the mammalian clock and energy metabolism," *PPAR Research*, vol. 2014, Article ID 653017, 6 pages, 2014.

[6] K. Padmanabhan, M. S. Robles, T. Westerling, and C. J. Weitz, "Feedback regulation of transcriptional termination by the mammalian circadian clock PERIOD complex," *Science*, vol. 337, no. 6094, pp. 599–602, 2012.

[7] K. Eckel-Mahan and P. Sassone-Corsi, "Epigenetic regulation of the molecular clockwork," *Progress in Molecular Biology and Translational Science*, vol. 119, pp. 29–50, 2013.

[8] K. Bozek, A. Relógio, S. M. Kielbasa et al., "Regulation of clock-controlled genes in mammals," *PLoS ONE*, vol. 4, no. 3, Article ID e4882, 2009.

[9] K. Bozek, A. L. Rosahl, S. Gaub, S. Lorenzen, and H. Herzel, "Circadian transcription in liver," *BioSystems*, vol. 102, no. 1, pp. 61–69, 2010.

[10] J. S. Menet, S. Pescatore, and M. Rosbash, "CLOCK: BMAL1 is a pioneer-like transcription factor," *Genes and Development*, vol. 28, no. 1, pp. 8–13, 2014.

[11] J. Wang, Y. Luo, K. Wang et al., "Clock-controlled StAR's expression and corticosterone production contribute to the endotoxemia immune response," *Chronobiology International*, vol. 32, no. 3, pp. 358–367, 2015.

[12] P. Marckmann, B. Sandström, and J. Jespersen, "Dietary effects on circadian fluctuation in human blood coagulation factor VII and fibrinolysis," *Atherosclerosis*, vol. 101, no. 2, pp. 225–234, 1993.

[13] S. Ohdo, K. Inoue, E. Yukawa, S. Higuchi, S. Nakano, and N. Ogawa, "Chronotoxicity of methotrexate in mice and its relation to circadian rhythm of DNA synthesis and pharmacokinetics," *Japanese Journal of Pharmacology*, vol. 75, no. 3, pp. 283–290, 1997.

[14] V. Y. Gorbacheva, R. V. Kondratov, R. Zhang et al., "Circadian sensitivity to the chemotherapeutic agent cyclophosphamide depends on the functional status of the CLOCK/BMAL1 transactivation complex," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 9, pp. 3407–3412, 2005.

[15] C. H. Ko and J. S. Takahashi, "Molecular components of the mammalian circadian clock," *Human Molecular Genetics*, vol. 15, no. 2, pp. R271–R277, 2006.

[16] S. Sahar and P. Sassone-Corsi, "Regulation of metabolism: the circadian clock dictates the time," *Trends in Endocrinology and Metabolism*, vol. 23, no. 1, pp. 1–8, 2012.

[17] N. M. Kettner, C. A. Katchy, and L. Fu, "Circadian gene variants in cancer," *Annals of Medicine*, vol. 46, no. 4, pp. 208–220, 2014.

[18] O. A. Podkolodnaya, "Molecular and genetic aspects of interactions of the circadian clock and the energy-producing substrate metabolism in mammals," *Russian Journal of Genetics*, vol. 50, no. 2, pp. 111–122, 2014.

[19] G. V. Vasiliev, V. M. Merkulov, V. F. Kobzev, T. I. Merkulova, M. P. Ponomarenko, and N. A. Kolchanov, "Point mutations within 663-666 bp of intron 6 of the human TDO2 gene, associated with a number of psychiatric disorders, damage the YY-1 transcription factor binding site," *FEBS Letters*, vol. 462, no. 1-2, pp. 85–88, 1999.

[20] J. V. Ponomarenko, T. I. Merkulova, G. V. Vasiliev et al., "rSNP_Guide, a database system for analysis of transcription factor binding to target sequences: application to SNPs and site-directed mutations," *Nucleic Acids Research*, vol. 29, no. 1, pp. 312–316, 2001.

[21] J. V. Ponomarenko, G. V. Orlova, T. I. Merkulova et al., "rSNP_Guide: an integrated database-tools system for studying SNPs and site-directed mutations in transcription factor binding sites," *Human Mutation*, vol. 20, no. 4, pp. 239–248, 2002.

[22] J. V. Ponomarenko, T. I. Merkulova, G. V. Orlova et al., "rSNP Guide, a database system for analysis of transcription factor binding to DNA with variations: application to genome annotation," *Nucleic Acids Research*, vol. 31, no. 1, pp. 118–121, 2003.

[23] O. Delaneau, J. Marchini, and 1000 Genomes Project Consortium, "Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel," *Nature Communications*, vol. 5, article 3934, 2014.

[24] S. T. Sherry, M.-H. Ward, M. Kholodov et al., "DbSNP: the NCBI database of genetic variation," *Nucleic Acids Research*, vol. 29, no. 1, pp. 308–311, 2001.

[25] M. Haeussler, B. J. Raney, A. S. Hinrichs et al., "Navigating protected genomics data with UCSC Genome Browser in a box," *Bioinformatics*, vol. 31, no. 5, pp. 764–766, 2015.

[26] D. R. Zerbino, S. P. Wilder, N. Johnson, T. Juettemann, and P. R. Flicek, "The ensembl regulatory build," *Genome Biology*, vol. 16, no. 1, article 56, 2015.

[27] A. Abbas, M. Lechevrel, and F. Sichel, "Identification of new single nucleotid polymorphisms (SNP) in alcohol dehydrogenase class IV ADH7 gene within a French population," *Archives of Toxicology*, vol. 80, no. 4, pp. 201–205, 2006.

[28] J. S. Amberger, C. A. Bocchini, F. Schiettecatte, A. F. Scott, and A. Hamosh, "OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders," *Nucleic Acids Research*, vol. 43, no. 1, pp. D789–D798, 2015.

[29] H. Mitsuyasu, K. Izuhara, X. Q. Mao et al., "Ile50Val variant of IL4R alpha upregulates IgE synthesis and associates with atopic asthma," *Nature Genetics*, vol. 19, no. 2, pp. 119–120, 1998.

[30] V. Rajendran, "Structural analysis of oncogenic mutation of isocitrate dehydrogenase 1," *Molecular BioSystems*, vol. 12, no. 7, pp. 2276–2287, 2016.

[31] M. Lopus, D. M. Paul, and R. Rajasekaran, "Unraveling the deleterious effects of cancer-driven stk11 mutants through conformational sampling approach," *Cancer Informatics*, vol. 15, pp. 35–44, 2016.

[32] D. Meshach Paul and R. Rajasekaran, "Exploration of structural and functional variations owing to point mutations in α-NAGA," *Interdisciplinary Sciences: Computational Life Sciences*, 2016.

[33] V. Rajendran and R. Sethumadhavan, "Drug resistance mechanism of PncA in Mycobacterium tuberculosis," *Journal of Biomolecular Structure and Dynamics*, vol. 32, no. 2, pp. 209–221, 2014.

[34] B. Senthilkumar and R. Rajasekaran, "Analysis of the structural stability among cyclotide members through cystine knot fold that underpins its potential use as a drug scaffold," *International Journal of Peptide Research and Therapeutics*, pp. 1–11, 2016.

[35] M. Ponomarenko, D. Rasskazov, O. Arkova et al., "How to use SNP_TATA_comparator to find a significant change in gene expression caused by the regulatory SNP of this gene's promoter via a change in affinity of the TATA-binding protein for this promoter," *BioMed Research International*, vol. 2015, Article ID 359835, 17 pages, 2015.

[36] L. K. Savinkova, M. P. Ponomarenko, P. M. Ponomarenko et al., "TATA box polymorphisms in human gene promoters and associated hereditary pathologies," *Biochemistry*, vol. 74, no. 2, pp. 117–129, 2009.

[37] M. Ponomarenko, V. Mironova, K. Gunbin, and L. Savinkova, "Hogness box," in *Brenner's Encyclopedia of Genetics*, S. Maloy and K. Hughes, Eds., vol. 3, pp. 491–494, Academic Press, Elsevier Inc, San Diego, Calif, USA, 2nd edition, 2013.

[38] O. V. Arkova, M. P. Ponomarenko, D. A. Rasskazov et al., "Obesity-related known and candidate SNP markers can significantly change affinity of TATA-binding protein for human gene promoters," *BMC Genomics*, vol. 16, supplement 13, article S5, 2015.

[39] M. P. Ponomarenko, O. Arkova, D. Rasskazov, P. Ponomarenko, L. Savinkova, and N. Kolchanov, "Candidate SNP markers of gender-biased autoimmune complications of monogenic diseases are predicted by a significant change in the affinity of TATA-binding protein for human gene promoters," *Frontiers in Immunology*, vol. 7, article 130, 2016.

[40] I. I. Turnaev, D. A. Rasskazov, O. V. Arkova et al., "Hypothetical SNP markers that significantly affect the affinity of the TATA-binding protein to *VEGFA, ERBB2, IGF1R, FLT1, KDR*, and *MET* oncogene promoters as chemotherapy targets," *Molecular Biology*, vol. 50, no. 1, pp. 141–152, 2016.

[41] G. M. Trovato, "Sustainable medical research by effective and comprehensive medical skills: overcoming the frontiers by predictive, preventive and personalized medicine," *EPMA Journal*, vol. 5, no. 1, article 14, 2014.

[42] R. M. Bannerman, L. M. Garrick, P. Rusnak-Smalley, J. E. Hoke, and J. A. Edwards, "Hemoglobin deficit: an inherited hypochromic anemia in the mouse," *Proceedings of the Society for Experimental Biology and Medicine*, vol. 182, no. 1, pp. 52–57, 1986.

[43] E. L. Unger, C. J. Earley, and J. L. Beard, "Diurnal cycle influences peripheral and brain iron levels in mice," *Journal of Applied Physiology*, vol. 106, no. 1, pp. 187–193, 2009.

[44] A.-H. Sun, Z.-M. Wang, S.-Z. Xiao, Z.-J. Li, J.-Y. Li, and L.-S. Kong, "Red cell basic ferritin concentration in sensorineural hearing loss," *ORL Journal for Oto-Rhino-Laryngology and Its Related Specialties*, vol. 53, no. 5, pp. 270–272, 1991.

[45] R. P. Allen, S. Auerbach, H. Bahrain, M. Auerbach, and C. J. Earley, "The prevalence and impact of restless legs syndrome on patients with iron deficiency anemia," *American Journal of Hematology*, vol. 88, no. 4, pp. 261–264, 2013.

[46] S. Verma, R. Gupta, M. Kudesia, A. Mathur, G. Krishan, and S. Singh, "Coexisting iron deficiency anemia and Beta thalassemia trait: effect of iron therapy on red cell parameters and hemoglobin subtypes," *ISRN Hematology*, vol. 2014, Article ID 293216, 5 pages, 2014.

[47] D. Thio, V. Prasad, P. Anslow, and P. Lennox, "Marrow proliferation as a cause of hearing loss in beta-thalassaemia major," *The Journal of Laryngology and Otology*, vol. 122, no. 11, pp. 1253–1256, 2008.

[48] M. Manetti, L. Ibba-Manneschi, C. Fatini et al., "Association of a functional polymorphism in the matrix metalloproteinase-12 promoter region with systemic sclerosis in an Italian population," *Journal of Rheumatology*, vol. 37, no. 9, pp. 1852–1857, 2010.

[49] N. L. Starodubtseva, V. V. Sobolev, A. G. Soboleva, A. A. Nikolaev, and S. A. Bruskin, "Genes expression of metalloproteinases (MMP-1, MMP-2, MMP-9, and MMP-12) associated with psoriasis," *Russian Journal of Genetics*, vol. 47, no. 9, pp. 1117–1123, 2011.

[50] G. M. Hunninghake, M. H. Cho, Y. Tesfaigzi et al., "MMP12, lung function, and COPD in high-risk populations," *The New England Journal of Medicine*, vol. 361, no. 27, pp. 2599–2608, 2009.

[51] H. J. Durrington, S. N. Farrow, A. S. Loudon, and D. W. Ray, "The circadian clock and asthma," *Thorax*, vol. 69, no. 1, pp. 90–92, 2014.

[52] E. M. El-Omar, M. Carrington, W.-H. Chow et al., "Interleukin-1 polymorphisms associated with increased risk of gastric cancer," *Nature*, vol. 404, no. 6776, pp. 398–402, 2000.

[53] Y. Wang, N. Kato, Y. Hoshida et al., "Interleukin-1β gene polymorphisms associated with hepatocellular carcinoma in hepatitis C virus infection," *Hepatology*, vol. 37, no. 1, pp. 65–71, 2003.

[54] K.-S. Wu, X. Zhou, F. Zheng, X.-Q. Xu, Y.-H. Lin, and J. Yang, "Influence of interleukin-1 beta genetic polymorphism, smoking and alcohol drinking on the risk of non-small cell lung cancer," *Clinica Chimica Acta*, vol. 411, no. 19-20, pp. 1441–1446, 2010.

[55] D. N. Martínez-Carrillo, E. Garza-González, R. Betancourt-Linares et al., "Association of IL1B -511C/-31T haplotype and *Helicobacter pylori* vacA genotypes with gastric ulcer and chronic gastritis," *BMC Gastroenterology*, vol. 10, article 126, 2010.

[56] F. Hayashi, M. Watanabe, T. Nanba, N. Inoue, T. Akamizu, and Y. Iwatani, "Association of the -31C/T functional polymorphism in the interleukin-1β gene with the intractability of Graves' disease and the proportion of T helper type 17 cells," *Clinical and Experimental Immunology*, vol. 158, no. 3, pp. 281–286, 2009.

[57] L. Strandberg, D. Mellström, Ö. Ljunggren et al., "IL6 and IL1B polymorphisms are associated with fat mass in older men: The MrOS Study Sweden," *Obesity*, vol. 16, no. 3, pp. 710–713, 2008.

[58] P. Borkowska, K. Kucia, S. Rzezniczek et al., "Interleukin-1beta promoter (-31T/C and -511C/T) polymorphisms in major recurrent depression," *Journal of Molecular Neuroscience*, vol. 44, no. 1, pp. 12–16, 2011.

[59] C. Ávila Moraes, T. Cambras, A. Diez-Noguera et al., "A new chronobiological approach to discriminate between acute and chronic depression using peripheral temperature, rest-activity, and light exposure parameters," *BMC Psychiatry*, vol. 13, article 77, 2013.

[60] O. Pivovarova, K. Jürchott, N. Rudovich et al., "Changes of dietary fat and carbohydrate content alter central and peripheral clock in humans," *Journal of Clinical Endocrinology and Metabolism*, vol. 100, no. 6, pp. 2291–2302, 2015.

[61] C. J. Carter, "Multiple genes and factors associated with bipolar disorder converge on growth factor and stress activated kinase pathways controlling translation initiation: implications for oligodendrocyte viability," *Neurochemistry International*, vol. 50, no. 3, pp. 461–490, 2007.

[62] H. Yamazaki, M. Takeoka, M. Kitazawa et al., "ASC plays a role in the priming phase of the immune response to type II collagen in collagen-induced arthritis," *Rheumatology International*, vol. 32, no. 6, pp. 1625–1632, 2012.

[63] I. C. Chikanza, P. Petrou, G. Kingsley, G. Chrousos, and G. S. Panayi, "Defective hypothalamic response to immune and inflammatory stimuli in patients with rheumatoid arthritis," *Arthritis and Rheumatism*, vol. 35, no. 11, pp. 1281–1288, 1992.

[64] E. Arnaud, V. Barbalat, V. Nicaud et al., "Polymorphisms in the 5′ regulatory region of the tissue factor gene and the risk of myocardial infarction and venous thromboembolism: the ECTIM and PATHROS studies," *Arteriosclerosis, Thrombosis, and Vascular Biology*, vol. 20, no. 3, pp. 892–898, 2000.

[65] E. Haus, "Chronobiology of hemostasis and inferences for the chronotherapy of coagulation disorders and thrombosis prevention," *Advanced Drug Delivery Reviews*, vol. 59, no. 9-10, pp. 966–984, 2007.

[66] K. Oishi, S. Koyanagi, and N. Ohkura, "Circadian mRNA expression of coagulation and fibrinolytic factors is organ-dependently disrupted in aged mice," *Experimental Gerontology*, vol. 46, no. 12, pp. 994–999, 2011.

[67] A. Kavlie, L. Hiltunen, V. Rasi, and H. P. B. Prydz, "Two novel mutations in the human coagulation factor VII promoter," *Thrombosis and Haemostasis*, vol. 90, no. 2, pp. 194–205, 2003.

[68] I. Colognesi, V. Pasquali, A. Foà et al., "Temporal variations of coagulation factor VII activity in mice are influenced by lighting regime," *Chronobiology International*, vol. 24, no. 2, pp. 305–313, 2007.

[69] J. Carvalho De Sousa, E. Bruckert, P. Giral et al., "Coagulation factor VII and plasma triglycerides. Decreased catabolism as a possible mechanism of factor VII hyperactivity," *Haemostasis*, vol. 19, no. 3, pp. 125–130, 1989.

[70] P. Marckmann, B. Sandström, and J. Jespersen, "Dietary effects on circadian fluctuation in human blood coagulation factor VII and fibrinolysis," *Atherosclerosis*, vol. 101, no. 2, pp. 225–234, 1993.

[71] I. A. Clark, K. A. Rockett, and D. Burgner, "Genes, nitric oxide and malaria in African children," *Trends in Parasitology*, vol. 19, no. 8, pp. 335–337, 2003.

[72] J. A. González-Martínez, G. Möddel, Z. Ying, R. A. Prayson, W. E. Bingaman, and I. M. Najm, "Neuronal nitric oxide synthase expression in resected epileptic dysplastic neocortex," *Journal of Neurosurgery*, vol. 110, no. 2, pp. 343–349, 2009.

[73] W. A. Hofstra and A. W. de Weerd, "The circadian rhythm and its interaction with human epilepsy: a review of literature," *Sleep Medicine Reviews*, vol. 13, no. 6, pp. 413–420, 2009.

[74] B. Kaya, S. Ünal, A. B. Karabulut, and Y. Türköz, "Altered diurnal variation of nitric oxide production in patients with panic disorder," *Tohoku Journal of Experimental Medicine*, vol. 204, no. 2, pp. 147–154, 2004.

[75] F. Al-Shakfa, S. Dulucq, I. Brukner et al., "DNA variants in region for noncoding interfering transcript of Dihydrofolate reductase gene and outcome in childhood acute lymphoblastic leukemia," *Clinical Cancer Research*, vol. 15, no. 22, pp. 6931–6938, 2009.

[76] A. J. Casal, V. J. P. Sinclair, A. M. Capponi, J. Nicod, U. Huynh-Do, and P. Ferrari, "A novel mutation in the steroidogenic acute regulatory protein gene promoter leading to reduced promoter activity," *Journal of Molecular Endocrinology*, vol. 37, no. 1, pp. 71–80, 2006.

[77] W. Plengpanich, W. Le Goff, S. Poolsuk, Z. Julia, M. Guerin, and W. Khovidhunkit, "CETP deficiency due to a novel mutation in the CETP gene promoter and its effect on cholesterol efflux and selective uptake into hepatocytes," *Atherosclerosis*, vol. 216, no. 2, pp. 370–373, 2011.

[78] K. Oka, L. M. Belalcazar, C. Dieker et al., "Sustained phenotypic correction in a mouse model of hypoalphalipoproteinemia with a helper-dependent adenovirus vector," *Gene Therapy*, vol. 14, no. 3, pp. 191–202, 2007.

[79] S. Hirayama, S. Soda, Y. Ito et al., "Circadian change of serum concentration of small dense LDL-cholesterol in type 2 diabetic patients," *Clinica Chimica Acta*, vol. 411, no. 3-4, pp. 253–257, 2010.

[80] C. Gabás-Rivera, R. Martínez-Beamonte, J. L. Ríos et al., "Dietary oleanolic acid mediates circadian clock gene expression in liver independently of diet and animal model but requires apolipoprotein A1," *Journal of Nutritional Biochemistry*, vol. 24, no. 12, pp. 2100–2109, 2013.

[81] A. Matsunaga, J. Sasaki, H. Han et al., "Compound heterozygosity for an apolipoprotein A1 gene promoter mutation and a structural nonsense mutation with apolipoprotein A1 deficiency," *Arteriosclerosis, Thrombosis, and Vascular Biology*, vol. 19, no. 2, pp. 348–355, 1999.

[82] J. Zukunft, T. Lang, T. Richter et al., "A natural *CYP2B6* TATA box polymorphism (-82T→C) leading to enhanced transcription and relocation of the transcriptional start site," *Molecular Pharmacology*, vol. 67, no. 5, pp. 1772–1782, 2005.

[83] A. L. Mereness, Z. C. Murphy, and M. T. Sellix, "Developmental programming by androgen affects the circadian timing system in female mice1," *Biology of Reproduction*, vol. 92, no. 4, article 88, 2015.

[84] S. Bianchi, R. Bigazzi, R. Nenci, and V. M. Campese, "Hyperinsulinemia, circadian variation of blood pressure and end-organ damage in hypertension," *Journal of Nephrology*, vol. 10, no. 6, pp. 325–333, 1997.

[85] D. E. Blask, R. T. Dauchy, E. M. Dauchy et al., "Light exposure at night disrupts host/cancer circadian regulatory dynamics: impact on the Warburg effect, lipid signaling and tumor growth prevention," *PLoS ONE*, vol. 9, no. 8, Article ID e102776, 2014.

[86] G. A. Laughlin, C. E. Dominguez, and S. S. C. Yen, "Nutritional and endocrine-metabolic aberrations in women with functional hypothalamic amenorrhea," *Journal of Clinical Endocrinology and Metabolism*, vol. 83, no. 1, pp. 25–32, 1998.

[87] A. M. Sieuwerts, M. Ansems, M. P. Look et al., "Clinical significance of the nuclear receptor co-regulator DC-SCRIPT in breast cancer: an independent retrospective validation study," *Breast Cancer Research*, vol. 12, no. 6, article R103, 2010.

[88] S. Philips, A. Richter, S. Oesterreich et al., "Functional characterization of a genetic polymorphism in the promoter of the ESR2 gene," *Hormones and Cancer*, vol. 3, no. 1-2, pp. 37–43, 2012.

[89] L. Binkhorst, J. S. L. Kloth, A. S. de Wit et al., "Circadian variation in tamoxifen pharmacokinetics in mice and breast cancer patients," *Breast Cancer Research and Treatment*, vol. 152, no. 1, pp. 119–128, 2015.

[90] S. E. Royston, N. Yasui, A. G. Kondilis, S. V. Lord, J. A. Katzenellenbogen, and M. M. Mahoney, "ESR1 and ESR2 differentially regulate daily and circadian activity rhythms in female mice," *Endocrinology*, vol. 155, no. 7, pp. 2613–2623, 2014.

[91] L. K. Savinkova, I. A. Drachkova, T. V. Arshinova, P. Ponomarenko, M. Ponomarenko, and N. Kolchanov, "An experimental verification of the predicted effects of promoter TATA-box polymorphisms associated with human diseases on interactions between the TATA boxes and TATA-binding protein," *PLoS ONE*, vol. 8, no. 2, Article ID e54626, 2013.

[92] K. Oishi, N. Ohkura, N. Amagai, and N. Ishida, "Involvement of circadian clock gene Clock in diabetes-induced circadian augmentation of plasminogen activator inhibitor-1 (PAI-1) expression in the mouse heart," *FEBS Letters*, vol. 579, no. 17, pp. 3555–3559, 2005.

[93] M. Elshazley, M. Sato, T. Hase et al., "The circadian clock gene BMAL1 is a novel therapeutic target for malignant pleural mesothelioma," *International Journal of Cancer*, vol. 131, no. 12, pp. 2820–2831, 2012.

[94] C.-X. He, N. Prevot, C. Boitard, P. Avner, and U. C. Rogner, "Inhibition of type 1 diabetes by upregulation of the circadian rhythm-related aryl hydrocarbon receptor nuclear translocator-like 2," *Immunogenetics*, vol. 62, no. 9, pp. 585–592, 2010.

[95] C.-X. He, P. Avner, C. Boitard, and U. C. Rogner, "Downregulation of the circadian rhythm related gene Arntl2 suppresses diabetes protection in Idd6 NOD.C3H congenic mice," *Clinical and Experimental Pharmacology and Physiology*, vol. 37, no. 12, pp. 1154–1158, 2010.

[96] A. Hashiramoto, T. Yamane, K. Tsumiyama et al., "Mammalian clock gene *Cryptochrome* regulates arthritis via proinflammatory cytokine TNF-α," *The Journal of Immunology*, vol. 184, no. 3, pp. 1560–1565, 2010.

[97] G. Savalli, W. Diao, S. Berger, M. Ronovsky, T. Partonen, and D. D. Pollak, "Anhedonic behavior in cryptochrome 2-deficient mice is paralleled by altered diurnal patterns of amygdala gene expression," *Amino Acids*, vol. 47, no. 7, pp. 1367–1377, 2015.

[98] L. Fang, Z. Yang, J. Zhou et al., "Circadian clock gene CRY2 degradation is involved in chemoresistance of colorectal cancer," *Molecular Cancer Therapeutics*, vol. 14, no. 6, pp. 1476–1487, 2015.

[99] H. Zhao, Z.-L. Zeng, J. Yang et al., "Prognostic relevance of Period1 (Per1) and Period2 (Per2) expression in human gastric cancer," *International Journal of Clinical and Experimental Pathology*, vol. 7, no. 2, pp. 619–630, 2014.

[100] Q. Cao, S. Gery, A. Dashti et al., "A role for the clock gene *Per1* in prostate cancer," *Cancer Research*, vol. 69, no. 19, pp. 7619–7625, 2009.

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

[102] A. Jilg, S. Lesny, N. Peruzki et al., "Temporal dynamics of mouse hippocampal clock gene expression support memory processing," *Hippocampus*, vol. 20, no. 3, pp. 377–388, 2010.

[103] T. Wang, P. Yang, Y. Zhan, L. Xia, Z. Hua, and J. Zhang, "Deletion of circadian gene Per1 alleviates acute ethanol-induced hepatotoxicity in mice," *Toxicology*, vol. 314, no. 2-3, pp. 193–201, 2013.

[104] V. Mehraj, J. Textoris, C. Capo, D. Raoult, M. Leone, and J.-L. Mege, "Overexpression of the per2 gene in male patients with acute Q fever," *Journal of Infectious Diseases*, vol. 206, no. 11, pp. 1768–1770, 2012.

[105] K. Miyazaki, M. Wakabayashi, Y. Hara, and N. Ishida, "Tumor growth suppression in vivo by overexpression of the circadian component, PER2," *Genes to Cells*, vol. 15, no. 4, pp. 351–358, 2010.

[106] Y. Shi, J. Cao, J. Gao et al., "Retinoic acid-related orphan receptor-α is induced in the setting of DNA damage and promotes pulmonary emphysema," *American Journal of Respiratory and Critical Care Medicine*, vol. 186, no. 5, pp. 412–419, 2012.

[107] M. Doulazmi, F. Frédéric, F. Capone, M. Becker-André, N. Delhaye-Bouchaud, and J. Mariani, "A comparative study of Purkinje cells in two RORα gene mutant mice: staggerer and RORα−/−," *Developmental Brain Research*, vol. 127, no. 2, pp. 165–174, 2001.

[108] A. Hamzaoui, H. Maalmi, A. Berraies et al., "Transcriptional characteristics of CD4 T cells in young asthmatic children: RORC and FOXP3 axis," *Journal of Inflammation Research*, vol. 4, pp. 139–146, 2011.

[109] K. Hamzaoui, A. Borhani Haghighi, I. B. Ghorbel, and H. Houman, "RORC and Foxp3 axis in cerebrospinal fluid of patients with Neuro-Behçet's Disease," *Journal of Neuroimmunology*, vol. 233, no. 1-2, pp. 249–253, 2011.

[110] M. Toyoshima, H. L. Howie, M. Imakura et al., "Functional genomics identifies therapeutic targets for MYC-driven cancer," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 24, pp. 9545–9550, 2012.

[111] N. Rodriguez, J. Yang, K. Hasselblatt et al., "Casein kinase I epsilon interacts with mitochondrial proteins for the growth and survival of human ovarian cancer cells," *EMBO Molecular Medicine*, vol. 4, no. 9, pp. 952–963, 2012.

[112] M. Flajolet, G. He, M. Heiman, A. Lin, A. C. Nairn, and P. Greengard, "Regulation of Alzheimer's disease amyloid-β formation by casein kinase I," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 10, pp. 4159–4164, 2007.

[113] C. D. Bryant, C. C. Parker, L. Zhou et al., "*Csnk1e* is a genetic regulator of sensitivity to psychostimulants and opioids," *Neuropsychopharmacology*, vol. 37, no. 4, pp. 1026–1035, 2012.

[114] M. C. Abba, H. Sun, K. A. Hawkins et al., "Breast cancer molecular signatures as determined by SAGE: correlation with lymph node status," *Molecular Cancer Research*, vol. 5, no. 9, pp. 881–890, 2007.

[115] J. E. Stajich, D. Block, K. Boulez et al., "The Bioperl toolkit: perl modules for the life sciences," *Genome Research*, vol. 12, no. 10, pp. 1611–1618, 2002.

[116] P. M. Ponomarenko, L. K. Savinkova, I. A. Drachkova et al., "A step-by-step model of TBP/TATA box binding allows predicting human hereditary diseases by single nucleotide polymorphism," *Doklady Biochemistry and Biophysics*, vol. 419, no. 1, pp. 88–92, 2008.

[117] V. V. Mironova, N. A. Omelyanchuk, P. M. Ponomarenko, M. P. Ponomarenko, and N. A. Kolchanov, "Specific/nonspecific binding of TBP to promoter DNA of the auxin response factor genes in plants correlated with ARFs function on gene transcription (activator/repressor)," *Doklady Biochemistry and Biophysics*, vol. 433, no. 1, pp. 191–196, 2010.

[118] S. Hahn, S. Buratowski, P. A. Sharp, and L. Guarente, "Yeast TATA-binding protein TFIID binds to TATA elements with both consensus and nonconsensus DNA sequences," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 86, no. 15, pp. 5718–5722, 1989.

[119] M. P. Ponomarenko, J. V. Ponomarenko, A. S. Frolov et al., "Identification of sequence-dependent DNA features correlating to activity of DNA sites interacting with proteins," *Bioinformatics*, vol. 15, no. 7-8, pp. 687–703, 1999.

[120] P. Bucher, "Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences," *Journal of Molecular Biology*, vol. 212, no. 4, pp. 563–578, 1990.

[121] D. Flatters and R. Lavery, "Sequence-dependent dynamics of TATA-box binding sites," *Biophysical Journal*, vol. 75, no. 1, pp. 372–381, 1998.

[122] M. P. Ponomarenko, L. K. Savinkova, Y. V. Ponomarenko, A. E. Kel', I. I. Titov, and N. A. Kolchanov, "Simulation of TATA box sequences in eukaryotes," *Molecular Biology*, vol. 31, no. 4, pp. 616–622, 1997.

[123] IUPAC-IUB Commission on Biochemical Nomenclature (CBN), "Abbreviations and symbols for nucleic acids, polynucleotides and their constituents," *Journal of Molecular Biology*, vol. 55, no. 3, pp. 299–310, 1971.

[124] A. J. Waardenberg, S. D. Basset, R. Bouveret, and R. P. Harvey, "CompGO: an R package for comparing and visualizing Gene Ontology enrichment differences between DNA binding experiments," *BMC Bioinformatics*, vol. 16, no. 1, article 275, 2015.

[125] I. Mogno, F. Vallania, R. D. Mitra, and B. A. Cohen, "TATA is a modular component of synthetic promoters," *Genome Research*, vol. 20, no. 10, pp. 1391–1397, 2010.

[126] I. Missala, U. Kassner, and E. Steinhagen-Thiessen, "A systematic literature review of the association of lipoprotein(a) and autoimmune diseases and atherosclerosis," *International Journal of Rheumatology*, vol. 2012, Article ID 480784, 10 pages, 2012.

[127] G. R. Abecasis, A. Auton, L. D. Brooks et al., "An integrated map of genetic variation from 1,092 human genomes," *Nature*, vol. 491, no. 7422, pp. 56–65, 2012.

[128] M. Kasowski, F. Grubert, C. Heffelfinger et al., "Variation in transcription factor binding among humans," *Science*, vol. 328, no. 5975, pp. 232–235, 2010.

[129] I. Ioshikhes, E. N. Trifonov, and M. Q. Zhang, "Periodical distribution of transcription factor sites in promoter regions and connection with chromatin structure," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 6, pp. 2891–2895, 1999.

[130] C.-Y. Chen, I.-S. Chang, C. A. Hsiung, and W. W. Wasserman, "On the identification of potential regulatory variants within genome wide association candidate SNP sets," *BMC Medical Genomics*, vol. 7, article 34, 2014.

[131] M. Barenboim and T. Manke, "ChroMoS: an integrated web tool for SNP classification, prioritization and functional interpretation," *Bioinformatics*, vol. 29, no. 17, pp. 2197–2198, 2013.

[132] A. Riva, "Large-scale computational identification of regulatory SNPs with rSNP-MAPPER," *BMC genomics*, vol. 13, supplement 4, article S7, 2012.

[133] J. V. Ponomarenko, G. V. Orlova, A. S. Frolov, M. S. Gelfand, and M. P. Ponomarenko, "SELEX_DB: a database on *in vitro* selected oligomers adapted for recognizing natural sites and for analyzing both SNPs and site-directed mutagenesis data," *Nucleic Acids Research*, vol. 30, no. 1, pp. 195–199, 2002.

[134] I. V. Deyneko, Y. M. Kalybaeva, A. E. Kel, and H. Blöcker, "Human-chimpanzee promoter comparisons: property-conserved evolution?" *Genomics*, vol. 96, no. 3, pp. 129–133, 2010.

[135] M. C. Andersen, P. G. Engstrom, S. Lithwick et al., "In silico detection of sequence variations modifying transcriptional regulation," *PLoS Computational Biology*, vol. 4, no. 1, article e5, 2008.

[136] C.-C. Chen, S. Xiao, D. Xie et al., "Understanding variation in transcription factor binding by modeling transcription factor genome-epigenome interactions," *PLoS Computational Biology*, vol. 9, no. 12, Article ID e1003367, 2013.

[137] G. Macintyre, J. Bailey, I. Haviv, and A. Kowalczyk, "is-rSNP: a novel technique for in silico regulatory SNP detection," *Bioinformatics*, vol. 26, no. 18, pp. i524–i530, 2010.

[138] A. P. Boyle, E. L. Hong, M. Hariharan et al., "Annotation of functional variation in personal genomes using RegulomeDB," *Genome Research*, vol. 22, no. 9, pp. 1790–1797, 2012.

[139] J. V. Ponomarenko, D. P. Furman, A. S. Frolov et al., "ACTIVITY: a database on DNA/RNA sites activity adapted to apply sequence-activity relationships from one system to another," *Nucleic Acids Research*, vol. 29, no. 1, pp. 284–287, 2001.

[140] L. O. Bryzgalov, E. V. Antontseva, M. Y. Matveeva et al., "Detection of regulatory SNPs in human genome using ChIP-seq ENCODE data," *PLoS ONE*, vol. 8, no. 10, Article ID e78833, 2013.

[141] N. L. Podkolodnyy, D. A. Afonnikov, Y. Y. Vaskin et al., "Program complex SNP-MED for analysis of single-nucleotide polymorphism (SNP) effects on the function of genes associated with socially significant diseases," *Russian Journal of Genetics: Applied Research*, vol. 4, no. 3, pp. 159–167, 2014.

[142] A. D. Johnson, R. E. Handsaker, S. L. Pulit, M. M. Nizzari, C. J. O'Donnell, and P. I. W. De Bakker, "SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap," *Bioinformatics*, vol. 24, no. 24, pp. 2938–2939, 2008.

[143] I. V. Deyneko, B. Bredohl, D. Wesely et al., "FeatureScan: revealing property-dependent similarity of nucleotide sequences," *Nucleic Acids Research*, vol. 34, pp. W591–W595, 2006.

[144] S. F. Saccone, R. Bolze, P. Thomas et al., "SPOT: a web-based tool for using biological databases to prioritize SNPs after a genome-wide association study," *Nucleic Acids Research*, vol. 38, no. 2, pp. W201–W209, 2010.

[145] Y. Fu, Z. Liu, S. Lou et al., "FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer," *Genome Biology*, vol. 15, no. 10, article 480, 2014.

[146] S. G. Coetzee, S. K. Rhie, B. P. Berman, G. A. Coetzee, and H. Noushmehr, "FunciSNP: an R/bioconductor tool integrating functional non-coding data sets with genetic association studies to identify candidate regulatory SNPs," *Nucleic Acids Research*, vol. 40, no. 18, article e139, 2012.

[147] D. A. Rasskazov, E. V. Antontseva, L. O. Bryzgalov et al., "rSNP_Guide-based evaluation of SNPs in promoters of the human APC and MLH1 genes associated with colon cancer," *Russian Journal of Genetics: Applied Research*, vol. 4, no. 4, pp. 245–253, 2014.

[148] J. Ponomarenko, T. Merkulova, G. Orlova, O. Fokin, E. Gorshkov, and M. Ponomarenko, "Mining DNA sequences to predict sites which mutations cause genetic diseases," *Knowledge-Based Systems*, vol. 15, no. 4, pp. 225–233, 2002.

[149] J. Ponomarenko, G. Orlova, T. Merkulova, G. Vasiliev, and M. Ponomarenko, "Mining genome variation to associate genetic disease with mutation alterations and ortho/paralogous polimorphysms in transcription factor binding site," *International Journal on Artificial Intelligence Tools*, vol. 14, no. 4, pp. 599–619, 2005.

[150] Y. Ni, A. W. Hall, A. Battenhouse, and V. R. Iyer, "Simultaneous SNP identification and assessment of allele-specific bias from ChIP-seq data," *BMC Genetics*, vol. 13, article 46, 2012.

[151] S. Leschner, I. V. Deyneko, S. Lienenklaus et al., "Identification of tumor-specific *Salmonella* Typhimurium promoters and their regulatory logic," *Nucleic Acids Research*, vol. 40, no. 7, pp. 2984–2994, 2012.

[152] J. Hu, J. W. Locasale, J. H. Bielas et al., "Heterogeneity of tumor-induced gene expression changes in the human metabolic network," *Nature Biotechnology*, vol. 31, no. 6, pp. 522–529, 2013.

[153] M. Hein and S. Graver, "Tumor cell response to bevacizumab single agent therapy in vitro," *Cancer Cell International*, vol. 13, no. 1, article 94, 2013.

[154] M. P. Ponomarenko, J. V. Ponomarenko, A. S. Frolov et al., "Oligonucleotide frequency matrices addressed to recognizing functional DNA sites," *Bioinformatics*, vol. 15, no. 7-8, pp. 631–643, 1999.

[155] M. G. Dozmorov, L. R. Cara, C. B. Giles, and J. D. Wren, "GenomeRunner web server: regulatory similarity and differences define the functional impact of SNP sets," *Bioinformatics*, vol. 32, 2016.

[156] C. Liu, B. Ho, C. Chen et al., "ePIANNO: ePIgenomics ANNOtation tool," *PLOS ONE*, vol. 11, no. 2, article e0148321, 2016.

[157] J. Bendl, M. Musil, J. Štourač et al., "PredictSNP2: a unified platform for accurately evaluating SNP effects by exploiting the different characteristics of variants in distinct genomic regions," *PLoS Computational Biology*, vol. 12, no. 5, Article ID e1004962, 2016.

[158] S. S. Yoo, C. Jin, D. K. Jung et al., "Putative functional variants of XRCC1 identified by RegulomeDB were not associated with lung cancer risk in a Korean population," *Cancer Genetics*, vol. 208, no. 1-2, pp. 19–24, 2015.

Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 9480276, 10 pages
http://dx.doi.org/10.1155/2016/9480276

Research Article

QuaBingo: A Prediction System for Protein Quaternary Structure Attributes Using Block Composition

Chi-Hua Tung,[1] Chi-Wei Chen,[2] Ren-Chao Guo,[2] Hui-Fuang Ng,[3] and Yen-Wei Chu[2,4]

[1]*Department of Bioinformatics, Chung-Hua University, Room S116, No. 707, Section 2, WuFu Road, Hsinchu 30012, Taiwan*
[2]*Institute of Genomics and Bioinformatics, National Chung Hsing University, 250 Kuo Kuang Road, Taichung 402, Taiwan*
[3]*Department of Computer Science, Universiti Tunku Abdul Rahman, Jalan Universiti, 31900 Kampar, Malaysia*
[4]*Biotechnology Center, Agricultural Biotechnology Center, Institute of Molecular Biology, Graduate Institute of Biotechnology, National Chung Hsing University, 250 Kuo Kuang Road, Taichung 402, Taiwan*

Correspondence should be addressed to Yen-Wei Chu; ywchu@nchu.edu.tw

Received 23 February 2016; Revised 30 June 2016; Accepted 20 July 2016

Academic Editor: Ryuji Hamamoto

Background. Quaternary structures of proteins are closely relevant to gene regulation, signal transduction, and many other biological functions of proteins. In the current study, a new method based on protein-conserved motif composition in block format for feature extraction is proposed, which is termed block composition. *Results*. The protein quaternary assembly states prediction system which combines blocks with functional domain composition, called QuaBingo, is constructed by three layers of classifiers that can categorize quaternary structural attributes of monomer, homooligomer, and heterooligomer. The building of the first layer classifier uses support vector machines (SVM) based on blocks and functional domains of proteins, and the second layer SVM was utilized to process the outputs of the first layer. Finally, the result is determined by the Random Forest of the third layer. We compared the effectiveness of the combination of block composition, functional domain composition, and pseudoamino acid composition of the model. In the 11 kinds of functional protein families, QuaBingo is 23% of Matthews Correlation Coefficient (MCC) higher than the existing prediction system. The results also revealed the biological characterization of the top five block compositions. *Conclusions*. QuaBingo provides better predictive ability for predicting the quaternary structural attributes of proteins.

1. Background

Proteins are responsible for a vast amount of biological synthesis, enzyme catalysis, transport of molecules, and functions in cells. In addition, their specific functions are closely associated with molecular structure. Protein structure can be divided into four levels, that is, from primary to quaternary structure. Many important biological functions must be achieved by polymerization of protein monomers to form oligomeric proteins or higher order multimeric proteins. The concept of protein quaternary structure was first presented by Bernal in 1958 [1, 2], in which he found that some protein compositions and structures were more complicated than others. These proteins were shown to be composed of several protein subunits to form biological macromolecules. The quaternary structures of protein subunits fold together by noncovalent bonds, and thus the structure classification can be delineated according to the type of subunit. If the protein complex consists of identical subunits, it is called a homooligomer; otherwise it is referred to as a heterooligomer. Classification based on the number of subunits can be divided into dimers, trimers, tetramers, and so forth [3]. Examples include (1) insulin, having the activity to form a homodimer; (2) tumor necrosis factor-α (tumor necrosis factor-alpha), to form a tight trimer; and (3) human hemoglobin protein is a heterotetramer, with two identical α subunits and two identical β subunits. An excellent review summarized what is known about the biological functions of nonhomologous homodimer and heterodimeric complexes [4]. For example, thymidylate synthase, a homodimeric protein, is highly conserved among distant species. The tertiary complex of thymidylate synthase has been revealed about the asymmetrical conformation of two homodimers (PDB ID: 4EB4). The closed and open forms of a molecule of the complex dimer

may affect the ligand binding strength [5]. In addition, HIV-1 reverse transcriptase is a well-known drug target for treating HIV infections (PDB ID: 3HVT) [6]. Heterodimerization of HIV-1 reverse transcriptase contains subunit P66 and P51 is required for DNA polymerase activity.

Although there has been significant progress in the analysis of protein structure with various experimental approaches, experimentation performed to determine protein structure is typically expensive and time-consuming. Consequently, it is necessary to develop a protein quaternary assembly states prediction system that will enable the analysis of protein structure and function using the current and rapidly increasing amount of sequence data. In previous studies, Garian predicted homodimers and nonhomodimers using a decision-tree and amino acid composition method involving the integration of AAindex. Zhang utilized support vector machines (SVM) and a weighted autocorrelation function in an attempt to identify the key features from the amino acid composition. These studies demonstrated that the primary structure indeed possessed needed information about quaternary structure formation [7, 8]. However, the general feature encoding method of amino acid composition will lose much important protein sequence information, such as physical and chemical properties of amino acids. Therefore, pseudoamino acid composition (PseAAC) was used to predict quaternary structure. This feature not only incorporates the sequence order effect but also reflects hydrophobic and hydrophilic properties [9]. Zhang et al. used PseAAC to develop sequence-segmented PseAAC and combined segments of the protein sequence and domain relationships in an effort to improve prediction results [10]. In recent years, functional domain composition was presented from an evolutionary and functional perspective, because proteins that share similar domain structures often have similar functions [11–13]. This method is suitable for applications in multiple categories of quaternary structural classification problems and can greatly improve prediction performance. However, a disadvantage is that some proteins may not contain any other known functional domains. In fact, the corresponding known functional domains are too few to represent proteins, which result in a classifier being unable to learn effectively. These problems are due to the current database still being incomplete.

The objective of this study is to construct an accurate prediction system for protein quaternary structure attributes. In addition to the previous studies, which have been shown to achieve high prediction accuracy of functional domain composition, the method of functional domains possesses problems that need to be overcome. Accordingly, we attempt to improve this feature extraction method based on a protein sequence homology region concept, that is, block composition, which was proposed to present the protein characteristics. Since the protein interaction binding sites usually have more surface area and a high exposure of hydrophobic solvent accessibility, we will combine amino acid solvent accessibility information and pseudoamino acid composition to calculate the sequence order effect. This system is a three-layer prediction classifier framework. The first layer classifier identifies the structure type of the unknown protein sequence which is, respectively, monomer, dimer, trimer, tetramer, pentamer, hexamer, octamer, decamer, and dodecamer. Then, the result of the first layer of each class serves as input for the second layer classifier, which is used to integrate different features, considering different protein features in the predictive ability of the corresponding advantages and disadvantages to enhance the accuracy of prediction. Finally, the third layer classifier determines the structure type of the query protein.

Cross-validation results show that the predicted results using block composition obtain the best results. Specifically, the overall average prediction accuracy rate is more than 90% in the 60% sequence similarity of each class. Functional domain composition and PseASA are lower by about 10% and 20%, respectively. The results prove that block composition is able to effectively identify quaternary structure assembly states. In addition, performance analysis of different types of function proteins revealed that QuaBingo exhibits superior predictive ability for enzymes, gene regulation, signal transduction, molecular binding, and other important proteins. An online web server is freely available at http://predictor.nchu.edu.tw/QuaBingo/.

2. Methods

2.1. Compilation of Datasets. The protein oligomer sequences used in this study come from the 3D Complex [14] protein quaternary structure classification database. This database provides protein structures, structure type, symmetrical patterns, and other pieces of relevant information. We searched homo- and heterooligomers of each class from the 3D Complex, and information regarding the corrected number of subunits was utilized to construct the database. The following steps were performed for processing: (1) remove oligomer sequences with lengths of less than 30 amino acids; (2) remove those sequences containing greater than or equal to three unknown amino acid; and (3) use the CD-HIT [15, 16] to remove redundant sequences in the database, that is, the sequence identity with 60%, for avoiding prediction bias. However, the classes of pentamer, octamer, decamer, and dodecamer used CD-HIT 90% for processing to avoid losing sufficient statistical significance. Finally, the database had 8,444 sequences, named Oli8444. This database was employed as the training dataset of the first and the second layer classifiers. Specifically, there were 3,273 monomers, 3,658 homooligomers, and 1,513 heterooligomers. In addition to monomers, the homo- and heterooligomers have eight individual subcategories, that is, dimer, trimer, tetramer, pentamer, hexamer, octamer, decamer, and dodecamer. Heptamer and undecamer sequences are not used due to little available information. The serial numbers of each category are listed in Supplementary Table S1 in Supplementary Material available online at http://dx.doi.org/10.1155/2016/9480276. In order to obtain more representative types of sequences, the training data in third layer classifier are processed by CD-HIT 40% from the Oli8444 training set to further remove sequences containing one or more types of oligomer and named Oli6926. However, the sequence has too few categories, such

as pentamer, octamer, decamer, and dodecamer, and is no longer subject to CD-HIT 40% treatment. The independent test is collected from nonlearning test sequences of Oli6926.

2.2. Block Composition (Block). A motif is a small and highly conserved sequence in the secondary structure, which is usually associated with protein function; there are multiple motifs in proteins. The Blocks database [17–19] is a protein motif database which is based on SWISS-PROT and Prosite to calculate ungapped multiple alignment of protein sequences present in short segments of high sequence similarity blocks. Because this feature extraction method is based on searching the sequence of the Blocks database, the method is termed block composition. The Blocks database currently contains 29,068 protein blocks. P_{Block} can be defined as 29,068 dimensional space vectors by (1). If the protein P can be compared to the corresponding block i in the Blocks database, B_i is 1; otherwise it is 0. The rule is defined by the following equation (2). One has

$$P_{\text{Block}} = \left[B_1 \;\; B_2 \;\; \cdots \;\; B_i \;\; \cdots \;\; B_{29068}\right]^T \quad (T \text{ is the transpose operator}), \tag{1}$$

$$B_i = \begin{cases} 1, & \text{when a hit is found for } P \text{ in the Blocks database} \\ 0, & \text{otherwise} \end{cases} \quad (i = 1, 2, \ldots, 29068). \tag{2}$$

2.3. Functional Domain Composition (FunD). Proteins usually consist of one or more functional domains. When the same functional domains are discovered in different proteins, this indicates that they may have the same evolutionary origin and function. Version v3.10 of CDD [20] contains 44,354 protein domains and families and includes several external source databases (Pfam [21], SMART [22], KOG [23], COG [23], PRK [24], and TIGR [25]). We use a conservative threshold with E-value <0.01 in order to identify what kinds of functional domains are found for query protein P. 44,354 proteins can be expressed as a feature vector P_{FunD} dimensional space by (3). If D_i is 1, this means that the ith domain in CDD is found for P, otherwise it is 0. The rule is defined by (3). One has

$$P_{\text{FunD}} = \left[D_1 \;\; D_2 \;\; \cdots \;\; D_i \;\; \cdots \;\; D_{44354}\right]^T \quad (T \text{ is the transpose operator}), \tag{3}$$

$$D_i = \begin{cases} 1, & \text{when a doman is found for } P \text{ in CDD} \\ 0, & \text{otherwise} \end{cases} \quad (i = 1, 2, \ldots, 44354). \tag{4}$$

2.4. Pseudoamino Acid Composition Based on Solvent Accessibility of Amino Acid (PseASA). Protein quaternary structure formed by interactions between two or more polypeptide chains and the interaction depend on surfaces of amino acids in contact with each other. Recent studies of protein hotspots suggest that solvent accessibility constitutes an important feature of protein interactions [26, 27]. Protein binding sites usually have a more exposed hydrophobic area and higher solvent accessibility. Therefore, we will apply this feature in encoding pseudoamino acid composition [9], named PseASA, to investigate the effect of the relationship between protein interactions and structure on the prediction system. First, the information regarding amino acid solvent accessibility is derived from NetSurfP version 1.1 [28] prediction data and divided into "exposed" or "buried" states. The discontinuous exposure and buried amino acid are linked into exposed protein sequence $P_e\ E_1E_2E_3E_4E_5 \cdots E_m$ and buried protein sequence $P_b\ B_1B_2B_3B_4B_5 \cdots B_n$ (m and n are the sequence lengths and may change with prediction data of different proteins). PseAAC-Builder [29] was used for feature encoding of pseudoamino acids. However, because of the consideration about overall accuracy of using protein features on Oli8444 dataset, QuaBingo did not use the PseASA feature (see Section 3; Tables 1 and 2).

Table 1: Performance of using different features with SVM in 10-fold cross-validation for monomer classification on the Oli8444 dataset.

	Monomeric protein			
	Sn (%)	Sp (%)	ACC (%)	MCC
Block				
Monomer	79.07	78.75	78.91	0.579
FunD				
Monomer	93.68	57.80	75.75	0.552
PseASA				
Monomer	70.15	56.33	63.24	0.268

2.5. The Three-Layer Architecture of Classifiers. SVM is generally used as a binary classifier that was initially applied to pattern recognition and other fields [30]. In the past, SVM has been successfully applied in various fields of classification problems, and the predictions of quaternary structure have also been found to achieve good results [8, 10, 31]. LibSVM is utilized in this study, and was developed by Chang and Lin [32].

The construction of the prediction system in the current study employs a three-tier architecture, the first layer of which uses SVM to create different characteristic rules of binary classification prediction model. Feature selection using python syntax written LibSVM package fselect.py [33] gives F-score based on the importance of each feature and then sorts the trained model by F-score. In order to avoid poor recognition and enormous computational time, the trained models are divided into four equal parts according to the F-score from high to low and remove 25% or less or more than 75% of the models. Finally, the construction of first layer classification model is completed by choosing better sensitivity, specificity, and Matthews Correlation Coefficient (MCC) based on 10-fold cross-validation accuracy of measurement. Due to 10-fold cross-validation results of first layer classification model, the predictive power of three kinds

Table 2: Performance of using different features with SVM in 10-fold cross-validation for homo- and heterooligomer classification on the Oli8444 dataset.

	Homooligomer				Heterooligomer			
	Sn (%)	Sp (%)	ACC (%)	MCC	Sn (%)	Sp (%)	ACC (%)	MCC
Block								
Dimer	83.18	82.83	83.00	0.660	66.30	97.00	86.73	**0.697**
Trimer	89.25	99.76	95.48	0.909	83.65	97.50	91.93	**0.835**
Tetramer	75.32	97.53	90.12	0.776	85.03	98.03	92.82	**0.853**
Pentamer	100.00	96.67	98.57	0.973	83.33	95.00	89.17	**0.813**
Hexamer	89.03	98.52	94.27	0.887	82.50	97.00	90.42	**0.812**
Octamer	95.71	72.62	84.17	0.709	90.00	85.33	87.67	**0.782**
Decamer	91.67	100.00	95.83	0.928	100.00	100.00	100.00	**1.000**
Dodecamer	95.50	98.00	96.75	0.941	86.00	94.67	90.33	**0.823**
Overall			92.27	0.848			91.13	**0.827**
FunD								
Dimer	52.49	90.26	71.37	0.462	74.74	90.16	85.01	**0.659**
Trimer	93.73	87.83	90.22	0.806	69.94	88.25	80.87	**0.600**
Tetramer	60.64	96.61	84.62	0.647	71.96	95.08	85.79	**0.706**
Pentamer	75.00	86.67	80.48	0.649	53.33	100.00	76.67	**0.572**
Hexamer	64.94	100.00	84.27	0.712	85.75	83.89	84.86	**0.698**
Octamer	48.81	100.00	74.40	0.566	44.67	100.00	72.33	**0.517**
Decamer	63.33	100.00	81.67	0.691	45.00	100.00	72.50	**0.473**
Dodecamer	63.00	100.00	81.50	0.680	69.00	100.00	84.50	**0.733**
Overall			81.07	0.652			80.32	**0.620**
PseASA								
Dimer	66.95	46.74	56.85	0.140	12.62	93.07	66.18	**0.094**
Trimer	39.16	85.93	66.93	0.288	36.08	82.75	63.98	**0.218**
Tetramer	30.11	91.52	71.05	0.280	33.37	83.49	63.34	**0.194**
Pentamer	64.17	70.00	67.62	0.343	86.67	66.67	76.67	**0.564**
Hexamer	65.76	60.37	62.78	0.262	61.63	80.37	71.85	**0.431**
Octamer	66.90	60.48	63.69	0.285	86.00	71.50	78.75	**0.604**
Decamer	81.67	93.33	87.50	0.785	90.00	85.00	87.50	**0.773**
Dodecamer	73.50	68.00	70.75	0.429	66.83	85.50	76.17	**0.556**
Overall			68.40	0.352			73.05	**0.429**

of characteristic rules for different classes of oligomers was known.

The second layer is the first layer using SVM optimization model predictions, the purpose of which is combining the individual features of each oligomer model outputs into one. Training the second layer integrated model approach is using 10-fold cross-validation test predictions of first layer as input and considering the strengths and weaknesses of the characteristics of different proteins in order to improve accuracy of prediction.

By comparing the data analysis ability of different machine learning algorithms, we finally selected Random Forest to construct the third layer classifier for the integration of these recognition results and determine the quaternary structure type of protein oligomer. Figure 1 is a flowchart of the predicting system.

2.6. Evaluation Measures. To assess the predictive performance of the classifier, we use the following formula. TP, FP, FN, and TN are true positives, false positives, false negatives, and true negatives, respectively. Sensitivity (Sn) on behalf of this type of protein oligomer reflects the percentage of correct predictions for that category. Specificity (Sp) on behalf of nonprotein oligomers of this type indicates the percentage of correct predictions of nonclass. Accuracy (ACC) is used to assess the overall predictive power of the prediction accuracy. Matthews Correlation Coefficient (MCC) values range from −1 to 1, in which the value of 1 represents a completely correct prediction, the value of 0 represents random prediction, and the value of −1 represents exactly the opposite prediction:

$$\mathrm{Sn} = \left(\frac{\mathrm{TP}}{\mathrm{TP}} + \mathrm{FN}\right) \times 100,$$

$$\mathrm{Sp} = \left(\frac{\mathrm{TN}}{\mathrm{TN}} + \mathrm{FP}\right) \times 100\%,$$

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \times 100\%,$$

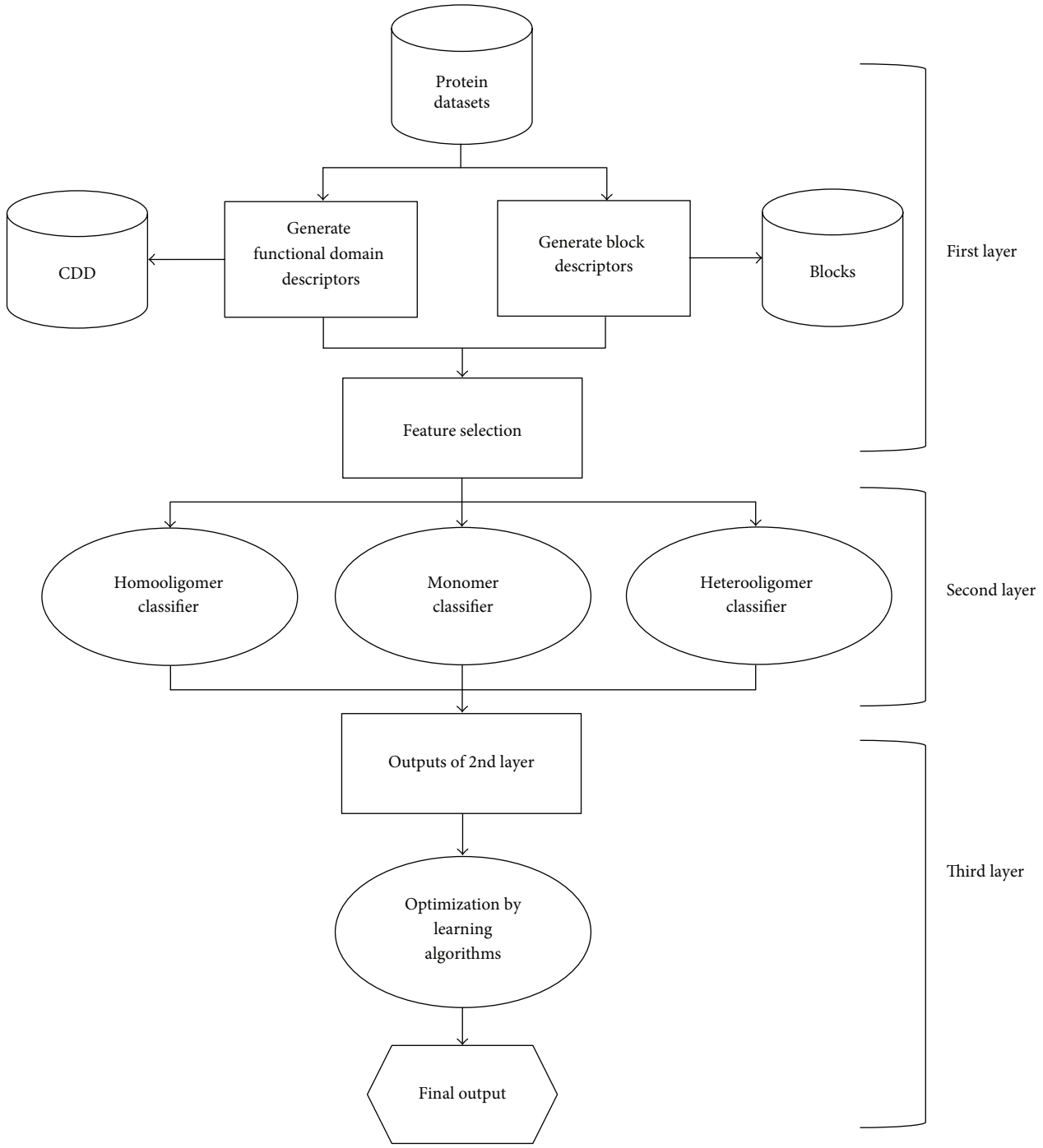


Figure 1: Flowchart of the three-layer architecture of classifiers.

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

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})},$$

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})},$$

$$F\text{-measure} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \quad (5)$$

For the third layer classifier evaluation criteria for the classification results, we used Kappa statistics and F-measure for viewing. Kappa statistics [34] are used to judge the classifier results, consistent with the random assortment. Its value is in the range of -1 to 1. When $K = 1$, it represents

Table 3: Performance comparison of model combination with SVM in 10-fold cross-validation for oligomer classification in the second layer.

	$F+P$		$B+P$		$B+F$		$B+F+P$	
	ACC (%)	MCC	ACC (%)	MCC	ACC (%)	MCC	ACC (%)	MCC
Monomer	75.75	0.552	78.91	0.579	82.64	0.663	82.64	0.663
Homooligomer								
Dimer	71.01	0.456	83.00	0.660	83.00	0.660	83.00	0.660
Trimer	90.22	0.806	95.48	0.909	95.48	0.909	95.47	0.907
Tetramer	84.21	0.638	90.12	0.776	93.41	0.854	93.41	0.854
Pentamer	80.48	0.649	98.57	0.973	98.57	0.973	98.57	0.973
Hexamer	84.27	0.712	94.27	0.887	96.94	0.939	96.94	0.939
Octamer	74.40	0.566	84.17	0.709	85.60	0.743	84.17	0.709
Decamer	85.83	0.759	95.83	0.928	98.33	0.971	98.33	0.971
Dodecamer	81.50	0.680	96.75	0.941	99.00	0.982	99.00	0.982
Overall	*81.49*	*0.658*	*92.27*	*0.848*	*93.79*	*0.879*	*93.61*	*0.874*
Heterooligomer								
Dimer	85.01	0.659	86.73	0.697	88.89	0.767	88.89	0.767
Trimer	80.87	0.600	91.93	0.835	91.93	0.835	92.07	0.836
Tetramer	85.79	0.706	92.82	0.853	94.48	0.889	94.48	0.889
Pentamer	77.50	0.577	89.17	0.799	93.33	0.886	93.33	0.886
Hexamer	84.86	0.698	90.42	0.812	90.42	0.812	93.72	0.875
Octamer	72.67	0.484	87.67	0.782	87.67	0.782	87.67	0.782
Decamer	92.50	0.873	97.50	0.958	100.00	1.000	97.50	0.958
Dodecamer	84.50	0.733	90.33	0.823	95.33	0.916	95.33	0.916
Overall	*82.96*	*0.666*	*90.82*	*0.820*	*92.76*	*0.861*	*92.88*	*0.863*

that the predicting results are different with random classifier prediction; $K = 0$ means predicting results are the same as random prediction; $K = -1$ represents that there is no effect and classification credibility. Here, we also use F-measure as the evaluation results of the standard classification. F-measure is a combination of precision and recall, with values from 0 to 1.

3. Results

3.1. Performance of Using Different Protein Features in the First Layer. In order to understand the different types of feature codes for the accuracy of the prediction structure, we trained the SVM classification model with 10-fold cross-validation evaluation model validity. Tables 1 and 2 show the 10-fold cross-validation prediction sensitivity, specificity, accuracy, and MCC on the monomer, homooligomer, and heterooligomer.

As can be seen from the results of the cross-validation, block composition in the monomer, homooligomer, and heterooligomer achieved an overall accuracy of 78.91%, 92.27%, and 91.13%, respectively. MCC was 0.579, 0.848, and 0.827, respectively. Since most of sensitivity performance has more than 80%, it indicates that a block composition method is indeed suitable for exhibition of protein characteristics and effectiveness of structure type classification. In the verification results of Functional domains (FunD) feature, the overall accuracy of monomer, homooligomer, and heterooligomer was 75.75%, 80.26%, and 79.93%, respectively. The results of FunD in homooligomer and heterooligomer were lower than the ones of block composition about 10%, while the sensitivity of homooctamer, heterooctamer, and heterodecamer are less than 50%. These results represent that FunD cannot be rendered for associated characteristics. The overall PseASA prediction accuracy is relatively low, that is, respectively, 68.40% and 73.05%. However, compared with the FunD, using PseASA method to predict heterooligomer, pentamer, octamer, and decamer is better at 86.67%, 86%, and 90% of sensitivity, respectively. In addition, the MCC of PseASA for prediction is generally lower, showing that the homology between the whole sequences is not high or that the same category of the sequence number and complexity increases, which makes it difficult to obtain correct predictions. Even if it does not contain pentamer, octamer, decamer, and dodecamer which have a high sequence homology, the overall accuracy of homo- and heterooligomers still reached 90.72% and 90.48%, respectively. To further enhance prediction accuracy, we used the second layer SVM to integrate the various features of the model output.

3.2. Performance of Model Combination to Enhance Oligomer Type Prediction Accuracy. The purpose of establishing the second layer is to integrate different predicted results of characteristic model in each category. We unitized different combinations of characteristic models, in which the model is constructed by three features referred to as B (Block), F (FunD), and P (PseASA). Table 3 displays that performance comparison of model combination in 10-fold cross-validation for oligomer classification in the second layer.

Table 4: Performance comparison of classification algorithms in 10-fold cross-validation and self-consistency test.

Algorithms	Test method					
	Cross-validation			Self-consistency		
	CCI (%)	Kappa	*F*-measure	CCI (%)	Kappa	*F*-measure
Bayes						
Bayes net	64.80	0.5017	0.608	65.02	0.5053	0.611
Naïve Bayes	39.91	0	0.228	39.91	0	0.228
Functions						
LibSVM	67.40	0.5288	0.616	68.60	0.5464	0.632
Logistic	67.28	0.5285	0.615	67.57	0.5326	0.619
Multilayer perceptron	64.01	0.4893	0.598	69.97	0.5694	0.657
Lazy						
IB1	51.91	0.3513	0.515	87.18	0.8218	0.869
IBk	58.45	0.4126	0.551	90.38	0.8682	0.902
KStar	62.43	0.463	0.581	88.10	0.8362	0.877
Meta						
AdaBoostM1	59.80	0.3909	0.493	59.80	0.3909	0.493
Bagging	66.27	0.5147	0.605	69.88	0.5671	0.65
Rules						
Conjunctive rule	59.80	0.3909	0.493	59.80	0.3909	0.493
Decision table	66.99	0.5189	0.601	67.22	0.5218	0.601
DTNB	67.12	0.5225	0.606	67.38	0.5268	0.61
Tree						
J48	66.45	0.5161	0.607	69.81	0.5639	0.646
Random forest	58.91	0.4306	0.566	90.02	0.8651	0.899
Random tree	54.65	0.3817	0.537	90.38	0.8682	0.902

*CCI is correctly classified instances.

In the result of the monomer combination $B + P$ with an accuracy of 78.91%, a difference of the combination of $F + P$ is about 3%. Using a combination of $B + F$ enhanced accuracy, improving from 78.91% to 82.64%. However, $B + F$ and $B + F + P$ combination exhibited less accuracy. The same situation also appears in the feature models combination for homo- and heterooligomers. Overall, $B + F$ model combinations can have better performance than using the single Block model. Most of the categories were improved from 1 to 6%. Therefore, this study will feature $B + F$ combination to construct the first layer and the second layer of the classification model.

3.3. Performance Comparison of Classification Algorithms in the Third Layer. In order to obtain unique results to determine an unknown protein quaternary structure type, we use a layer of the classifier to process the output of the second layer. By comparing different types of algorithms on power of data analysis and problem solving ability, we selected the better algorithm for constructing the third layer classifier. Studies using six types of typical algorithms are tested, that is, Bayes, Functions, Lazy, Rules, Trees, and Meta. The Oli6926 dataset is used in this training. We also used the two authentication methods, 10-fold cross-validation and self-consistency, to assess the learning effectiveness of the classifier.

In the results of 10-fold cross-validation, Correctly Classified Instances (CCI) of LibSVM and Logistic were 67.40% and 67.28%, respectively (Table 4). Kappa statistics was 0.5288 and 0.5285, respectively. And the *F*-measure was 0.616 and 0.615, respectively. These two algorithms have best predicted results. However, we found that the predictive accuracy and statistical value of LibSVM and Logistic are higher because most correct predictions which occurred in the large quaternary categories and in minor categories predictions, like pentamer, hexamer, and octamer, are completely ignored. Other algorithms, such as decision table and Bagging, also have a similar situation. Conversely, the accuracy of Random Forest, Random Tree, and IBk was 58.91%, 54.65%, and 58.45%, respectively. Kappa was 0.4306, 0.3817, and 0.4126, respectively. *F*-measure was 0.566, 0.537, and 0.551, respectively. Although the results of these three algorithms are not perfect, they are not susceptible to imbalance of data numbers.

The results of 10-fold cross-validation of LibSVM and Logistic in the self-consistency test were not significantly increased. Relative under the self-consistency verification, Random Forest, Random Tree, and IBk correctly predicted ratio reached about 90%, since they have good recognition capability for the known information. The prediction performance of Random Forest and IBk was similar in self-consistency which could achieve the highest value of 0.856 MCC. Since the cross-validation and prediction results of

Table 5: Comparison of results of different functional categories of proteins on QuaBingo and QuatIdent.

Protein categories	QuaBingo				QuatIdent			
	Sn (%)	Sp (%)	ACC (%)	MCC	Sn (%)	Sp (%)	ACC (%)	MCC
Immunity system	40.46	96.28	68.37	0.367	20.61	96.66	58.64	0.199
Enzyme	57.21	97.33	**77.27**	0.545	38.18	97.98	68.08	0.426
Cell cycle	44.44	96.53	70.49	0.410	14.82	97.80	56.31	0.176
Chaperone	45.95	96.62	71.28	0.426	20.27	98.99	59.63	0.313
Gene regulation	58.36	97.40	**77.88**	0.558	21.75	98.19	59.97	0.276
Transport proteins	57.80	97.36	77.58	0.552	21.67	97.86	59.77	0.258
Single transduction	59.16	97.45	**78.30**	0.566	11.97	98.42	55.19	0.167
Viral protein	42.73	96.42	69.57	0.391	10.00	98.75	54.38	0.156
Membrane protein	57.81	97.36	**77.59**	0.552	16.41	98.49	57.45	0.229
Molecular binding	63.37	97.71	**80.54**	0.611	27.11	98.47	62.79	0.351
Hormone	36.08	96.01	66.04	0.321	28.87	97.29	63.08	0.305
Others	60.03	97.50	78.77	0.575	17.18	98.61	57.89	0.247
Overall	51.95	97.00	74.47	0.490	20.74	98.13	59.43	0.259

Table 6: Top five features of block composition of oligomers.

Oligomer type	Top 5 features				
	1	2	3	4	5
Monomer	IPB002225A	IPB002347A	IPB000817A	IPB002347D	IPB013549A
Homooligomer					
Dimer	IPB000817A	IPB004045	IPB013572B	IPB001647	IPB003449A
Trimer	IPB007691D	IPB006052A	IPB006056A	IPB006175A	IPB006175B
Tetramer	IPB002347D	IPB003560D	IPB002198B	IPB002347B	IPB002347E
Pentamer	IPB007334A	IPB001931A	IPB013124E	IPB008681A	IPB012599D
Hexamer	IPB001564C	IPB001753C	IPB001980A	IPB001564A	IPB001564B
Octamer	IPB001354C	IPB013341B	IPB002682	IPB001354A	IPB001354B
Decamer	IPB000866A	IPB000866B	IPB013740	IPB003394A	IPB002587G
Dodecamer	IPB002177A	IPB002177B	IPB008331B	IPB014035B	IPB007664A
Heterooligomer					
Dimer	IPB003026B	IPB008386B	IPB000315A	IPB000219A	IPB012565
Trimer	IPB002353B	IPB012565	IPB003990A	IPB001003B	IPB003026B
Tetramer	IPB003026B	IPB012565	IPB010004A	IPB001664D	IPB002398F
Pentamer	IPB001280E	IPB003484D	IPB012420	IPB004333C	IPB006711D
Hexamer	IPB002919A	IPB003038	IPB008019A	IPB001591A	IPB001762
Octamer	IPB007659A	IPB004977B	IPB006574B	IPB002971G	IPB003539A
Decamer	IPB013124E	IPB002662B	IPB003417A	IPB000732A	IPB000817A
Dodecamer	IPB002682	IPB000353B	IPB001003B	IPB003597B	IPB006217A

Random Forest algorithms for minor categories were good, we finally chose the Random Forest classification algorithm as the third layer classifier in QuaBingo.

3.4. Performance Analysis. In order to understand the prediction capabilities of QuaBingo for different functional protein structures in the cell, we compared it with a known quaternary structure prediction tool QuatIdent [12] using an independent test. As shown in Table 5, the predicted result of the average sensitivity of QuaBingo was 51.95%. For the protein categories in the enzyme, gene regulation, membrane protein, single transduction, and molecular binding, there was better prediction of ACC from 77% to 80%. In the QuatIdent, the average sensitivity was 20.74%. These results illustrated the predicting method which is composed of functional domain and PsePSSM cannot obtain a correct identification result for most quaternary protein structures.

3.5. The Top Five Features of Block Composition of Oligomer on Oli8444. The feature extraction method of block composition is simple, which implies that a lot of useful information can be gained to help discover mechanisms of protein aggregation and serial modes. We will optimize block composition by feature selection, according to the degree of importance of each characteristic value, giving an F-score numerical score. The top five features are shown in Table 6. For example, the

IPB006052A of block composition in the top five features is TNF (Tumor Necrosis Factor) family of conserved sequence, which is found in trimeric CD40 ligand (PDB ID: 1ALY) in the training data and also found in the human Collagen X sequences (PDB ID: 1GR3). Human Collagen X needs to rely on the C1q domain to form a stable homotrimer. In existing data annotation, C1q and TNF-like domains overlap, and there are a number of important positions on the sequence of amino acids with high conservation and similar topology [35]. Much literature has confirmed that these amino acids play an important role in the formation of a hydrophobic core stability trimeric structure and formation of biologically active protein complexes [27, 35, 36]. In addition, many other features of block composition are associated with a particular function of protein. Thus, feature selection not only reduces the number of features in block composition but also can effectively identify characteristic patterns obviously related to the protein molecule aggregation phenomenon and hence distinguish quaternary structure among different oligomers.

3.6. Case Study. Thymidylate synthase (TS; EC 2.1.1.45) is an enzyme that can converts deoxyuridine monophosphate into deoxythymidine monophosphate and has an important position for necessary cell function about DNA replication and damage repair. The inhibition of TS is a way of cancer treatment that involves using inhibitors to interfere with DNA biosynthesis and create a disturbance in growth of cancer. TS is known that conserved protein from *E. coli* to human. Here, QuaBingo provides the testing results for several TS homologs, including 2KCE (*E. coli*), 4IQQ (*C. elegans*), 2TSR (rat), 4EB4 (mouse), 1HVY (human), and 1I00 (human). The testing results show that QuaBingo can correctly predict the quaternary structure, as homodimer, with TS phylogenetic distant homologs, and the sensitivity performance was 100%. This demonstrates that the QuaBingo may work within the example of phylogenetic homologous proteins.

4. Conclusions

In this study, we propose a feature extraction method based on a block of conserved protein sequence for the classification of protein quaternary structure. This method can overcome the problems of feature extraction encountered by functional domain composition: (1) some proteins may not contain any other known functional domains; and (2) corresponding known functional domains are too few to represent proteins. It is worth noting that the first problem has not yet been encountered in our proposed method, and the second problem was comprehensively solved using QuaBingo. The 10-fold cross-validation results showed that the overall accuracy of block composition of homo- and heterooligomers is 92.27% and 91.13%, respectively. Moreover, they are all 10% higher than the functional domain composition. These results demonstrate that the block composition can extract important and biologically meaningful features and thus enhance the prediction of protein quaternary structure.

Although many proteins exist as monomers, they may interact with another protein to form polymers or may further assemble to become a biologically relevant tetramer or octamer. Currently, most of these problems have not been solved through scientific research or verified by adequate information. In the future, as more and more data are added to pertinent databases, an accurate prediction system could be established that would greatly assist relevant research development. An online web server is freely available at http://predictor.nchu.edu.tw/QuaBingo/.

Competing Interests

The authors declare no competing interests.

Authors' Contributions

Chi-Wei Chen and Ren-Chao Guo wrote the experimental programs, participated in the experimental design, and constructed the QuaBiogo website. Ren-Chao Guo compiled the data set, participated in the experimental design, and wrote the paper. Chi-Hua Tung, Hui-Fuang Ng, and Yen-Wei Chu conceived of the study, participated in its design and coordination, and drafted the paper. All authors read and approved the paper.

Acknowledgments

This research was supported by Ministry of Science and Technology, Taiwan, under Grant no. 105-2221-E-216-021. The authors would like to thank Professor Jyung-Hurng Liu who provided the information of 3D Complex protein quaternary structure classification database.

References

[1] J. D. Bernal, "General introduction: structure arrangements of macromolecules," *Discussions of the Faraday Society*, vol. 25, pp. 7–18, 1958.

[2] H. Sund and K. Weber, "The quaternary structure of proteins," *Angewandte Chemie International Edition in English*, vol. 5, no. 2, pp. 231–245, 1966.

[3] G. Petsko and D. Ringe, *Protein Structure and Function*, New Science Press, London, UK, 2004.

[4] S. Jones and J. M. Thornton, "Principles of protein-protein interactions," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 93, no. 1, pp. 13–20, 1996.

[5] A. Dowierciał, P. Wilk, W. Rypniewski, W. Rode, and A. Jarmuła, "Crystal structure of mouse thymidylate synthase in tertiary complex with dUMP and raltitrexed reveals N-terminus architecture and two different active site conformations," *BioMed Research International*, vol. 2014, Article ID 945803, 7 pages, 2014.

[6] S. J. Smerdon, J. Jäger, J. Wang et al., "Structure of the binding site for nonnucleoside inhibitors of the reverse transcriptase of human immunodeficiency virus type 1," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 91, no. 9, pp. 3911–3915, 1994.

[7] R. Garian, "Prediction of quaternary structure from primary structure," *Bioinformatics*, vol. 17, no. 6, pp. 551–556, 2001.

[8] S.-W. Zhang, Q. Pan, H.-C. Zhang, Y.-L. Zhang, and H.-Y. Wang, "Classification of protein quaternary structure with support vector machine," *Bioinformatics*, vol. 19, no. 18, pp. 2390–2396, 2003.

[9] K.-C. Chou and Y.-D. Cai, "Predicting protein quaternary structure by pseudo amino acid composition," *Proteins: Structure, Function, and Bioinformatics*, vol. 53, no. 2, pp. 282–289, 2003.

[10] S.-W. Zhang, W. Chen, F. Yang, and Q. Pan, "Using Chou's pseudo amino acid composition to predict protein quaternary structure: a sequence-segmented PseAAC approach," *Amino Acids*, vol. 35, no. 3, pp. 591–598, 2008.

[11] X. Yu, C. Wang, and Y. Li, "Classification of protein quaternary structure by functional domain composition," *BMC Bioinformatics*, vol. 7, no. 1, article 187, 2006.

[12] H.-B. Shen and K.-C. Chou, "QuatIdent: a web server for identifying protein quaternary structural attribute by fusing functional domain and sequential evolution information," *Journal of Proteome Research*, vol. 8, no. 3, pp. 1577–1584, 2009.

[13] W. K. Kim, J. Park, and J. K. Suh, "Large scale statistical prediction of protein-protein interaction by potentially interacting domain (PID) pair," *Genome Informatics*, vol. 13, pp. 42–50, 2002.

[14] E. D. Levy, J. B. Pereira-Leal, C. Chothia, and S. A. Teichmann, "3D complex: a structural classification of protein complexes," *PLoS Computational Biology*, vol. 2, no. 11, pp. 1395–1406, 2006.

[15] W. Li and A. Godzik, "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics*, vol. 22, no. 13, pp. 1658–1659, 2006.

[16] Y. Huang, B. Niu, Y. Gao, L. Fu, and W. Li, "CD-HIT Suite: a web server for clustering and comparing biological sequences," *Bioinformatics*, vol. 26, no. 5, pp. 680–682, 2010.

[17] S. Henikoff and J. G. Henikoff, "Automated assembly of protein blocks for database searching," *Nucleic Acids Research*, vol. 19, no. 23, pp. 6565–6572, 1991.

[18] S. Henikoff, J. G. Henikoff, and S. Pietrokovski, "Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations," *Bioinformatics*, vol. 15, no. 6, pp. 471–479, 1999.

[19] J. G. Henikoff, E. A. Greene, S. Pietrokovski, and S. Henikoff, "Increased coverage of protein families with the blocks database servers," *Nucleic Acids Research*, vol. 28, no. 1, pp. 228–230, 2000.

[20] A. Marchler-Bauer, A. R. Panchenko, B. A. Shoemarker, P. A. Thiessen, L. Y. Geer, and S. H. Bryant, "CDD: a database of conserved domain alignments with links to domain three-dimensional structure," *Nucleic Acids Research*, vol. 30, no. 1, pp. 281–283, 2002.

[21] M. Punta, P. C. Coggill, R. Y. Eberhardt et al., "The Pfam protein families database," *Nucleic Acids Research*, vol. 40, no. 1, pp. D290–D301, 2012.

[22] I. Letunic, T. Doerks, and P. Bork, "SMART 7: recent updates to the protein domain annotation resource," *Nucleic Acids Research*, vol. 40, no. 1, pp. D302–D305, 2012.

[23] R. L. Tatusov, N. D. Fedorova, J. D. Jackson et al., "The COG database: an updated vesion includes eukaryotes," *BMC Bioinformatics*, vol. 4, article 41, 2003.

[24] W. Klimke, R. Agarwala, A. Badretdin et al., "The national center for biotechnology information's protein clusters database," *Nucleic Acids Research*, vol. 37, supplement 1, pp. D216–D223, 2009.

[25] D. H. Haft, J. D. Selengut, and O. White, "The TIGRFAMs database of protein families," *Nucleic Acids Research*, vol. 31, no. 1, pp. 371–373, 2003.

[26] Y. Ofran and B. Rost, "Protein-protein interaction hotspots carved into sequences," *PLoS Computational Biology*, vol. 3, no. 7, article e119, 2007.

[27] P. Venier, L. Varotto, U. Rosani et al., "Insights into the innate immunity of the Mediterranean mussel *Mytilus galloprovincialis*," *BMC Genomics*, vol. 12, no. 1, article 69, 2011.

[28] B. Petersen, T. N. Petersen, P. Andersen, M. Nielsen, and C. Lundegaard, "A generic method for assignment of reliability scores applied to solvent accessibility predictions," *BMC Structural Biology*, vol. 9, no. 1, article 51, 2009.

[29] P. Du, X. Wang, C. Xu, and Y. Gao, "PseAAC-builder: a cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions," *Analytical Biochemistry*, vol. 425, no. 2, pp. 117–119, 2012.

[30] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[31] W. Chen, S.-W. Zhang, Y.-M. Cheng, and Q. Pan, "Prediction of protein-protein interaction types using the decision templates based on multiple classier fusion," *Mathematical and Computer Modelling*, vol. 52, no. 11-12, pp. 2075–2084, 2010.

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

[33] Y.-W. Chen and C.-J. Lin, "Combining SVMs with various feature selection strategies," in *Feature Extraction*, pp. 315–324, Springer, Berlin, Germany, 2006.

[34] B. Di Eugenio and M. Glass, "The kappa statistic: a second look," *Computational Linguistics*, vol. 30, no. 1, pp. 95–101, 2004.

[35] A. Brass, K. E. Kadler, J. T. Thomas, M. E. Grant, and R. P. Boot-Handford, "The fibrillar collagens, collagen VIII, collagen X and the C1q complement proteins share a similar domain in their C-terminal non-collagenous regions," *FEBS Letters*, vol. 303, no. 2-3, pp. 126–128, 1992.

[36] L. Shapiro and P. E. Scherer, "The crystal structure of a complement-1q family protein suggests an evolutionary link to tumor necrosis factor," *Current Biology*, vol. 8, no. 6, pp. 335–340, 1998.

Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 7863706, 9 pages
http://dx.doi.org/10.1155/2016/7863706

Research Article

Reconstruction of the Fatty Acid Biosynthetic Pathway of *Exiguobacterium antarcticum* B7 Based on Genomic and Bibliomic Data

Regiane Kawasaki,[1] Rafael A. Baraúna,[1] Artur Silva,[1] Marta S. P. Carepo,[2] Rui Oliveira,[2] Rodolfo Marques,[2] Rommel T. J. Ramos,[1] and Maria P. C. Schneider[1]

[1]*Genomics and Systems Biology Center, Institute of Biological Sciences, Federal University of Pará, 66075-110 Belém, PA, Brazil*
[2]*Rede de Química e Tecnologia/Centro de Química Fina e Biológica, Chemistry Department, Universidade Nova de Lisboa, 2829-516 Costa da Caparica, Portugal*

Correspondence should be addressed to Rafael A. Baraúna; r.a.barauna@gmail.com

Received 10 November 2015; Accepted 16 June 2016

Academic Editor: Yongsheng Bai

Exiguobacterium antarcticum B7 is extremophile Gram-positive bacteria able to survive in cold environments. A key factor to understanding cold adaptation processes is related to the modification of fatty acids composing the cell membranes of psychrotrophic bacteria. In our study we show the *in silico* reconstruction of the fatty acid biosynthesis pathway of *E. antarcticum* B7. To build the stoichiometric model, a semiautomatic procedure was applied, which integrates genome information using KEGG and RAST/SEED. Constraint-based methods, namely, Flux Balance Analysis (FBA) and elementary modes (EM), were applied. FBA was implemented in the sense of hexadecenoic acid production maximization. To evaluate the influence of the gene expression in the fluxome analysis, FBA was also calculated using the $\log_2$FC values obtained in the transcriptome analysis at 0°C and 37°C. The fatty acid biosynthesis pathway showed a total of 13 elementary flux modes, four of which showed routes for the production of hexadecenoic acid. The reconstructed pathway demonstrated the capacity of *E. antarcticum* B7 to *de novo* produce fatty acid molecules. Under the influence of the transcriptome, the fluxome was altered, promoting the production of short-chain fatty acids. The calculated models contribute to better understanding of the bacterial adaptation at cold environments.

1. Introduction

Bacteria are increasingly used in industrial processes to produce chemicals, foods, and drugs, among other products [1]. The main biochemical pathways of bacteria may be manipulated and optimized to more efficiently produce compounds of industrial interest in various areas; for example, metabolic pathways of *Corynebacterium glutamicum* are rationally engineered to produce L-amino acids on an industrial scale [2]. To accomplish this task, specific tools are used such as FMM (from metabolite to metabolite) [3] Cytoscape [4], CellDesigner [5], SBW (Systems Biology Workbench) [6], COPASI (COmplex PAthway SImulator) [7], and COBRA (COnstraints Based Reconstruction and Analysis) toolbox [8].

The genomes of several bacterial strains have been sequenced and annotated and have been used in combination with biochemical and physiological data to reconstruct metabolic networks at a genome-scale [9]. Recently, genomic models were reconstructed for some bacterial species aiming to increase the amount and quality of data that has been annotated in either the literature or databases [10–12]. The draf t network generated from the annotated genome still requires significant manual curation for a comprehensive and accurate metabolic representation of the organism [13].

The need to develop automatic or at least semiautomatic methods to reconstruct metabolic networks from genome annotation is increasing because the number of complete genome sequences available is growing fast. Recent studies [13, 14] have highlighted the problems associated with

genome annotations and databases, which perform automatic reconstructions and, thus, require manual assessment. The currently available 96-step protocol, proposed by Thiele and Palsson [15], is a well-established process for the assembly, curation, and validation of metabolic reconstruction. This protocol is combined with computational tools, including the visualization and numerical calculation software package, MATLAB® (MathWorks, USA).

Constraint-based modeling is frequently used as final validation step of the reconstructed network. This step is extremely useful for simulating the phenotypic behavior under different physiological environments [16–18] thereby allowing assessing if the reconstructed network represents well the *in vivo* cellular system. Microbial adaptation to cold environments is one of the applications of these methods.

Psychrotrophic microorganisms have an optimal growth temperature higher than 15°C but are also able to grow and adapt to extremely cold environments, with temperatures of approximately 0°C [19]. Thus, the unique physicochemical characteristics of their habitat and the biological apparatus developed by these microorganisms to survive under these conditions render these organisms valuable sources of biotechnological processes. The cellular response to cold by psychrotrophic bacteria may be studied from a general standpoint with the advent of omics methods. Recently, the B7 strain of *Exiguobacterium antarcticum* was isolated from Ginger Lake sediments located in the Antarctic Peninsula Region (69°30′S, 65°W). This lake was formed due to the warming in the region, which led to partial melting of ice caps [20]. The genome of this strain was sequenced [21], and its response to cold was evaluated through differential expression of its genome at 37°C and 0°C using omics methods [22]. One of the mechanisms of cold adaptation of all psychrophilic or psychrotrophic organisms is the change in the chemical structures of the membrane phospholipids. The fatty acid chains become shorter and unsaturated at low temperatures. Accordingly, the fluidity of the membrane is kept intact [19, 23].

Bacterial *de novo* synthesis of fatty acids is regulated by the protein FapR [24], which is responsible for activating/disabling a regulon consisting of four operons in *E. antarcticum* B7. In cold, two of these operons are repressed (*fabH1-fabF* and *fabI*), and the expression levels of the other two remain unaltered [22]. The chemical components, which are included in this regulon, must be reconstructed and evaluated and then associated with their respective genetic elements to further understand this metabolic pathway and to reach more complete conclusions about its importance for adaptation to cold [23, 24]. Bioinformatics methods are used for *in silico* reconstruction of metabolic pathways [25].

In this work we present the *in silico* reconstruction of the fatty acid biosynthesis pathway of the *Exiguobacterium antarcticum* B7, based on linear programming (FBA) and convex cone method (elementary modes). The influence of transcriptome in FBA calculation was also evaluated.

2. Materials and Methods

2.1. Data Collection. The genomic data of *E. antarcticum* B7 in the formats .gbk and .fasta were collected from NCBI under accession number NC_018665. The metabolic pathway of the fatty acid biosynthesis was initially evaluated in the KEGG (Kyoto Encyclopedia of Genes and Genomes) database [26]. When necessary, the visualization of the genome of the bacterium was performed using Artemis software [27].

2.2. Preliminary Reconstruction. This step was performed following two methods: one semiautomatic and the other automatic. The tools within the KEGG databases were used essentially in the semiautomatic method. The.fasta file with the *E. antarcticum* B7 genome was submitted to the online tool KAAS (KEGG Automatic Annotation Server) [28], available at http://www.genome.jp/kaas-bin/kaas_main. The parameters chosen to run this software were as follows: (a) bidirectional Best Hit (BBH) Method, recommended for complete genomes, performs the search for orthologous genes between a specific group of organisms, and (b) prokaryote, the set of genes chosen, should be representative of the target organism, in this case, the bacterium *E. antarcticum* B7. Following the processing, a text file was generated (query.ko). Each line of this file is formed by two parameters: the first consists of the sequence identified (gene), and the second, when present, consists of the KO assignment, termed *K* number. This value indicates orthologous groups encoding the same enzymatic activity. Afterwards, the file generated is passed through a filter, an auxiliary computer software program (script) Python developed for the present study, which only selects *K* numbers and individually and increasingly commands per line into a new file (new_query.ko). This file was used as entry in the option User Data Mapping of the Pathway Mapping tool of the KEGG.

The automatic method essentially consisted of submitting the .fasta file of the *E. antarcticum* B7 genome to the online tool RAST (Rapid Annotation using Subsystems Technology) [29], available at http://rast.nmpdr.org/, to generate the drafts of the metabolic network and of the fatty acid biosynthesis pathway of the target microorganism. The final draft of this step was generated from the combination of the resulting pathways of the semiautomatic and automatic models. The common pathways were maintained, while surplus compounds, enzymes, and reactions, that is, present in some, but absent in others, were not directly excluded but were instead reserved for the curated step.

2.3. Manual Curation. The following steps were completed in this manual curation stage, following the protocol explained above. (i) Draft refinement: this phase began with the analysis of enzymes and reactions, components of the fatty acid biosynthesis pathway, by reading books and articles specifically on the subject. The objective was to diagnose the absence or presence of more than one element of the study pathway. The online databases KEGG, ENZYME [30], and SEED [31] were consulted to ratify the enzymes and the structures of the reactions.

(ii) Assessment of the stoichiometry and reversibility of the reactions: in this step, all model reactions were assessed and stoichiometrically corrected, if necessary. The biochemical data on the organism are very important to determine the reversibility of the reaction. For this purpose, the databases

Table 1: Genes, locus tags, and EC numbers identified in the draft of the fatty acid biosynthesis pathway of *E. antarcticum* B7. The features displayed were generated using the methods: semiautomatic and automatic.

Semiautomatic method			Automatic method		
Gene	*Locus tag*	EC number	Gene	*Locus tag*	EC number
accA	Eab7_2059	6.4.1.2	accA	Eab7_2059	6.4.1.2
accB	Eab7_0870	6.4.1.2	accB	Eab7_0870	6.4.1.14
accC	Eab7_0871	6.4.1.2	accC	Eab7_0871	6.4.1.2
accD	Eab7_2060	6.4.1.2	accD	Eab7_2060	6.4.1.14
fabD	Eab7_1760	2.3.1.39	fabD	Eab7_1760	2.3.1.39
fabH1	Eab7_1911	2.3.1.180	fabH1	Eab7_1911	2.3.1.180
fabF	Eab7_1910	2.3.1.179	fabF	Eab7_1910	2.3.1.179
fabG	Eab7_1795	1.1.1.100	fabG	Eab7_1795	1.1.1.100
fabZ	Eab7_2463	4.2.1.59	fabI	Eab7_1885	1.3.1.10
fabI	Eab7_1885	1.3.1.10			
fabK	Eab7_0377	1.3.1.9			
—	Eab7_2235	1.14.19.2			

(KEGG, SEED, and ENZYME) and the tool eQuilibrator [32] were used to analyze the reversibility of the reactions. Thus, the thermodynamic constraints were respected.

(iii) Addition of gene data and reaction location: the Artemis tool was used to identify the genes of the reactions (enzymes) from their locations in the genome assessed using the draft generated.

(iv) Assessment of Gene-Protein-Reaction (GPR) associations: in this step, the function of each gene is indicated. GPR associations were identified using databases of the organism and specific literature.

(v) Definition of constraints: some constraints were defined in the model in this manual curation step, including stoichiometric and thermodynamic constraints (through the reversibility and irreversibility of fluxes).

2.4. Metabolic Model Design. The metabolic model designed and refined following the manual curation step was converted into a mathematical representation, termed a stoichiometric matrix, which encouraged the development of a wide variety of computational tools to analyze network properties.

The constraints of capacity, which are the upper and lower limits defining the maximum and minimum fluxes allowed for the reactions, were added in this step. The inputs of the stoichiometric matrix are the coefficients of the metabolites in the reactions with negative values for consumed metabolites (substrates) and positive values when the metabolites are produced or secreted (products) (Additional File 1 see Supplementary Material available online at http://dx.doi.org/10.1155/2016/7863706).

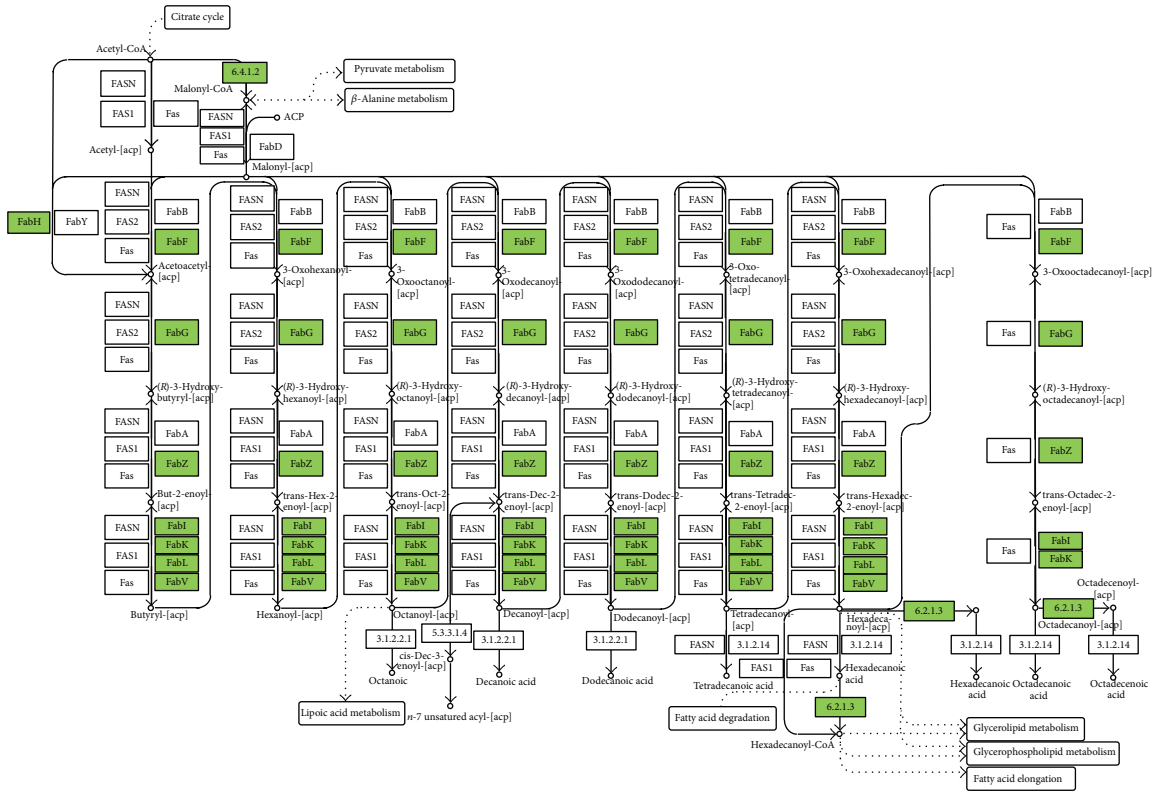
2.5. Metabolic Pathway Validation. The computational model sought to examine the metabolic capabilities and to evaluate the system properties they may perform under the constraints imposed on the cell. Thus, the final step in the reconstruction process consisted of assessing, evaluating, and validating the fatty acid biosynthesis pathway of *E. antarcticum* B7. The validation of that metabolic model was performed using simulation and flux analysis. The fatty acid biosynthesis pathway is well described in the literature because it is a highly conserved process among organisms, which facilitated its complete definition. Thus, most gaps had already been filled during the manual curation process.

3. Results and Discussion

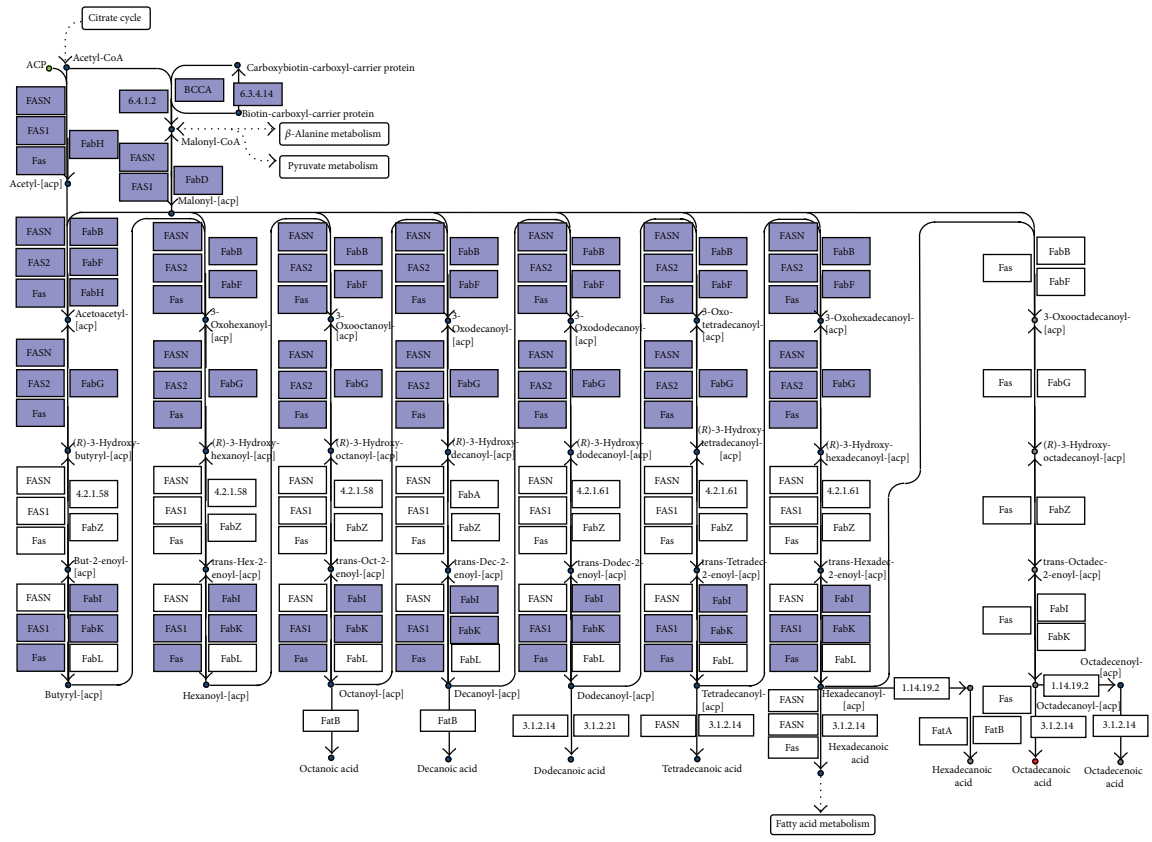
The expectation to understand the relationship between the genome and the physiology of a particular organism was a key incentive for reconstructing metabolic networks. Protocol adaptations using semiautomatic and automatic methods are necessary to reconstruct the metabolic networks of organisms with few reported data on their metabolic capabilities, including *E. antarcticum* B7.

3.1. Pathway Reconstruction Using the Semiautomatic Method. The draft of the metabolic network of *E. antarcticum* B7 was retrieved from the KEGG database [26]. The KEGG Metabolic Pathway tool was used to highlight the fatty acid biosynthesis pathway from the resulting draft of the metabolic network. The genes annotated and identified using KEGG and their respective enzymes are shown in Figure 1(a) in green, and the others are listed in white boxes. Table 1 shows the genes, locus tags, and enzymes identified using the KEGG Metabolic Pathway tool. A total of 11 *locus tags* associated with their respective genes and Enzyme Commission (EC) numbers were identified; only the locus tag Eab7_2235 has no added gene associated with it.

3.2. Pathway Reconstruction Using the Automatic Method. The RAST/SEED tool does not provide graphic display of the metabolic map draft as KEGG; for this purpose, it uses a standard table to list the 247 metabolic pathways that compose the network, regardless of whether they were identified in the genome of the microorganism. RAST identified 25 reactions, 40 compounds, and 20 EC numbers in the fatty acid biosynthesis pathway of *E. antarcticum* B7. Table 1 outlines the genes, locus tags, and enzymes identified at this step. Annotated genes identified by RAST and their respective enzymes



(a)



(b)

FIGURE 1: Drafts of the fatty acid biosynthesis pathway of *E. antarcticum* B7 bacteria. The drafts were designed using the following methods: (a) semiautomatic method, generated by the KEGG database, and (b) automatic method, generated by online tool RAST. Colored boxes indicate the possibility of the presence of enzymes in the pathway.

Table 2: Relationships between components of the *E. antarcticum* B7 fatty acid biosynthesis pathway following curation. The signals ⇒ and ⇔ indicate irreversible and reversible reactions, respectively.

Gene	Locus tag	EC number	Enzyme	Reaction	Fold change ($\log_2$FC)
accA	Eab7_2059	6.4.1.2	Acetyl-CoA carboxylase carboxyl transferase alpha subunit	ATP + acetyl-CoA + HCO_3^- ⇒ ADP + orthophosphate + malonyl-CoA	0.4562
accB	Eab7_0870	6.4.1.2	Acetyl-CoA carboxylase biotin-carboxyl carrier protein	ATP + acetyl-CoA + HCO_3^- ⇒ ADP + orthophosphate + malonyl-CoA	−0.05773
accC	Eab7_0871	6.4.1.2 6.4.1.14	Acetyl-CoA carboxylase, biotin carboxylase subunit	ATP + acetyl-CoA + HCO_3^- ⇒ ADP + orthophosphate + malonyl-CoA	−0.5623
accD	Eab7_2060	6.4.1.2	Acetyl-CoA carboxylase carboxyl transferase beta subunit	ATP + acetyl-CoA + HCO_3^- ⇒ ADP + orthophosphate + malonyl-CoA	−0.2811
fabD	Eab7_1760	2.3.1.39	ACP S-malonyl transferase	Malonyl-CoA + ACP ⇔ CoA + malonyl-(acp)	0.9448
fabH1	Eab7_1911	2.3.1.180	3-Oxoacyl-ACP synthase III	Acetyl-CoA + malonyl-(acp) ⇔ acetoacetyl-(acp) + CoA + CO_2	0.8512
fabF	Eab7_1910	2.3.1.179	3-Oxoacyl-ACP synthase II	Acetyl-(acp) + malonyl-(acp) ⇒ acetoacetyl-(acp) + CO_2 + ACP	0.6942
fabG	Eab7_1895	1.1.1.100	3-Oxoacyl-ACP reductase	Acetoacetyl-(acp) + NADPH + H^+ ⇔ (R)-3-hydroxybutanoyl-(acp) + $NADP^+$	0.8523
fabZ	Eab7_2463	4.2.1.59	3-Hydroxyacyl-ACP dehydratase	(R)-3-hydroxybutanoyl-(acp) ⇔ but-2-enoyl-(acp) + H_2O	0.12902
fabI	Eab7_1885	1.3.1.9 1.3.1.10	Enoyl-ACP reductase I	But-2-enoyl-(acp) + NADH + H^+ ⇔ butyryl-(acp) + NAD^+	0.2969
—	Eab7_2235	1.14.19.2	Acyl-ACP desaturase	Hexadecanoyl-(acp) + acceptor_reduced + O_2 ⇒ hexadecenoyl-(acp) + acceptor + $2H_2O$	1.3768

are colored in purple, and the others are listed in white boxes (Figure 1(b)).

The analysis of both fatty acid biosynthetic pathway drafts shows that the draft generated using KEGG apparently has the most complete flux, except for enzyme 6.3.4.14, which is exclusively present in the draft resulting from the RAST tool. The draft generated using RAST has a gap in which the enzymes FabA and FabB are not included in the pathway elongation process. The flux for the production of hexadecenoic acid is also absent from the pathway generated using RAST.

Artemis software was used to confirm the presence of all genes selected through the automatic and semiautomatic methods. The genes accABCD, fabD, fabH1, fabF, fabG, fabZ, and fabI and the locus tag Eab7_2235 were described in the genome of *E. antarcticum* B7, except fabK gene, which was detected only by the automatic method.

The KGML file produced by KEGG was submitted to the software KEGGtranslator [33] to be converted into a SBML (System Biology Markup Language) file [34]. This file was converted into an Excel spreadsheet using MATLAB functions. The files in SBML format and the Excel spreadsheets are the most used formats in metabolic reconstructions. The reactions and metabolites of the preliminary network generated using KEGG could be visualized in the spreadsheet.

The data generated using the RAST/SEED tool were analyzed and added to the first step of the process, supplementing the data collected using KEGG. The files generated with both platforms were used to manage the manual curation data.

The larger number of genes identified using KEGG (12) compared to those found using RAST/SEED (9) may be explained because the former uses orthology (KEGG Orthology (KO)) through protein homology to identify the so-called metabolite genes [35] in a genome, which facilitates finding gene-protein-reaction (GPR) associations.

3.3. Manual Curation of the Metabolic Pathway of De Novo Fatty Acid Synthesis. The reactions of the fatty acid synthesis pathway were annotated and refined. The metabolites were organized into two compartments (cytoplasm and extracellular compartment) based on the location of the enzymes associated with each pathway. The cofactors and the reversibility of the reactions were compiled from the data published in the literature and online tools (ENZYME and BRENDA). The EC number was noted, and the genes were identified. A summary of those results is shown in Table 2. Thermodynamic analysis of the reactions revealed that malonyl-CoA synthesis from acetyl-CoA (AcCoA) is an irreversible process; similar to the process regarding the *fabF* gene, the Eab7_2235 locus tag, and the extracellular metabolites the other processes are reversible.

It is very important to assess the quality of the annotated genome submitted to the online tools during curation. The literature categorically states that the quality of the reconstructed network directly depends on the annotated genome of the organism. The rule is to use the latest updated version of the annotated genome [36–38].

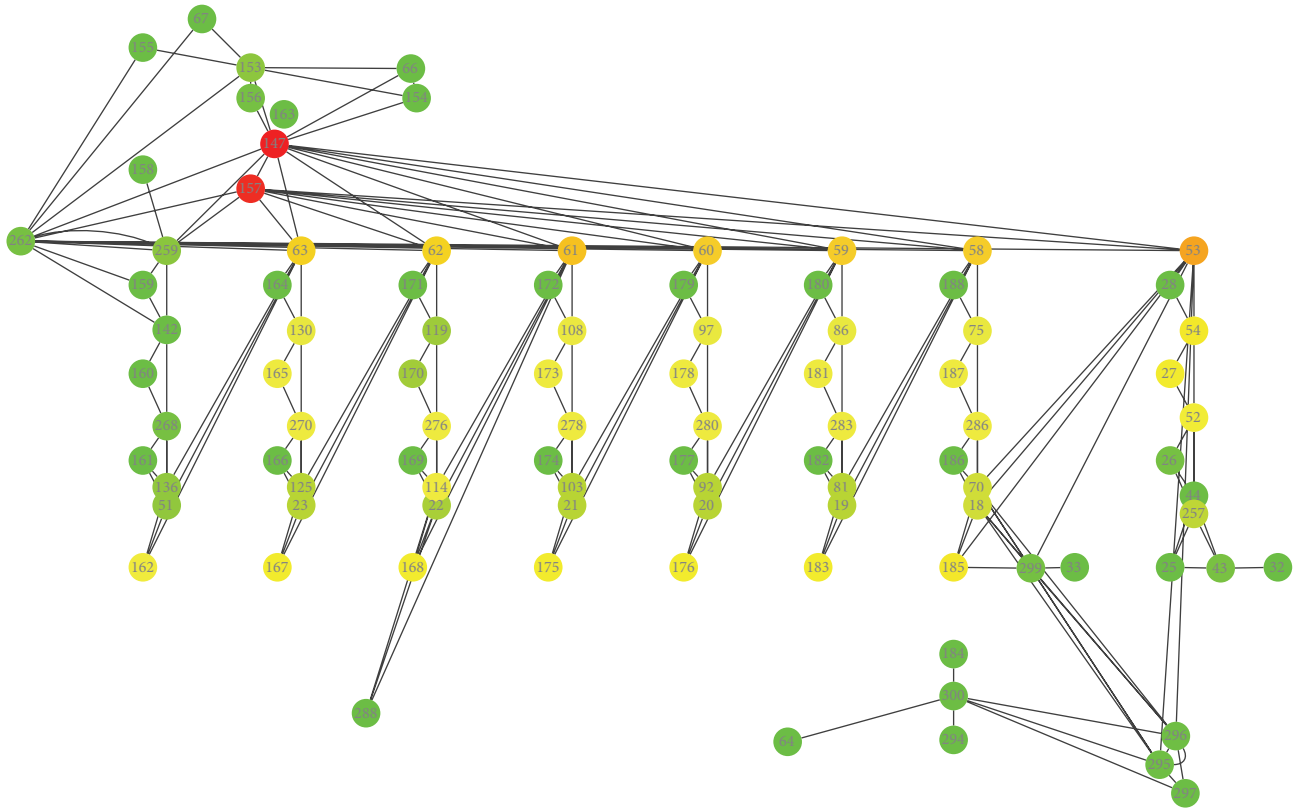


Figure 2: Layout of the fatty acid biosynthesis pathway generated using Cytoscape. Green vertices: fewer connections. Yellow vertices: regular number of connections. Red vertices: large number of connections. This network is a free model scale.

3.4. Curated Metabolic Network. The previous step added constraints regarding the stoichiometry (chemical balance), thermodynamics (reversibility of reactions), and physiology (cofactors used) of the study model. The result was a system of equations that describes the cell metabolism according to the metabolites of interest. The mathematical representation of this model essentially consisted of describing the performance of the fatty acid biosynthesis pathway using a stoichiometric matrix. This data structure consists of 54 metabolites and 59 reactions, resulting in a 54×59 stoichiometric matrix (Additional File 1). Additional File 2 shows the list of biochemical reactions identified in the curated model.

The reconstructed pathway model was converted into SBML in the used MATLAB toolbox. The SBML file was validated using the tool SBML validator and was then submitted to the tool Cytoscape, which generated the network of Figure 2. The gene-protein-reaction (GPR) representation therein describes the degree of connectivity of each enzyme in the pathway. Vertices with few connections are in green, the vertices with regular numbers of connections are in yellow, and the vertices with large numbers of connections are in red. The network connectivity obeys a scale-free model [39].

3.5. Flux Balance Analysis (FBA). The FBA was coded in MATLAB implementing a constrained linear program using the GLPK (GNU Linear Programming Kit) linear optimization library [8]. All fluxes were calculated in percentage of the input flux of AcCoA (reaction 39), which was fixed to 100. Hexadecenoic acid is the key metabolic product; thus the respective flux (reaction 41) was set as the maximization target for FBA. To improve convergence, upper and lower bounds were $[0, 100]$ for irreversible reactions and $[-100, 100]$ for reversible reaction. The final optimized fluxes are shown in Additional File 3. The target maximum, reaction R41 in Additional File 3, was 7.69, which may be read as 7.69 moles of hexadecenoic acid produced for every 100 moles of AcCoA consumed per unit time per unit cell mass.

Figure 3(a) shows the generated flux plot, which shows the variation occurring between the response fluxes, with the majority, approximately 37, showing positive values smaller than 20, while 15 are above that range.

3.6. Influence of Transcriptome in FBA Calculation. Log base 2-fold change values ($\log_2$FC) obtained *in vitro* by Dall'Agnol and colleagues [22] were used to evaluate the influence of differential expression in the FBA calculation. These values were obtained by comparison of RPKM (reads per kilobase per million reads sequenced) generated in the transcriptome of the bacterium at 0°C and 37°C. The $\log_2$FC of genes that composes the fatty acid biosynthesis pathway is shown in Table 2.

As presented by Dall'Agnol and colleagues [22], the aerobic energetic metabolism of *E. antarcticum* B7 at 0°C is repressed, and a fraction of the acetyl-CoA is probably used as

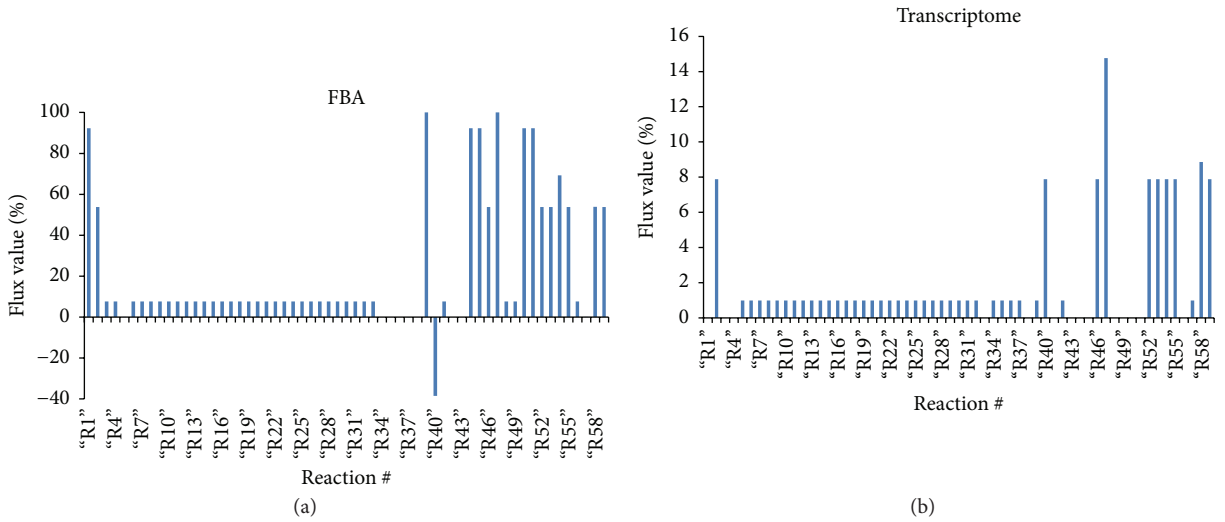


Figure 3: Flux graph generated using the software MATLAB, which uses methods based on FBA constraints. The vertical axis represents the fluxes calculated from the input of 100 moles of acetyl-CoA (AcCoA). The horizontal axis represents the reactions participating in the fatty acid biosynthesis pathway. The blue bars determine the output percentage for each pathway reaction. (a) FBA for maximizing the production of hexadecenoic acid. (b) FBA calculated with the $\log_2$FC values obtained from the transcriptome.

a substrate to synthesize short-chain fatty acids in cold. The synthesis begins with the conversion of acetyl-CoA to malonyl-CoA catalyzed by the multimeric enzyme encoded by the genes *accABCD*. In FBA analysis, only half of the input of acetyl-CoA (100 mM) is converted to malonyl-CoA, which binds to the acyl carrier protein (ACP) at 0°C (Figure 3(b)). As stated earlier, the other half of acetyl-CoA is probably used for energy generation.

The remaining route is cyclical, being the reactions catalyzed by enzymes encoded by *fabF*, *fabG*, *fabZ*, and *fabI* genes. At each round, two carbons are added to the growing chain of fatty acid. In these reactions, the flow of metabolites remains unchanged until the fatty acid molecule reaches a size of eight carbons (octadecanoyl-ACP in reaction 38) where the percentage of the flow amount decreases (Figure 3(b)). These results are consistent with the previously published data which affirm that bacteria decrease their fatty acid chains to survive in cold. These short fatty acid molecules will be converted into membrane phospholipids in order to maintain the fluidity of this biological barrier.

3.7. Elementary Flux Modes. The calculation of elementary modes was performed in MATLAB using the Metatool toolbox [40] (*modo_elementar.m* code in the Supplemental Information). A total of 13 elementary flux modes were found for the fatty acid biosynthesis pathway of *E. antarcticum* B7 (Additional File 4). Of these, only 4 elementary modes (2, 5, 8, and 11) have a positive nonzero coefficient for reaction 41, which indicates that the target product hexadecenoic acid may only be generated by one of these four elementary modes. The routes identified in EM2 and EM5 begin at the second reaction (R2) which is catalyzed by the enzyme FabD. The value of R2 for both elementary modes indicates a considerable production of malonyl-CoA. The remaining reactions of EM2 and EM5 are presented in Additional File 4. In EM8 and EM11, the routes begin from R1. The value 1 for this reaction indicates a lower activity, reflecting in a lower production of hexadecenoic acid. Regarding reaction 41, the values of elementary modes 2 and 5, in this case 1, are higher than the values of 8 and 11 (0.142857 each), indicating that both elementary modes 2 and 5 produce 1 mol hexadecenoic acid when they are active, while elementary modes 8 and 11 produce 0.142857 moles.

4. Conclusions

The first metabolic pathway of *E. antarcticum* B7, reconstructed following the steps defined in this work, suggests that the protocol used is a suitable tool for further metabolic reconstruction studies. Almost all the first steps of the process were automated; however, manual curation was, as usual, laborious because it required an intensive search for available data.

The metabolic pathway of fatty acid biosynthesis was representative and consistent under the limits and boundary conditions set. The FBA and elementary mode methods were used to examine the hexadecenoic acid production potential of the reconstructed pathway. The application of constraint-based modeling revealed being very useful to assess network operation plasticity, even if the intracellular kinetics are unknown. The *in silico* analysis performed using FBA enabled a quantitative assessment of hexadecenoic acid production potential.

Finally, a key issue involves deciding when to stop the process and to consider the reconstruction finalized, at least temporarily. This decision is usually based on the reconstruction purpose. The most complete metabolic model currently available is the *E. coli* model, which has been researched and refined for over 10 years [41–45]. Other studies constantly updating their models are *Homo sapiens*, with three

reconstructions [46–48], and *S. cerevisiae*, with more than a dozen reconstructions, including two in 2013 [49, 50]. The protocol reported in the present study may be used to compile several data pieces available in the literature aimed at proposing possible metabolic pathways, thereby enabling deeper research of the metabolism under study.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

This study was supported by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

References

[1] J. W. Lee, D. Na, J. M. Park, J. Lee, S. Choi, and S. Y. Lee, "Systems metabolic engineering of microorganisms for natural and non-natural chemicals," *Nature Chemical Biology*, vol. 8, no. 6, pp. 536–546, 2012.

[2] J. Becker and C. Wittmann, "Systems and synthetic metabolic engineering for amino acid production—the heartbeat of industrial strain development," *Current Opinion in Biotechnology*, vol. 23, no. 5, pp. 718–726, 2012.

[3] C.-H. Chou, W.-C. Chang, C.-M. Chiu, C.-C. Huang, and H.-D. Huang, "FMM: a web server for metabolic pathway reconstruction and comparative analysis," *Nucleic Acids Research*, vol. 37, no. 2, pp. W129–W134, 2009.

[4] 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.

[5] A. Funahashi, Y. Matsuoka, A. Jouraku, M. Morohashi, N. Kikuchi, and H. Kitano, "CellDesigner 3.5: a versatile modeling tool for biochemical networks," *Proceedings of the IEEE*, vol. 96, no. 8, pp. 1254–1265, 2008.

[6] F. T. Bergmann and H. M. Sauro, "SBW—a modular framework for systems biology," in *Proceedings of the 37th Winter Simulation Conference (WSC '06)*, pp. 1637–1645, December 2006.

[7] S. Hoops, S. Sahle, R. Gauges et al., "COPASI—a COmplex PAthway SImulator," *Bioinformatics*, vol. 22, no. 24, pp. 3067–3074, 2006.

[8] J. Schellenberger, R. Que, R. M. T. Fleming et al., "Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0," *Nature Protocols*, vol. 6, no. 9, pp. 1290–1307, 2011.

[9] M. Terzer, N. D. Maynard, M. W. Covert, and J. Stelling, "Genome-scale metabolic networks," *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, vol. 1, no. 3, pp. 285–297, 2009.

[10] 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, article 661, 2013.

[11] M. Fondi, I. Maida, E. Perrin et al., "Genome-scale metabolic reconstruction and constraint-based modelling of the Antarctic bacterium *Pseudoalteromonas haloplanktis* TAC125," *Environmental Microbiology*, vol. 17, no. 3, pp. 751–766, 2015.

[12] 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, article 661, 2013.

[13] A. M. Feist, M. J. Herrgård, I. Thiele, J. L. Reed, and B. Ø. Palsson, "Reconstruction of biochemical networks in microorganisms," *Nature Reviews Microbiology*, vol. 7, no. 2, pp. 129–143, 2009.

[14] J. L. Reed, I. Famili, I. Thiele, and B. Ø. Palsson, "Towards multidimensional genome annotation," *Nature Reviews Genetics*, vol. 7, no. 2, pp. 130–141, 2006.

[15] I. Thiele and B. Ø. Palsson, "A protocol for generating a high-quality genome-scale metabolic reconstruction," *Nature Protocols*, vol. 5, no. 1, pp. 93–121, 2010.

[16] J. S. Edwards, R. U. Ibarra, and B. O. Palsson, "In silico predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data," *Nature Biotechnology*, vol. 19, no. 2, pp. 125–130, 2001.

[17] R. Mahadevan, B. Ø. Palsson, and D. R. Lovley, "In situ to in silico and back: elucidating the physiology and ecology of *Geobacter spp.* using genome-scale modelling," *Nature Reviews Microbiology*, vol. 9, no. 1, pp. 39–50, 2011.

[18] J. D. Trawick and C. H. Schilling, "Use of constraint-based modeling for the prediction and validation of antimicrobial targets," *Biochemical Pharmacology*, vol. 71, no. 7, pp. 1026–1035, 2006.

[19] R. Y. Morita, "Psychrophilic bacteria," *Bacteriological Reviews*, vol. 39, no. 2, pp. 144–167, 1975.

[20] R. Mulvaney, N. J. Abram, R. C. A. Hindmarsh et al., "Recent Antarctic Peninsula warming relative to Holocene climate and ice-shelf history," *Nature*, vol. 489, no. 7414, pp. 141–144, 2012.

[21] A. R. Carneiro, R. T. J. Ramos, H. Dall'Agnol et al., "Genome sequence of *Exiguobacterium antarcticum* B7, isolated from a biofilm in Ginger Lake, King George Island, Antarctica," *Journal of Bacteriology*, vol. 194, no. 23, pp. 6689–6690, 2012.

[22] H. Dall'Agnol, R. A. Baraúna, P. de Sá et al., "Omics profiles used to evaluate the gene expression of *Exiguobacterium antarcticum* B7 during cold adaptation," *BMC Genomics*, vol. 15, no. 1, article 986, 2014.

[23] A. Casanueva, M. Tuffin, C. Cary, and D. A. Cowan, "Molecular adaptations to psychrophily: the impact of 'omic' technologies," *Trends in Microbiology*, vol. 18, no. 8, pp. 374–381, 2010.

[24] G. E. Schujman, L. Paoletti, A. D. Grossman, and D. de Mendoza, "FapR, a bacterial transcription factor involved in global regulation of membrane lipid biosynthesis," *Developmental Cell*, vol. 4, no. 5, pp. 663–672, 2003.

[25] E. Pitkänen, J. Rousu, and E. Ukkonen, "Computational methods for metabolic reconstruction," *Current Opinion in Biotechnology*, vol. 21, no. 1, pp. 70–77, 2010.

[26] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, vol. 27, no. 1, pp. 29–34, 1999.

[27] T. Carver, S. R. Harris, M. Berriman, J. Parkhill, and J. A. McQuillan, "Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data," *Bioinformatics*, vol. 28, no. 4, pp. 464–469, 2012.

[28] Y. Moriya, M. Itoh, S. Okuda, A. C. Yoshizawa, and M. Kanehisa, "KAAS: an automatic genome annotation and pathway reconstruction server," *Nucleic Acids Research*, vol. 35, no. 2, pp. W182–W185, 2007.

[29] R. K. Aziz, D. Bartels, A. Best et al., "The RAST Server: rapid annotations using subsystems technology," *BMC Genomics*, vol. 9, article 75, 2008.

[30] A. Bairoch, "The ENZYME database in 2000," *Nucleic Acids Research*, vol. 28, no. 1, pp. 304–305, 2000.

[31] R. Overbeek, R. Olson, G. D. Pusch et al., "The SEED and the rapid annotation of microbial genomes using Subsystems Technology (RAST)," *Nucleic Acids Research*, vol. 42, no. 1, pp. D206–D214, 2014.

[32] A. Flamholz, E. Noor, A. Bar-Even, and R. Milo, "EQuilibrator—the biochemical thermodynamics calculator," *Nucleic Acids Research*, vol. 40, no. 1, pp. D770–D775, 2012.

[33] C. Wrzodek, A. Dräger, and A. Zell, "KEGGtranslator: visualizing and converting the KEGG PATHWAY database to various formats," *Bioinformatics*, vol. 27, no. 16, pp. 2314–2315, 2011.

[34] M. Hucka, A. Finney, H. M. Sauro et al., "The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models," *Bioinformatics*, vol. 19, no. 4, pp. 524–531, 2003.

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

[36] J. J. Hamilton and J. L. Reed, "Software platforms to facilitate reconstructing genome-scale metabolic networks," *Environmental Microbiology*, vol. 16, no. 1, pp. 49–59, 2014.

[37] G. J. E. Baart and D. E. Martens, "Genome-scale metabolic models: reconstruction and analysis," *Methods in Molecular Biology*, vol. 799, pp. 107–126, 2012.

[38] V. Lacroix, L. Cottret, P. Thébault, and M.-F. Sagot, "An introduction to metabolic networks and their structural analysis," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 5, no. 4, pp. 594–617, 2008.

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

[40] T. Pfeiffer, I. Sánchez-Valdenebro, J. C. Nuño, F. Montero, and S. Schuster, "METATOOL: for studying metabolic networks," *Bioinformatics*, vol. 15, no. 3, pp. 251–257, 1999.

[41] E. Almaas, B. Kovács, T. Vicsek, Z. N. Oltvai, and A.-L. Barabási, "Global organization of metabolic fluxes in the bacterium *Escherichia coli*," *Nature*, vol. 427, no. 6977, pp. 839–843, 2004.

[42] J. S. Edwards and B. Ø. Palsson, "The *Escherichia coli* MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 10, pp. 5528–5533, 2000.

[43] I. M. Keseler, J. Collado-Vides, S. Gama-Castro et al., "EcoCyc: a comprehensive database resource for *Escherichia coli*," *Nucleic Acids Research*, vol. 33, pp. D334–D337, 2005.

[44] J. D. Orth, T. M. Conrad, J. Na et al., "A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism-2011," *Molecular Systems Biology*, vol. 7, article 535, 2011.

[45] A. M. Feist and B. Ø. Palsson, "The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*," *Nature Biotechnology*, vol. 26, no. 6, pp. 659–667, 2008.

[46] N. C. Duarte, S. A. Becker, N. Jamshidi et al., "Global reconstruction of the human metabolic network based on genomic and bibliomic data," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 6, pp. 1777–1782, 2007.

[47] T. Y. Kim, S. B. Sohn, Y. B. Kim, W. J. Kim, and S. Y. Lee, "Recent advances in reconstruction and applications of genome-scale metabolic models," *Current Opinion in Biotechnology*, vol. 23, no. 4, pp. 617–623, 2012.

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

[49] B. D. Heavner, K. Smallbone, N. D. Price, and L. P. Walker, "Version 6 of the consensus yeast metabolic network refines biochemical coverage and improves model performance," *Database*, vol. 2013, Article ID bat059, 2013.

[50] T. Österlund, I. Nookaew, S. Bordel, and J. Nielsen, "Mapping condition-dependent regulation of metabolism in yeast through genome-scale modeling," *BMC Systems Biology*, vol. 7, article 36, 2013.

Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 1687623, 7 pages
http://dx.doi.org/10.1155/2016/1687623

Research Article

Social Determinants of Chronic Prostatitis/Chronic Pelvic Pain Syndrome Related Lifestyle and Behaviors among Urban Men in China: A Case-Control Study

Yan Wang,[1] Chen Chen,[1] Changcai Zhu,[1] Liang Chen,[1] Qingrong Han,[2] and Huarong Ye[3]

[1]*Department of Preventive Medicine, School of Public Health, Wuhan University of Science and Technology, 2 Huangjia Lake West Road, Hongshan District, Wuhan, Hubei 430065, China*
[2]*Yiling Hospital, 31 East Lake Avenue, Yiling District, Yichang, Hubei 443100, China*
[3]*China Resources & WISCO General Hospital, 209 Metallurgy Road, Qingshan District, Wuhan, Hubei 430080, China*

Correspondence should be addressed to Changcai Zhu; zcc621120@163.com

Received 15 December 2015; Accepted 29 June 2016

Academic Editor: Yongsheng Bai

Purpose. In order to find key risk factors of chronic prostatitis/chronic pelvic pain syndrome (CP/CPPS) among urban men in China, an age-matched case-control study was performed from September 2012 to May 2013 in Yichang, Hubei Province, China. *Methodology.* A total of 279 patients and 558 controls were recruited in this study. Data were collected by a self-administered questionnaire, including demographics, diet and lifestyle, psychological status, and a physical exam. Conditional logistic regression model was used to analyze collected data. *Results.* Chemical factors exposure, night shift, severity of mood, and poor self-health cognition were entered into the regression model, and result displayed that these four factors had odds ratios of 1.929 (95% CI, 1.321–2.819), 1.456 (95% CI, 1.087–1.949), 1.619 (95% CI, 1.280–2.046), and 1.304 (95% CI, 1.094–1.555), respectively, which suggested that these four factors could significantly affect CP/CPPS. *Conclusion.* These results suggest that many factors affect CP/CPPS, including biological, social, and psychological factors.

1. Introduction

Chronic prostatitis/chronic pelvic pain syndrome (CP/CPPS) is a chronic pain disorder, which is characterized by the presence of noninfectious pelvic or perineal pain lasting longer than 3 months. The International Prostatitis Collaborative Network of the National Institutes of Health (IPCN-NIH) has provided detailed criteria for diagnosing CP/CPPS [1, 2]. According to various epidemiological studies using different methodologies, the prevalence of CP/CPPS varies from approximately 8% to 20% worldwide [3–6]. A population-based survey estimated that the prevalence of CP/CPPS-like symptoms in China is 4.5% [4].

Although there have been many basic and clinical research studies, the exact etiology, pathophysiology, and mechanism of CP/CPPS remain indeterminate. This syndrome is currently considered to be a multifactorial medical condition and requires a multimodal treatment approach [7]. New diagnostic/therapeutic criteria targeted to the urinary, psychosocial, organ-specific, infection, neurological/systemic, and tenderness (UPOINT) system were developed by Shoskes et al. in 2009 to classify patients suffering with CP/CPPS and, more importantly, to direct appropriate therapy [8]. Multimodal therapy based on the UPOINT phenotype system greatly improves the symptoms of CP/CPPS [9, 10].

Currently, CP syndromes represent an important healthcare problem worldwide [11]. Furthermore, many studies have suggested that CP places a large financial burden on patients and society. Chronic prostatitis increases healthcare expenditures directly and indirectly (e.g., unemployment). The average total costs (direct and indirect) for 3 months of CP treatment is USD 1099 per person for resource consumption, with an expected annual total cost per person of USD 4397 [12]. In China, treatment for CP is relatively costly (USD 1151 or 8059 CNY per person) [13].

A cross-sectional study about CP/CPPS patients has been done previously; it was reported that there were many potential factors that might have an influence on CP/CPPS, including smoking, drinking tea, sedentariness, overstress, economic pressure, and self-health cognition [14, 15]. However, as a cross-sectional study they could not give more information about risk factors and affected degree. Therefore, based on the results of the cross-sectional study several factors were selected and a case-control study was conducted for further study of risk factors of CP/CPPS.

2. Materials and Methods

2.1. Study Setting and Target Population. This case-control study was conducted in Yiling District, Yichang, Hubei Province, China. The study included one administrative district, one economic development zone, eight towns, and three townships. With a total area of 3424 square kilometers and a population of 546,500, this district is the most populous administrative region of Yichang. Most of the residents are drivers or laborers. The target populations of this study are composed of male patients diagnosed with CP/CPPS for the first time in Yiling Hospital and healthy men without CP/CPPS symptoms in Yiling District. Data were collected from 279 patients with CP/CPPS between September 2012 and May 2013. Five hundred and fifty-eight males matched by age (1 : 2) without CP/CPPS symptoms were enrolled into the control group.

2.2. Selection Criteria for Cases and Controls. Diagnosis of patients with CP/CPPS was based on the IPCN-NIH consensus [1] following a critical medical interview, a digital rectal examination, a prostate secretion examination after prostate massage, and urinalysis. Patients who suffered from pyuria and other genitourinary symptoms that may be associated with benign prostatic hyperplasia, chronic bacterial prostatitis, acute prostatitis, or genitourinary diseases other than CP/CPPS were excluded from the study to minimize the measurement bias. The control subjects were healthy men without CP/CPPS symptoms who participated in a physical examination in Yiling Hospital. People who had been diagnosed with benign prostatic hyperplasia, chronic bacterial prostatitis, acute prostatitis, or other genitourinary diseases were excluded from the controls. Every case was matched with two controls by age within ±2 years.

2.3. Ethical Approval. The study was approved by the Medical Ethics Committee of Wuhan University of Science and Technology. Written informed consent was obtained from the study subjects who were assured of confidentiality by the use of anonymous questionnaires. Verbal consent was also sought from community leaders prior to the focus group discussions.

2.4. Data Collection. An anonymous questionnaire was designed by experts on health statistics and urology, and all of the collectors were trained before the questionnaire survey. According to the results of a presurvey, the questionnaire and plan were modified. The questionnaire was proved to be valid. Finally, the questionnaire consisted of five major domains of items, including demographics (age, degree of education, occupation, medical insurance, and average monthly income), lifestyle (frequency of eating fast food, time of using a mobile per day, smoking, drinking, drinking tea, and sedentariness), working situation (occupational hazards, night shift), psychological status (severity of bad mood, stress of work, economic stress of family, self-health cognition, and spousal relationship), and a physical exam. The item of occupational hazards determined by type of work was asked by questionnaire investigators. The questionnaire was self-administered after informed consent unless the participant was illiterate.

2.5. Statistical Analysis. The collected questionnaires were collated manually for the first time. They were then checked again while the data were entered into the database set up by Epi database. Data were analyzed using descriptive statistical methods, the chi-square test, and conditional logistic regression analysis. The Statistical Package for Sciences (SPSS) software version 17.0 (SPSS Inc., Chicago, IL, USA) was used for data analysis. The significance level was set at $P < 0.05$.

3. Results

3.1. Sociodemographic Characteristics. A total of 279 patients and 558 controls were recruited to participate in this retrospective survey. All of the participants completed the questionnaire. The age of all of the subjects ranged from 24 to 59 years. The mean age of all of the subjects was 43.30 ± 7.92 years. The mean ages of patients and controls were 43.57±8.09 years and 43.16 ± 7.84 years, respectively.

For cases, 72.04% of subjects were aged between 30 and 49 years. Only one subject was illiterate, and most (82.79%) had a middle school or high school diploma. A total of 78.14% of the patients were employed as skilled laborers. A total of 86.73% of the patients had medical insurance for urban workers except for 3 who were reported as self-paying when seeing a doctor. More than 80% of the patients had an average monthly income less than 3000 CNY (483 USD), and only 10 patients had an average monthly salary of more than 5000 CNY (805 USD) (Table 1).

Similarly, 72.22% of the controls were aged between 30 and 49 years. Most of the controls were employed as skilled laborers. A total of 86.73% of the controls had medical insurance for urban workers, whereas five subjects were reported as self-paying when seeing a doctor. Almost 80% of the control subjects' average monthly income was less than 3000 CNY (483 USD). All of the controls received school education, and 82.79% had a secondary school or high school diploma (Table 1).

3.2. Chi-Square Test for Single-Factor Analysis. Table 2 shows the values of significant risk factors. Table 3 shows the results of single-factor analysis. Several factors including "time of using a mobile per day," "smoking," "drinking," "drinking tea," and "sedentariness" did not have significant difference between patients and controls. The odds of CP/CPPS for subjects who were exposed to chemical factors in occupational workplace increased by approximately 104% (95%

Table 1: Sociodemographic characteristics of respondents (n = 837).

Characteristic	Value	Cases n (%)	Controls n (%)
Age	24~	15 (5.38%)	42 (7.53%)
	30~	70 (25.09%)	136 (24.37%)
	40~	131 (46.95%)	267 (47.85%)
	50~	63 (22.58%)	113 (20.25%)
Degree of education	No formal education	1 (0.36%)	0 (0.00%)
	Primary school	2 (0.72%)	19 (3.40%)
	Junior high school	105 (37.63%)	179 (32.08%)
	Senior high or technical secondary school	126 (45.16%)	258 (46.24%)
	Junior college or above	45 (16.13%)	102 (18.28%)
Occupation	Unskilled; for example, trader, farming	20 (7.17%)	34 (6.09%)
	Skilled labor; for example, driver, blue-collar worker	218 (78.14%)	450 (80.65%)
	Professional; for example, teacher, healthcare worker, office worker	6 (2.15%)	20 (3.58%)
	Others	35 (12.54%)	54 (9.68%)
Medical insurance	Medical insurance for urban workers	242 (86.73%)	484 (86.73%)
	Rural cooperative medical service	30 (10.75%)	57 (10.21%)
	Commercial insurance	3 (1.08%)	6 (1.08%)
	Self-paying	3 (1.08%)	5 (0.90%)
	Others	1 (0.36%)	6 (1.08%)
Average monthly income	Less than 1000 CNY	40 (14.34%)	71 (12.72%)
	1000–2000 CNY	106 (37.99%)	206 (36.92%)
	2000–3000 CNY	78 (27.96%)	161 (28.85%)
	3000–4000 CNY	31 (11.11%)	62 (11.11%)
	4000–5000 CNY	14 (5.02%)	30 (5.38%)
	More than 5000 CNY	10 (3.58%)	28 (5.02%)

Table 2: Values of significant risk factors.

Variable	Value		
CP/CPPS	No = 0	Yes = 1	
Occupational hazards	No factors = 0	Physical factors = 1	Chemical factors = 2
	Biological factors = 3	Other factors = 4	
Night shift	Yes = 1	No = 2	
Frequency of eating fast food	Frequent = 1	Once in a while = 2	Never = 3
Time of using a mobile per day	Less than half an hour = 1	Half an hour to one hour = 2	
	One hour to two hours = 3	Two hours to three hours = 4	
	Three hours to four hours = 5	More than four hours = 6	
Severity of mood (e.g., sadness, anxiety, depression)	Not a bit = 1	A bit = 2	Medium = 3
	Very serious = 4	Extremely serious = 5	
Stress of work	Not a bit = 1	A bit = 2	Medium = 3
	Very serious = 4	Extremely serious = 5	
Economic stress of family	Not a bit = 1	A bit = 2	Medium = 3
	Very serious = 4	Extremely serious = 5	
Self-health cognition	Beyond comparison = 1	Very good = 2	Good = 3
	Common = 4	Bad = 5	
Spousal relationship	Very good = 1	Common = 2	Bad = 3

CP/CPPS: chronic prostatitis/chronic pelvic pain syndrome.

confidence interval [CI], 1.416–2.941). A night shift caused the odds of CP/CPPS to increase by approximately 53% for workers. The odds of CP/CPPS increased with the frequency of eating fast food (odds ratio, 1.32; 95% CI, 1.022–1.703). The time of using a mobile phone per day also affected the odds of CP/CPPS with positive correlation (odds ratio, 1.152; 95% CI, 1.027–1.293), which means, when spending more time on mobile phone, there will be more risk on suffering

Table 3: Results of the chi-square test for single-factor analysis.

Factors	*B*	SE	Wald	df	Sig.	Exp (*B*)	95.0% CI for Exp (*B*)	
No occupational hazards	—	—	15.372	4	0.004	—	—	—
Chemical factors	0.713	0.187	14.620		**0.000**	**2.040**	1.416	2.941
Not on night shift	−0.426[a]	0.144	8.769	1	**0.003**	**0.653**	0.493	0.866
Low frequency of eating fast food	−0.278[a]	0.130	4.540	1	**0.033**	**0.758**	0.587	0.978
Time of using a mobile per day	0.142	0.059	5.793	1	**0.016**	**1.152**	1.027	1.293
Severity of mood (e.g., sadness, anxiety, depression)	0.525	0.115	20.848	1	**0.000**	**1.691**	1.350	2.119
Stress of work	0.280	0.083	11.317	1	**0.001**	**1.323**	1.124	1.557
Economic stress of family	0.155	0.076	4.147	1	**0.042**	**1.167**	1.006	1.354
Self-health cognition	0.325	0.086	14.355	1	**0.000**	**1.384**	1.170	1.637
Spousal relationship	0.414	0.141	8.593	1	**0.003**	**1.513**	1.147	1.996

[a]Two negative values were obtained because the variables "not on night shifts" and "low frequency of eating fast food" were two protective factors of CP/CPPS.

Table 4: Results of conditional logistic regression for multiple-factor analysis.

Factors	*B*	SE	Wald	df	Sig.	Exp (*B*)	95.0% CI for Exp (*B*)	
No occupational hazards	—	—	11.919	4	0.018	—	—	—
Physical factors	0.269	0.255	1.117	1	0.291	1.309	0.794	2.158
Chemical factors	0.657	0.193	11.546	1	**0.001**	**1.929**	1.321	2.819
Biological factors	0.417	0.651	0.411	1	0.522	1.518	0.424	5.438
Others	0.360	0.250	2.074	1	0.150	1.433	0.878	2.339
Not on night shift	−0.376[a]	0.149	6.365	1	**0.012**	**0.687**	0.513	0.920
Severity of mood (e.g., sadness, anxiety, depression)	0.482	0.120	16.213	1	**0.000**	**1.619**	1.280	2.046
Self-health cognition	0.265	0.090	8.741	1	**0.003**	**1.304**	1.094	1.555

[a]A negative value was obtained because the variable "not on night shifts" was a protective factor of CP/CPPS.

CP/CPPS. The risk of CP/CPPS increased with the degree of mood (e.g., sadness, anxiety, and depression), stress of work, economic stress of family, the level of self-health cognition, and spousal relationship, with odds ratios of 1.691 (95% CI, 1.350–2.119), 1.323 (95% CI, 1.124–1.557), 1.167 (95% CI, 1.006–1.354), 1.384 (95% CI, 1.170–1.637), and 1.513 (95% CI, 1.147–1.996), respectively.

3.3. Conditional Logistic Regression for Multiple-Factor Analysis. Nine significant factors selected by the chi-square test were used to build the regression model. Occupational hazards were set as classification variables, using the forward Wald method. Probability values for stepwise entry and removal were 0.05 and 0.10, respectively. Finally, four factors were included in the regression model (Table 4). Chemical factors, night shift, degree of mood (e.g., sadness, anxiety, and depression), and poor self-health cognition increased the odds of CP/CPPS by 93%, 46%, 62%, and 30%, respectively.

4. Discussion

Prostatitis has become increasingly more common, and age is not a limiting factor. Given the complexity of prostatitis, a systematic classification was provided by the NIH, including category I (acute bacterial prostatitis), category II (chronic bacterial prostatitis), category III (chronic bacterial prostatitis/CPPS), and category IV (asymptomatic inflammatory prostatitis) [16]. Among these four categories, chronic bacterial prostatitis/CPPS has become a recognized intractable disease. In China, a previous study indicated that most urological surgeons considered chronic bacterial prostatitis/CPPS as a clinical syndrome, and different treatment protocols were used to relieve pain, improve voiding symptoms, and improve quality of life [17]. Treatment protocols for bacterial prostatitis/CPPS that are used by urological surgeons include drug therapy (95%), changing lifestyle (88.9%), and psychotherapy (79.9%). Drugs include botanical drugs (84.5%), adrenergic alpha-antagonists (79%), and antibiotics (64%) [17]. Based on the results of a review of Medline articles, many individual therapies, including antibiotics, anti-inflammatory medications, neuromodulators, alpha blockers, pelvic floor physical therapy, and cognitive behavior therapy, have been evaluated in the treatment of CP/CPPS. Each therapy has been found to have varying efficiency in alleviating symptoms [18]. In a clinical study, the effect of combination therapy was analyzed in a single specialized prostatitis clinic; the result showed that a clinically appreciable reduction of ≥6 points of the total NIH-CPSI score was achieved in 77.5% of patients subjected to combination therapy for a period of 6 months [19]. Multimodal therapy that includes pharmacotherapy, baths, prostate massage, and pelvic floor physical therapy may help patients to control the symptoms of CP/CPPS [11]. Another study in China showed that 65% of CP patients undergo long-term routine treatment 12 times per year, and

most CP patients are not satisfied with the effectiveness of the costly treatment [13].

In addition, the quality of life obviously declines in patients with CP/CPPS. Wenninger et al. [20] evaluated the effect of chronic nonbacterial prostatitis on the quality of life and functional status. They found that the mean Sickness Impact Profile score in men with chronic nonbacterial prostatitis was 7.5, which was greater than that for the general population. Additionally, the most severe effect of CP/CPPS appeared to be on social interaction in their study [20].

Thus, many epidemiologic studies have been done to find key risk factors of the disease and help people change their lifestyle to reduce the risk of CP/CPPS. Lan et al. [21] carried out a multicenter case-control study between June 2005 and May 2008 in China. They showed that urinary system infection, frequent masturbation, a cold climate, prostatomegaly, mental stress, high altitude, little exercise, and alcohol addiction might be risk factors of CP/CPPS [21]. This study also found that severity of mood (e.g., sadness, anxiety, and depression) might be related to CP/CPPS. Zhao et al. [22] conducted a retrospective case-control study of clinical data from 322 CP/CPPS patients (case group) and 341 non-CP/CPPS patients (control group). They showed an association between foreskin length and the odds of CP/CPPS. When the foreskin length covered up more than half of the glans penis, there were greater odds for CP/CPPS [22]. A literature review performed by Pontari and Ruggieri showed that the symptoms of CP/CPPS appeared to result from interplay between psychological factors and dysfunction in the immune, neurological, and endocrine systems [23]. Another study performed in northwest China suggested that oxidative stress and cytokines might be involved in the pathological process and aggravation of symptoms [24]. These results suggested that further experimental study, like cellular and molecular level research, should be done. This study could not explain whether exposure to cold was a risk factor because patients and controls came from the same region. A multinational observational study indicated that factors of the severity of symptoms of CP/CPPS varied between regions [25]. This previous study showed that effects of exposure to cold ($P = 0.1856$) and abdominal symptoms ($P = 0.1119$) were highest in Finland, those of education level ($P = 0.0151$), sexual activity ($P = 0.0574$), and erectile dysfunction ($P = 0.0151$) were highest in Germany/Switzerland, and those of age ($P = 0.0698$) were highest in Italy [25].

Dietary habit is often considered to have a considerable effect on CP/CPPS. According to our chi-square test results, among the lifestyle factors in the questionnaire, only "frequency of eating fast food" was significant. Finally, this factor was not included in the regression model. Many foods, such as spicy food, coffee, alcoholic beverages, and tea, can exacerbate the symptoms of patients with CP/CPPS, while others, such as docusate, psyllium, water, herbal teas, and polycarbophil, can ameliorate symptoms [26]. Another case-control study showed that the risk factors of CP include spicy food and drinking alcohol [27]. In our study, the degree of mood (e.g., sadness, anxiety, and depression) was significant in single-factor analysis and multiple-factor analysis. Many previous studies obtained similar results that depression might be involved in the development and clinical course of CP/CPPS [28–32]. In fact, depression and CP/CPPS may share, at least in part, several common pathophysiological mechanisms [7, 33]. It was demonstrated that the prostate gland responds to emotional stimulation through the autonomic nervous system; an experimental evidence also supports the theory that psychological stress may contribute to dysfunction of the prostate [34]. However, other studies have suggested that CP patients experience an increased risk of depressive disorders compared with non-CP patients [30], which meant psychological problems occurred after the disease. Our study also showed that night shifts might increase the risk of suffering from CP/CPPS by approximately 46%. When working at night, workers needed to overcome much more difficulties, such as fatigue, sleepiness, loneliness, and inattention. It was reported that staying up late was a risk factor of CP/CPPS [27]. In our study, although chemical factors had an effect on CP/CPPS; this factor needs to be studied further because the chemical substances were unknown. Some chemical substances and/or their metabolites might have a negative effect on the prostate when they penetrate the human body's protective barrier and enter the body. But this needs further studies because occupational hazards were only based on job duties without fast-field analysis, and the chemical substances were unknown. Self-health cognition was also a significant factor, but there might have been bias. Diseases, especially those that can reduce the quality of life, can change one's self-health cognition to a great extent. Therefore, despite the statistical significance of self-health cognition, it had no practical significance.

There were some limitations in our study. First, because this study was a retrospective study, it could not provide sufficient evidence of a causal relationship between risk factors and CP/CPPS. Second, the questionnaire was self-designed, despite its reliability, and bias might have been present. Third, chemical occupational factors have not been divided into particular toxicity or hazard, which may be confusing.

5. Conclusions

Many studies have shown a relationship between CP/CPPS and potential risk factors. An increasing number of researchers support the viewpoint that CP/CPPS is a clinical syndrome with an unclear or unknown pathogenesis. Our study shows that chemical factors, night shifts, the moods of sadness, anxiety, depression, and poor self-health cognition may affect CP/CPPS. Although there are many limitations in this study, our results might provide instructive information for patients and urologists.

Disclosure

Yan Wang and Chen Chen are coauthors.

Competing Interests

None of the authors declare competing financial interests.

Authors' Contributions

Yan Wang, Changcai Zhu, Liang Chen, and Chen Chen designed the questionnaire and performed the survey and data analysis; Qingrong Han and Huarong Ye carried out the physical examinations; Yan Wang and Changcai Zhu wrote the paper. All authors have read and approved the final version of the paper and agree with the order of presentation of the authors. Yan Wang and Chen Chen equally contributed to this paper.

Acknowledgments

The authors would like to thank all of the workers who took part in this research, especially the doctors at Yiling Hospital, Yichang, Hubei Province. This study was funded by three research projects, called the Male Reproductive Health Status and Intervention Countermeasures in Three Gorges Region of Yichang (WJ2015Z087), a Case-Control Study on Influential Factors of Chronic Prostatitis among Iron and Steel Enterprise Male Workers (WJ2015MB256), which were provided by the Health and Family Planning Commission of Hubei Province, and a grant from the Undergraduates Innovation Fund of Hubei Province (201310488041), which was provided by Hubei Provincial Department of Education.

References

[1] J. N. Krieger, L. Nyberg, and J. C. Nickel, "NIH consensus definition and classification of prostatitis," *The Journal of the American Medical Association*, vol. 282, no. 3, pp. 236–237, 1999.

[2] B. A. Mahal, J. M. Cohen, S. A. Allsop et al., "The role of phenotyping in chronic prostatitis/chronic pelvic pain syndrome," *Current Urology Reports*, vol. 12, no. 4, pp. 297–303, 2011.

[3] R. O. Roberts, M. M. Lieber, T. Rhodes, C. J. Girman, D. G. Bostwick, and S. J. Jacobsen, "Prevalence of a physician-assigned diagnosis of prostatitis: the olmsted county study of urinary symptoms and health status among men," *Urology*, vol. 51, no. 4, pp. 578–584, 1998.

[4] C.-Z. Liang, H.-J. Li, Z.-P. Wang et al., "The prevalence of prostatitis-like symptoms in China," *The Journal of Urology*, vol. 182, no. 2, pp. 558–563, 2009.

[5] C. E. C. C. Ejike and L. U. S. Ezeanyika, "Prevalence of chronic prostatitis symptoms in a randomly surveyed adult population of urban-community-dwelling Nigerian males," *International Journal of Urology*, vol. 15, no. 4, pp. 340–343, 2008.

[6] J. Q. Clemens, R. T. Meenan, M. C. O'Keeffe Rosetti, T. Kimes, and E. A. Calhoun, "Prevalence of and risk factors for prostatitis: population based assessment using physician assigned diagnoses," *The Journal of Urology*, vol. 178, no. 4, part 1, pp. 1333–1337, 2007.

[7] J. C. Nickel, D. A. Shoskes, and F. M. E. Wagenlehner, "Management of chronic prostatitis/chronic pelvic pain syndrome (CP/CPPS): the studies, the evidence, and the impact," *World Journal of Urology*, vol. 31, no. 4, pp. 747–753, 2013.

[8] D. A. Shoskes, J. C. Nickel, R. R. Rackley, and M. A. Pontari, "Clinical phenotyping in chronic prostatitis/chronic pelvic pain syndrome and interstitial cystitis: a management strategy for urologic chronic pelvic pain syndromes," *Prostate Cancer and Prostatic Diseases*, vol. 12, no. 2, pp. 177–183, 2009.

[9] D. A. Shoskes, J. C. Nickel, and M. W. Kattan, "Phenotypically directed multimodal therapy for chronic prostatitis/chronic pelvic pain syndrome: a prospective study using UPOINT," *Urology*, vol. 75, no. 6, pp. 1249–1253, 2010.

[10] X. Guan, C. Zhao, Z.-Y. Ou et al., "Use of the UPOINT phenotype system in treating Chinese patients with chronic prostatitis/chronic pelvic pain syndrome: a prospective study," *Asian Journal of Andrology*, vol. 17, no. 1, pp. 120–123, 2015.

[11] S.-J. Xia, D. Cui, and Q. Jiang, "An overview of prostate diseases and their characteristics specific to Asian men," *Asian Journal of Andrology*, vol. 14, no. 3, pp. 458–464, 2012.

[12] E. A. Calhoun, M. McNaughton Collins, M. A. Pontari et al., "The economic impact of chronic prostatitis," *Archives of Internal Medicine*, vol. 164, no. 11, pp. 1231–1236, 2004.

[13] C.-Z. Liang, H.-J. Li, Z.-P. Wang et al., "Treatment of chronic prostatitis in Chinese men," *Asian Journal of Andrology*, vol. 11, no. 2, pp. 153–156, 2009.

[14] R.-L. Gong, T. Lv, Q.-R. Han et al., "A correlation analysis of male reproductive system disorders and behavioral factors," *Chinese Journal of Disease Control & Prevention*, vol. 19, no. 8, p. 3, 2015.

[15] L. Wang, T. Lyu, Q. Han, R. Gong, and C. Zhu, "Effects of psychological factor on urogenital system health among childbearing-aged men in the three-gorges region," *Chinese Journal of Family Planning*, vol. 22, no. 12, p. 4, 2014.

[16] J. C. Nickel, L. M. Nyberg, and M. Hennenfent, "Research guidelines for chronic prostatitis: consensus report from the first National Institutes of Health International Prostatitis Collaborative Network," *Urology*, vol. 54, no. 2, pp. 229–233, 1999.

[17] K. Zhang, B. Xu, Y.-X. Xiao et al., "Chinese urologists' practice patterns of diagnosing and treating chronic pelvic pain syndrome: a questionnaire survey," *Journal of Peking University. Health sciences*, vol. 46, no. 4, pp. 578–581, 2014.

[18] A. S. Polackwich and D. A. Shoskes, "Chronic prostatitis/chronic pelvic pain syndrome: a review of evaluation and therapy," *Prostate Cancer and Prostatic Diseases*, vol. 19, no. 2, pp. 132–138, 2016.

[19] V. Magri, E. Marras, A. Restelli, F. M. E. Wagenlehner, and G. Perletti, "Multimodal therapy for category III chronic prostatitis/chronic pelvic pain syndrome in UPOINTS phenotyped patients," *Experimental and Therapeutic Medicine*, vol. 9, no. 3, pp. 658–666, 2015.

[20] K. Wenninger, J. R. Heiman, I. Rothman, J. P. Berghuis, and R. E. Berger, "Sickness impact of chronic nonbacterial prostatitis and its correlates," *The Journal of Urology*, vol. 155, no. 3, pp. 965–968, 1996.

[21] T. Lan, Y.-M. Wang, Y. Chen et al., "Multicenter study of the risk factors for chronic prostatitis," *Medical Journal of Chinese People's Liberation Army*, vol. 35, no. 1, p. 4, 2010.

[22] Y.-Y. Zhao, D.-L. Xu, F.-J. Zhao et al., "Redundant prepuce increases the odds of chronic prostatitis/chronic pelvic pain syndrome (CP/CPPS)," *Asian Journal of Andrology*, vol. 16, no. 5, pp. 774–777, 2014.

[23] M. A. Pontari and M. R. Ruggieri, "Mechanisms in prostatitis/chronic pelvic pain syndrome," *Journal of Urology*, vol. 172, no. 3, pp. 839–845, 2004.

[24] T. Lan, Y. Wang, Y. Chen et al., "Influence of environmental factors on prevalence, symptoms, and pathologic process of chronic prostatitis/chronic pelvic pain syndrome in northwest China," *Urology*, vol. 78, no. 5, pp. 1142–1149, 2011.

[25] F. M. E. Wagenlehner, M. Spangenberg, V. Magri et al., "Risk factors analysis in patients with chronic prostatitis/chronic pelvic pain syndrome (CP/CPPS) from Finland, Germany, Italy and Switzerland: results of a multinational observational study," *European Urology Supplements*, vol. 12, no. 1, p. 1, 2013.

[26] A. S. Herati, B. Shorter, A. K. Srinivasan et al., "Effects of foods and beverages on the symptoms of chronic prostatitis/chronic pelvic pain syndrome," *Urology*, vol. 82, no. 6, pp. 1376–1380, 2013.

[27] L. Qing-dong, *Case-control study of risk factors of chronic prostatitis [M.S. thesis]*, Central South University, Changsha, China, 2011.

[28] J. Q. Clemens, S. O. Brown, and E. A. Calhoun, "Mental health diagnoses in patients with interstitial cystitis/painful bladder syndrome and chronic prostatitis/chronic pelvic pain syndrome: a case/control study," *Journal of Urology*, vol. 180, no. 4, pp. 1378–1382, 2008.

[29] D. A. Tripp, J. C. Nickel, J. R. Landis, L. W. Yan, and J. S. Knauss, "Predictors of quality of life and pain in chronic prostatitis/chronic pelvic pain syndrome: findings from the National Institutes of Health Chronic Prostatitis Cohort Study," *BJU International*, vol. 94, no. 9, pp. 1279–1282, 2004.

[30] S.-D. Chung, C.-C. Huang, and H.-C. Lin, "Chronic prostatitis and depressive disorder: a three year population-based study," *Journal of Affective Disorders*, vol. 134, no. 1–3, pp. 404–409, 2011.

[31] G.-X. Zhang, W.-J. Bai, T. Xu, and X.-F. Wang, "A preliminary evaluation of the psychometric profiles in Chinese men with chronic prostatitis/chronic pelvic pain syndrome," *Chinese Medical Journal*, vol. 124, no. 4, pp. 514–518, 2011.

[32] S. G. Ahn, S. H. Kim, K. I. Chung, K. S. Park, S. Y. Cho, and H. W. Kim, "Depression, anxiety, stress perception, and coping strategies in Korean military patients with chronic prostatitis/chronic pelvic pain syndrome," *Korean Journal of Urology*, vol. 53, no. 9, pp. 643–648, 2012.

[33] J. C. Nickel, "Understanding chronic prostatitis/chronic pelvic pain syndrome (CP/CPPS)," *World Journal of Urology*, vol. 31, no. 4, pp. 709–710, 2013.

[34] H. K. Ja, W. K. Soo, and J.-S. Paick, "Quality of life and psychological factors in chronic prostatitis/chronic pelvic pain syndrome," *Urology*, vol. 66, no. 4, pp. 693–701, 2005.

Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 8479587, 9 pages
http://dx.doi.org/10.1155/2016/8479587

Research Article

A Shortest Dependency Path Based Convolutional Neural Network for Protein-Protein Relation Extraction

Lei Hua[1] and Chanqin Quan[2]

[1]*Department of Computer and Information Sciences, Hefei University of Technology, Hefei 230009, China*
[2]*Department of Computer and Information Sciences, Kobe University, Kobe 6578501, Japan*

Correspondence should be addressed to Lei Hua; hualeilxf@163.com

Received 4 March 2016; Revised 4 June 2016; Accepted 15 June 2016

Academic Editor: Rita Casadio

The state-of-the-art methods for protein-protein interaction (PPI) extraction are primarily based on kernel methods, and their performances strongly depend on the handcraft features. In this paper, we tackle PPI extraction by using convolutional neural networks (CNN) and propose a shortest dependency path based CNN (sdpCNN) model. The proposed method (1) only takes the sdp and word embedding as input and (2) could avoid bias from feature selection by using CNN. We performed experiments on standard Aimed and BioInfer datasets, and the experimental results demonstrated that our approach outperformed state-of-the-art kernel based methods. In particular, by tracking the sdpCNN model, we find that sdpCNN could extract key features automatically and it is verified that pretrained word embedding is crucial in PPI task.

1. Introduction

Biomedical relations play an important role in biologic processes and are widely researched in the field of biomedical natural language processing (BioNLP). PPI task aims to extract protein interactions; for example, in sentence *"The distribution of actin filaments is altered by profilin overexpression,"* the interaction between protein entities *"actin"* and *"profilin"* would be extracted. A number of databases, such as BIND [1], MINT [2], and IntAct [3], had been created to store structured interactions. However, the biomedical literature regarding protein interactions is expanding rapidly, making it difficult for these databases to keep up with the latest protein-protein interactions. Consequently, effective and automatic protein-protein relation extraction systems become more significant.

Previous researches have illustrated the effectiveness of the shortest dependency path (sdp) between entities for relation extraction in many fields [4–7]. For example, in PPI task, [8] proposed an edit-distance kernel based on sdp and classified the relations by SVM. Reference [9] has made a detailed investigation into the relevant work of relation extraction and elaborated the important role of sdp in relation extraction. However, how to preprocess the sdp (e.g., using a variety of kernels) and how to combine different features (e.g., part-of-speech, *n*-grams, and parser tree) still are open problems. In this work, the proposed approach takes raw sdp as the only input, and it can learn features automatically. And thus, different from previous researches, manual feature selection and feature combination are not necessary in our approach.

Many efforts have been done on PPI task, especially the kernel based methods. Most of these methods take the PPI task as a binary classification problem by determining whether there is an interaction between the two entities. The kernels include bag-of-words kernel [10], all-path kernel [11], subset-tree kernel [12], edit-distance kernel [8], and graph kernel [13], and they have shown effectiveness in PPI task. Considering that single kernel partly calculates the similarity of two instances, hybrid kernel [14–17] has been proposed and demonstrated much better performance than single kernel. Kernel methods are effective, because they integrate a large amount of manually selected features. The problem of existing kernel based method is how to combine different features; in most cases, sophisticated design is required.

Deep learning methods have achieved remarkable results in computer vision [18] and speech recognition [19], and due

to much of the effective work involved in neural network language models (NNLM) [20, 21], recently, some work has focused on neural network especially CNN for natural language processing (NLP) problem. Using CNN to extract features for NLP was previously researched by the authors in [22]; they considered the tasks including part-of-speech (POS), chunking, name entity recognition (NER), and semantic role labeling (SRL) as sequential labeling problems. In recent years, researches have proposed the use of CNN to extract features for relation extraction. Reference [23] combined the word representation, lexical level features, and word features and used the CNN model to learn the sentence-level features; the features were then concatenated into a vector and fed to a Softmax layer to classify the relationship. Reference [24] shared a similar idea to [23]; the authors proposed a new logistic loss function and a pairwise method to train their CNN model.

However, the CNN based methods described by [23, 24] usually take whole sentence or the context between two target entities as input. The problem of these methods is that such representations fail to describe the relationships of two target entities far in sentence distance, and the irrelevant information may also be considered due to the long distance. Considering the described problems and the complexity of PPI task, in this work, we use dependency parsing to analyze the sentence for generating the sdp at first to capture semantical and syntactical features and then send sdp to sdpCNN for classification.

Comparing with the prior work, the contributions of our work can be concluded as follows:

(1) We propose a new model (sdpCNN) to tackle PPI task and show that sdpCNN model built on word embedding is effective in extracting protein-protein relations.

(2) We demonstrate that sdpCNN with pretrained word embedding performs much better than randomly generated word embedding and state-of-the-art kernel based methods. It could be concluded that the well pretrained word embedding is important in PPI task.

(3) The proposed model is able to extract key features automatically such that the manual feature selection procedure can be avoided.

2. Material and Methods

In this section, we firstly introduce word embedding, and then we describe the proposed sdpCNN model in detail. The proposed model consists of three parts: the sdp extraction, sdpCNN based feature extraction, and multilayer perceptron (MLP) based classification.

2.1. Introduction for Word Embedding. Word embedding is a feature learning technique in NLP where words or phrases from the vocabulary are mapped to vectors of real numbers in a low-dimensional space relative to the vocabulary size. Many methods have been proposed to train the word embedding, but most of the methods are based on the distributional hypothesis: words that occur in similar contexts tend to have similar meanings. Given this hypothesis, the trained word embeddings would be close to each other in vector space when the words contain similar meanings (Figure 1 shows visualization of word embedding by t-SNE [25]).

In this work, we use public available pretrained word embedding (300-dimension), trained on 100 billion words from Google News by word2vec [21] (https://code.google.com/archive/p/word2vec/), to build the proposed sdpCNN model.

Compared with traditional *"one-hot"* representation, pretrained word embedding brings about three advantages. (1) It could capture semantic information and weaken word gap problem; for example, in Figure 1, interaction verbs (interaction verbs usually indicate the relation among entities and thus they are important in PPI task) *"affects"* and *"enhance"* are clustered together; however, in traditional *"one-hot"* representation, the verbs *"affects"* and *"enhance"* are completely different. (2) Data sparseness problem could be avoided since all words are mapped into low-dimensional vectors. (3) Pretrained word embedding is trained on large unlabeled corpora, and thus it could enlarge the coverage of vocabulary and decrease the number of unknown words.

2.2. Shortest Dependency Path (sdp) Extraction. Semantic dependency parsing had been frequently used to dissect sentence and to capture word semantic information close in context but far in sentence distance. To extract the relationship between two entities, the most direct approach is to use sdp. The motivation of using sdp is based on the observation that the sdp between entities usually contains necessary information to identify their relationship [9]. For example, in Figure 2, the word *"affects"* in sdp provides useful information for classifying two target proteins, and the dependency relationship such as "*nsubj*" (the dependency relation "*nsubj*" represents "*nominal subject*," and the governor of this relation is always a verb, because interaction verbs are crucial in PPI task; thus, this dependency relation is important in PPI task; more detailed descriptions for relation "*nsubj*" can be found in [26]) between words "*profilin*" and "*affects*" also adds supplemental information for classification.

To reduce the sparseness and ensure the generalization of features, we replace two target proteins with special symbols "*Protein1*" and "*Protein2*," respectively, and thus we can get a sdp *"Protein1-nsubj-affects-dobj-properties-prep-of-Protein2"* from Figure 2.

2.3. sdpCNN Model for Feature Extraction. Figure 3 shows the architecture of the proposed sdpCNN model. In the first step, the model transforms a sdp into a matrix representation by looking up pretrained word embedding; and then, a convolution layer is applied to this matrix to automatically extract the features. The following max-pooling operation generates the most useful local features. At last, the extracted features are fed to a multilayer perceptron (MLP) with a hidden layer and a Softmax classifier.

For notation, we use $\mathbf{D} \in R^{|V| \times d}$ to represent pretrained word embedding, where V is the vocabulary of corpora

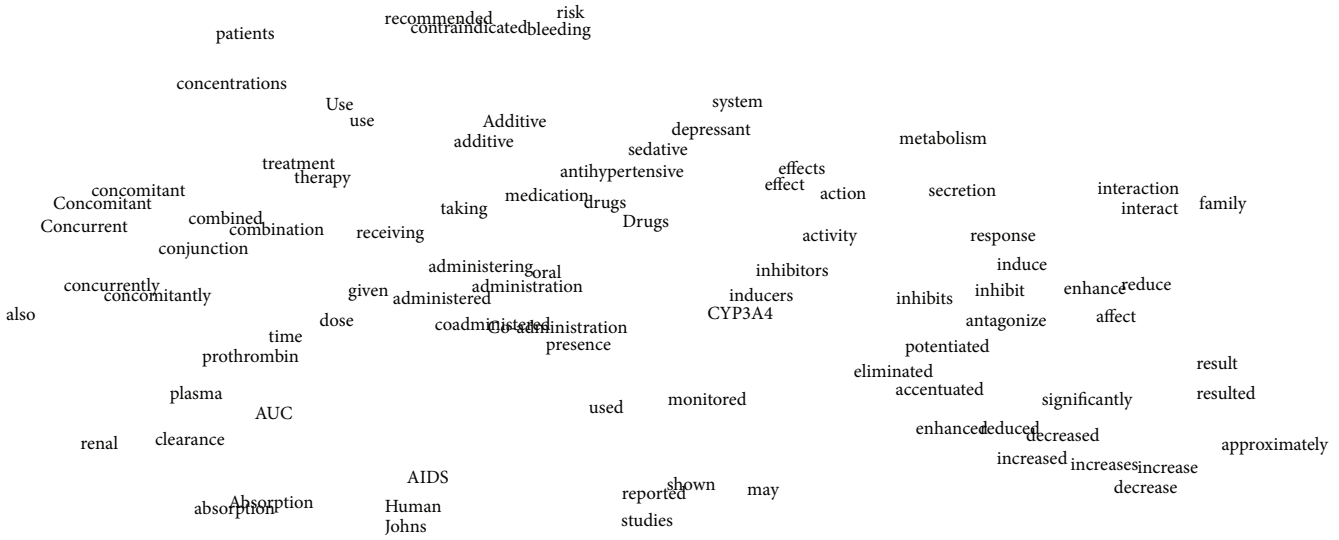


Figure 1: Visualization of word embedding by t-SNE. The words are highly frequent in PPI task. The original word embedding for each word is a 300-dimension vector; all of these words are reduced to 2 dimensions by t-SNE.

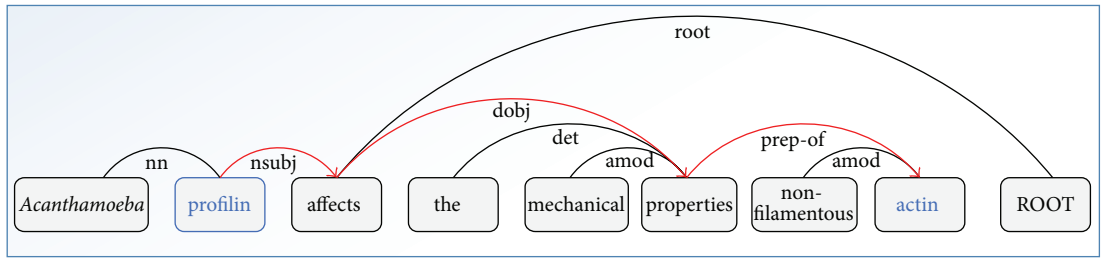


Figure 2: The dependency parsing result of sentence "*Acanthamoeba profilin affects the mechanical properties of non-filamentous actin.*" The words in blue are the two target proteins, and the sdp between the proteins is represented by the red arrows. Tags such as "*nsubj*" and "*dobj*" are the dependency relations between two words.

and d is the dimension of word embedding. Suppose $\mathbf{x} = \{x_1, x_2, x_3, \ldots, x_N\}$ is an input sdp with length N (we fix the length of input path as N by truncating or padding special symbol *"PADDING"*). When we assign each word in sdp $\mathbf{x}$ with a corresponding row vector from $\mathbf{D}$, we would get a matrix representation $\mathbf{P} \in R^{N \times d}$ for input sdp (yellow part in Figure 3).

The convolutional operation would be considered to apply filter $\mathbf{W} \in R^{h \times d}$ to the h-word window in input sdp $\mathbf{x}$. An h-word window in input sdp can be represented as $\mathbf{P}_{i,i+h-1} \in R^{h \times d}$ (yellow part surrounded with red rectangle in Figure 3) by connecting row i to $i + h - 1$ in $\mathbf{P}$. A feature c_i can be generated by

$$c_i = f\left(W \odot \mathbf{P}_{i,i+h-1} + b_1\right), \tag{1}$$

where f is an activation function such as hyperbolic tangent (tanh), b_1 is the bias term, and $\odot$ is element-wise multiplication. By applying filter to each word window of the input sdp, the model will produce a new feature which we call feature map $\mathbf{c}$ in

$$\mathbf{c} = \left[c_1, c_2, \ldots, c_{N-h+1}\right]. \tag{2}$$

Max-pooling operation (see (3)) takes the maximum value over all the word windows in feature map $\mathbf{c}$ which brings about two advantages: (1) it could extract the most important local features and (2) it reduces the computational complexity by reducing the feature dimension. Hence,

$$c^* = \max\left(\mathbf{c}\right). \tag{3}$$

As each filter produces a feature c^*, multiple filters will generate multiple features. Suppose M is the number of the filters; the model would get fixed-size distributed features $\mathbf{r} = [c_1^*, c_2^*, \ldots, c_M^*]$, where c_i^* is the ith feature generated by ith filter.

2.4. MLP for Classification. A MLP model is employed to calculate the probability of each class. Given the distributed representation $\mathbf{r}$, the full-connection weight matrix $\mathbf{W}_2 \in R^{H \times M}$, the number of hidden layers H, and the bias term b_2, the output of full-connection layer $\mathbf{O} \in R^{H \times 1}$ is calculated by

$$\mathbf{O} = f\left(\mathbf{W}_2 \mathbf{r} + b_2\right). \tag{4}$$

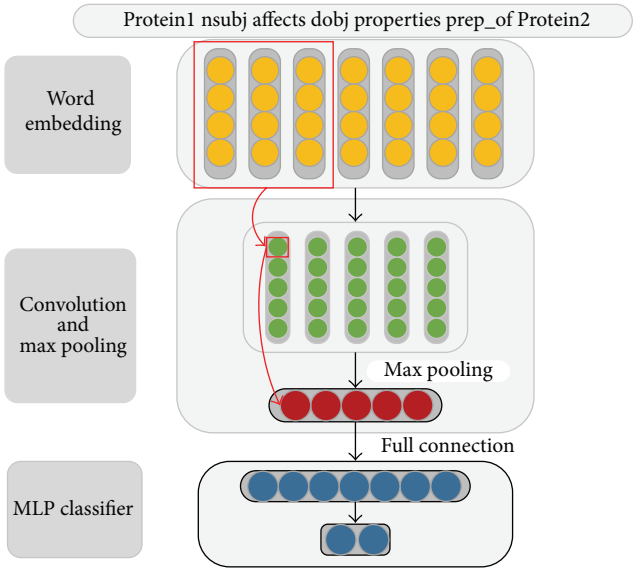


Figure 3: The framework of sdpCNN model with 3-word window. In this example, the input sdp has 7 words (dependency relations such as "*nsubj*" and "*dobj*" are also considered as words), each word embedding is 4 dimensions, and 5 filters are used. The yellow part is the matrix representation for an input sdp; each column in the green part represents the feature map generated by a filter through (1) and (2); and the red part represents the max-pooling results by taking the maximum value over each column in the green part by (3). The arrows in red show the process of generating a feature map c^*. The blue part is a MLP classifier with a full-connection layer and a Softmax layer.

Before applying Softmax layer for classification, the original feature space is transformed into confidence space. The input for Softmax layer $\mathbf{I} \in R^{C\times 1}$ is described by

$$\mathbf{I} = \mathbf{W}_3\mathbf{O}, \tag{5}$$

where $\mathbf{W}_3 \in R^{C\times H}$ is a transformation matrix and C is the number of classes. This task is binary classification, so C is 2.

Each value in $\mathbf{I}$ represents the confidence of the current sample that belongs to each class. A Softmax layer normalizes the confidence to $[0, 1]$. Given $\mathbf{I} = [i_1, i_2, \ldots, i_C]$, the output of Softmax layer $\mathbf{S} = [s_1, s_2, \ldots, s_C]$. The Softmax operation can be calculated by (6). Both s_j and $p(j \mid \mathbf{x})$ represent probability of sdp $\mathbf{x}$ that belongs to class j. Hence,

$$s_j = p\left(j \mid \mathbf{x}\right) = \frac{e^{i_j}}{\sum_{k=1}^{C} e^{i_k}}. \tag{6}$$

2.5. Training Procedure. There are several parameters that need to be updated during the training: the multifilter $\mathbf{W}$, the full-connection weight $\mathbf{W}_2$, the transformation matrix $\mathbf{W}_3$, and the bias terms b_1 and b_2. All of the parameters are represented by $\theta = (\mathbf{W}, \mathbf{W}_2, \mathbf{W}_3, b_1, b_2)$. We apply Negative Log-Likelihood (NLL) in (7) ($\mathbf{y}_i \in \{0, 1\}$ is annotated label for the input sdp $\mathbf{x}_i$) as loss function. In order to minimize the loss function, we use gradient descent (GD) based method to learn the network parameters. For each input pair $(\mathbf{x}_i, \mathbf{y}_i)$, we calculate the gradient (using the chain rules) of each parameter relative to loss and update each parameter with learning rate λ by (8). It is notable that fixed learning rate λ would lead to unstable loss in training. In this work, we use an improved GD based algorithm, Adadelta [27], to update the parameters in each training step. Adadelta is able to dynamically adjust the learning rate. Hence,

$$\text{loss} = -\log p\left(\mathbf{y}_i \mid \mathbf{x}_i\right), \tag{7}$$

$$\theta = \theta - \lambda\frac{\partial \text{loss}}{\theta}. \tag{8}$$

Table 1: Data statistics for Aimed and BioInfer datasets.

Datasets	Positive	Negative
BioInfer	2512	7010
Aimed	995	4812

Table 2: Hyperparameter settings for Aimed and BioInfer.

Datasets	N	d	h	M	H
BioInfer	30	300	3	100	500
Aimed	20	300	3	100	500

3. Results

3.1. Experimental Setup

3.1.1. Datasets. Two standard datasets (both datasets are available at http://corpora.informatik.hu-berlin.de/), Aimed and BioInfer [28], are used to evaluate our model. Aimed was manually tagged by [9] which included about 200 medical abstracts with around 1900 sentences and was considered as a standard dataset for PPI task. BioInfer was developed by Turku BioNLP group (see details at http://bionlp.utu.fi/) which contained about 1100 sentences. If there is an interaction between the two entities, we consider this instance as a positive one; otherwise, we consider it as a negative one (in Table 1). Text preprocessing includes sentence splitting, word segmentation, and dependency parsing (Stanford parser was utilized).

3.1.2. Word Embedding Initialization. In experiments, we compare the performances of pretrained embedding with randomly initialized word embedding. When the words that appeared in the datasets are not included in the pretrained word embedding, we follow [29] and initialize word embedding by randomly sampling from $[-a, a]$, where a is the variance of pretrained word embedding trained by word2vec. For random part, all of the words are initialized by sampling from $[-a, a]$.

3.1.3. Model Hyperparameters Settings. We experimentally choose the hyperparameters for the model on BioInfer and Aimed datasets shown in Table 2. The Discussion gives details on parameter selection as well as the impact of the parameters.

Table 3: The comparison with other kernel based methods on PPI task. *Random sdpCNN model*: sdpCNN model with randomly initialized word embedding. *Pretrained sdpCNN model*: sdpCNN model built on pretrained word embedding.

Method	BioInfer			Aimed		
	P	*R*	*F*	*P*	*R*	*F*
Random sdpCNN model (baseline)	69.6	77.8	73.4	54.5	75.2	62.7
Pretrained sdpCNN model	**73.4**	77.0	**75.2**	**64.8**	67.8	**66.0**
sdp based methods						
Walk-weighted subsequence kernel	61.8	54.2	57.6	61.4	53.3	56.6
Graph kernel	—	—	—	52.9	61.8	56.4
SDP-CPT	—	—	62.4	—	—	58.1
Tree kernel	—	—	62.8	—	—	51.4
Edit-distance kernel	—	—	—	58.4	61.2	59.6
Hybrid kernel based methods						
Hybrid kernel	65.7	71.1	68.1	55.0	68.8	60.8
Multiple features and parser	—	—	67.6	—	—	64.2
Multiple kernel	57.0	77.3	65.8	57.7	71.1	64.4

3.1.4. Evaluation Metrics. We use precision (P), recall (R), and F-score (F) to evaluate the performances of our sdpCNN model. F is the harmonic mean of recall and precision which is defined by (9). 10-cross-validation (10-fold CV) method is used to calculate the average F-scores. Hence,

$$F = \frac{2 \times P \times R}{P + R}. \tag{9}$$

3.2. Performance Comparison. We evaluate our system and compare the performance with state-of-the-art kernel based methods. We start from a baseline model with randomly initialized word embedding, and then we evaluate our model with the pretrained word embedding. Table 3 shows the comparison results in detail.

We firstly compare the performance with other sdp based methods, and then we compare the results with hybrid kernels based methods. The descriptions for methods in Table 3 are as follows:

Walk-Weighted Subsequence Kernel [30]. Generating sdp at first and then integrating the proposed e-walk and v-walk kernels for classification.

Graph Kernel [13]. Encoding the dependency parser results into a graph, proposing an all-path graph kernel by leveraging sdp; at last, least squares support vector machine is used for classification.

SDP-CPT [4]. Using both sdp and directed constituent parser tree for classification.

Tree Kernel [6]. On the bias of SDP-CPT, considering the modal verb phrases and appositive dependency features.

Edit-Distance Kernel [8]. A semisupervised machine learning approach (TSVM) with edit-distance kernel based on sdp.

Hybrid Kernel [14]. A combination of bag-of-words (BOW) kernel, subset-tree (ST) kernel, and graph kernel.

Multiple Features and Parser [31]. A combination of rich features including bag-of-words features, sdp features, and graph features.

Multiple Kernel [32]. A weighted multiple kernel by combining parser tree, graph features, POS, and sdp.

As we can see, kernel methods listed in Table 3 usually require sophisticated design and complex feature combination, and feature engineering still accounts for a large proportion of these systems. In this work, we avoid manual features selection and features combination by using CNN. In addition, the features used in these kernel based methods are all discrete; therefore, the *"word gap"* problem is inevitable, while, by leveraging word embedding and CNN, we can train our model in continuous space and avoid hard assignment.

The main differences of the sdp based methods listed in Table 3 are how sdps were used and how similarity functions were calculated. For example, the most direct way is to encode sdp into *"one-hot"* representation and use SVM for classification [4, 6]. Another way is by using edit-distance kernel [8] to calculate the similarity of two sdps through Levenshtein distance. Compared with these sdp based methods in Table 3, even the baseline model achieved competitive results. Furthermore, pretrained sdpCNN model improved the F-scores by 12.4 and 6.4 compared with tree kernel [6] and edit-distance kernel [8] on BioInfer and Aimed datasets, respectively.

It has been verified that a combination of multiple kernels could improve the effectiveness of kernel based PPI extraction methods. Kernels such as tree kernel, graph kernel, and bag-of-words kernel are commonly used in hybrid kernel based methods. Compared with the methods listed in Table 3, the baseline model alone yielded competitive results and improved the F-scores by 5.3 on BioInfer dataset when compared with [14]. By integrating pretrained word embedding, our pretrained sdpCNN model exceeded 7.1 and 1.6 compared with [14, 32] on BioInfer and Aimed datasets in Table 3. The experimental results showed that, with the appropriate expression (the sdp in this work) of the relationship,

the sdpCNN model built on word embedding can get much better results than the combination of a variety of features (or kernels).

For better understanding extracted features by sdpCNN, Figure 3 illustrates the way of generating a feature map c^* in sdpCNN model. By following the negative direction of the red arrows in Figure 3, we can find which word window contributes most to the final classifier. Considering the example in Figure 3, the 3-word window ("*Proteins nsubj affects*") circled with a red rectangle is key item. We define the word in the middle of the key word window as key-word, and thus the word "*nsubj*" in the middle of the 3-word window "*Proteins nsubj affects*" in Figure 3 is key-word. Each filter produces a key-word; consequently, M filters will generate M key-words. In our experiments, we noticed that interaction verbs such as "*inhibits*," "*cause*," and "*bind*" were often chosen as key-words by sdpCNN model. Generally, the construction of an interaction verbs dictionary manually requires a great deal of time and effort, but our model can extract these verbs automatically.

Moreover, the experimental results also showed that the proposed method achieved considerably higher precision (73.4 on BioInfer dataset and 64.8 on Aimed dataset) than the existing approaches.

3.3. Evaluation on Different Scales of Training Data. In order to investigate the effect of different scales of training data, we split the original datasets by different ratios. Figure 4 shows the changes of performance on different scales of training data. As we can see, the performance varied significantly depending on the size of training and test corpus, and F-scores changed from 75.1 to 48.2 on BioInfer dataset and 71.1 to 36.2 on Aimed dataset when proportion of test data ranged from 0.1 to 0.9; too few training data would have the risk of loss of data information; as a result, the trained sdpCNN model cannot well generalize the original data which would lead to poor performance.

3.4. Discussion. In this section, we firstly investigate the impact of hyperparameters and provide general parameters settings for sdpCNN. After that, we compare the performances among the four proposed methods in Table 5. At last, we manually analyze the errors of sdpCNN alone with the possible solutions to errors.

3.4.1. The Influence of Different Hyperparameters Settings. Consider the following:

(1) Window size h: a 3-word window is commonly used in many related works [22–24]; we tested a 2-word window on both Aimed and BioInfer datasets. On Aimed dataset, the results remained essentially unchanged; however, when tested on BioInfer dataset, F-scores reduced by 5. We also tested a 4-word window, while, in this experiment, performances are markedly inferior on both datasets, which means a 4-word window is too long to capture the structure information.

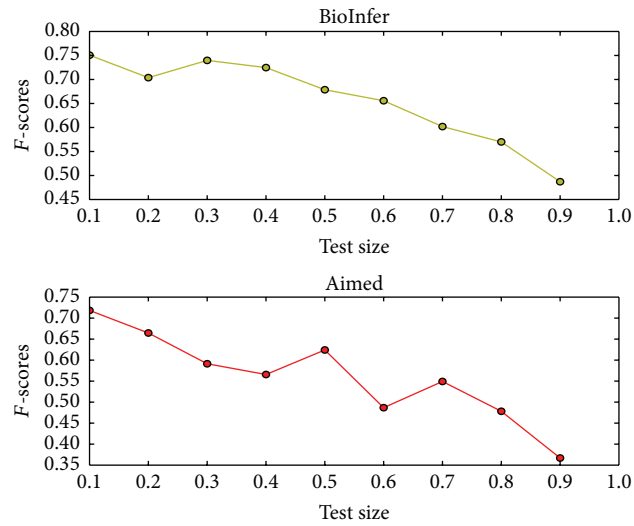


FIGURE 4: Changes of performance on different scales of training data. The x-axis represents the proportion of test data, and the y-axis corresponds to F-scores. The pretrained sdpCNN model is used in this experiment.

(2) The length of fixed-size sdp N: the lengths of most paths (more than 95%) in Aimed dataset are less than 20, while, in BioInfer dataset, most of the path lengths (more than 95%) are less than 30. And thus we set N with 20 and 30 on Aimed and BioInfer datasets, respectively.

(3) The filters size M: due to the limited size of corpora, when the filters size is too big, the model is prone to overfitting; we heuristically choose M as 100 in our experiments.

(4) The number of full-connection layer units H: based on the idea of [33], the appropriate increment of full-connection layer units could improve the performance. But too many units also suffer from overfitting, so we set H with 500 in this experiment.

3.4.2. Comparisons among the Four Proposed Models

Random sdpCNN Model versus Pretrained sdpCNN Model. From Table 3, we can find that the pretrained sdpCNN model performed much better than random sdpCNN model and improved the F-scores by 1.8 and 3.3 on BioInfer and Aimed datasets, respectively. Intuitively, the pretrained word embedding could capture the semantic information of words, which means words with similar semantics are clustered together in the vector space (Figure 1). Table 4 shows the examples of neighboring words of target words based on cosine similarity; we can see that word, for example, "*affect*," shares a similar meaning with words "*impacting*," "*jeopardize*," and so forth. However, when we randomly allocated the word embedding, semantic information among words would be discarded; as a result, random sdpCNN model might correctly classify the sentence "*Protein1 affects Protein2*" but fails on the sentence "*Protein1 impacts Protein2*" although both sentences indicate interactions. Random sdpCNN model is somewhat similar

Table 4: The top 5 neighboring words of target words based on cosine similarity (the variants of the target words, such as "*induced*," "*inducing*," and "*depended*," are not included in this table).

Target words	1	2	3	4	5
Induce	*elicit*	*suppress*	*provoke*	*potentiate*	*engender*
Affect	*impacting*	*jeopardize*	*hinder*	*impair*	*imperil*
Bind	*vise*	*attach*	*untie*	*glue*	*entangle*
Depend	*rely*	*hinge*	*predicated*	*affect*	*dictate*
Prevent	*deter*	*avoid*	*discourage*	*forestall*	*avert*

Table 5: The results of the four proposed models on PPI task. *Combined model*: using both randomly initialized and pretrained word embedding as inputs and concatenating the outputs of max-pooling layer as features for MLP. *Random (update) sdpCNN model*: initializing word embedding randomly and updating word embedding during training.

Method	BioInfer	Aimed
	F	*F*
Combined model	**75.3**	**66.6**
Pretrained sdpCNN model	75.2	66.0
Random sdpCNN model	73.4	62.7
Random (update) sdpCNN model	74.1	63.3

to the "*one-hot*" model; the trained random sdpCNN model can be well applied to the test data only when train and test instances contain common words which means this model is too dependent on cooccurrence of words and lacks good generalization ability. However, as a benefit from sdp and CNN, the structure information could be well preserved; therefore, random sdpCNN model still achieved comparable results. More specifically, it could be concluded that pretrained sdpCNN model can capture both semantic information and structural information, while the random sdpCNN model could only keep structural information. Both semantic information and structural information play important roles in PPI task.

Random sdpCNN Model versus Random (Update) sdpCNN Model. In random (update) sdpCNN model, we considered word embedding as hyperparameters and updated it in the training procedure. The experimental results showed that the random (update) sdpCNN model had a slight improvement (0.7 and 0.6 *F*-scores improvements on BioInfer and Aimed datasets, resp.) compared with the random sdpCNN model. Intuitively, the random (update) sdpCNN model can adapt to the specific task by fine-tuning word embedding which means word embedding can learn task specific patterns. However, when compared with pretrained sdpCNN model, the model's performances reduced by 1.1 and 2.7 on BioInfer and Aimed datasets. The good performance on pretrained sdpCNN model is understandable due to the fact that the pretrained word embedding is trained on large corpora which ensures that the pretrained sdpCNN model could obtain abundant semantic information. Moreover, because the pretrained sdpCNN model does not need to update word embedding, the training time consumption could be reduced.

Combined Model versus Pretrained sdpCNN Model. To better learn the representation of the raw sdp input, we also proposed a model that combined the pretrained and random word embedding (see details in Table 5). The combined model improved the *F*-scores by 0.6 on Aimed corpus and kept the performance on BioInfer corpus when compared with pretrained sdpCNN model. However, it is also notable that the combined model would take more than two times the cost on training time. There is always a trade-off between time and performance.

Among these four models, pretrained sdpCNN model is more time-saving (relative to combined model and random (update) sdpCNN model), robust (relative to random (update) sdpCNN model), and effective (relative to random (update) sdpCNN model and random sdpCNN model). In conclusion, a CNN model built on high-quality pretrained word embedding could be considered as an effective alternative in PPI task.

3.4.3. Errors Analysis. Confined to the complexity and diversity of the biomedical expressions, extracting relations from biological articles remains a big challenge. In this subsection, we carefully analyze the errors of sdpCNN and list the three typical errors as follows:

(1) When an input sentence is too long, the Stanford dependency analysis tool is prone to errors, and because our model is built on sdp the propagation of errors would lead to poor performance of sdpCNN.

(2) When irrelevant interaction verbs are included in sdp, as mentioned before, interaction verbs strongly suggest interactions; as a result, the model would make a mistake.

(3) Randomly initialized word embedding would also hurt the system's performance. In our system, the dependency relations such as "*nsubj*" and "*prep-of*" are all considered as input words, and such words are not likely to be included in pretrained word embedding, and thus these words are randomly assigned with vectors. As a result, *"nsubj"* and *"prep-of"* might be far from each other in vector space.

For example, for two input paths *"Protein1-nsubj-bind-nsubj-Protein2"* and *"Protein1-nsubj-bind-prep-of-Protein2,"* both paths indicate interactions; however, the sdpCNN model could only distinguish the first one.

The possible solutions for the mentioned errors are described as follows: the first error could be weakened by integrating the context between two target entities, because the context could provide supplementary information when standard tools fail to capture dependency relations among words. As for the second error, a possible solution is to introduce position information, because, in most of the time, the relevant interaction verbs locate in the middle of two target entities. For randomly initialized word embedding problem, we might take word embedding as hyperparameter and update it during the training. Meanwhile, word embedding used in this work is trained on large unlabeled Google News; it would be better to train word embedding on large biological articles to enrich semantic information.

4. Conclusion

In this paper, we have described a sdpCNN model built on word embedding for PPI task. Experiments demonstrated that our method outperformed the state-of-the-art kernel based methods. The main contribution of the proposed method is the integration of word embedding, sdp, and CNN. Word embedding is able to capture semantic information and effectively weaken word gap problem. By applying sdp and CNN, the proposed model could make full use of structure information and avoid manual feature selection. Our experimental results also indicated that (1) the raw sdp input is crucial to describe protein-protein relationship in PPI task; (2) the CNN model is useful to capture the local features and structure information; (3) high-quality pretrained word embedding is important in PPI task. Through error analysis, we notice that there still is room for improvement. In our future work, we would like to train our own word embedding and design our PPI system by making full use of context information, position information, and sdp.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

This research has been partially supported by the National Natural Science Foundation of China under Grant no. 61472117 and the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry.

References

[1] G. D. Bader, D. Betel, and C. W. V. Hogue, "BIND: the biomolecular interaction network database," *Nucleic Acids Research*, vol. 31, no. 1, pp. 248–250, 2003.

[2] A. Zanzoni, L. Montecchi-Palazzi, M. Quondam, G. Ausiello, M. Helmer-Citterich, and G. Cesareni, "MINT: a molecular interaction database," *FEBS Letters*, vol. 513, no. 1, pp. 135–140, 2002.

[3] S. Kerrien, B. Aranda, L. Breuza, and A. Bridge, "The intact molecular interaction database in 2012," *Nucleic Acids Research*, vol. 38, article D525, 2012.

[4] L. Qian and G. Zhou, "Tree kernel-based protein-protein interaction extraction from biomedical literature," *Journal of Biomedical Informatics*, vol. 45, no. 3, pp. 535–543, 2012.

[5] C. Quan, M. Wang, and F. Ren, "An unsupervised text mining method for relation extraction from biomedical literature," *PLoS ONE*, vol. 9, no. 7, Article ID e102039, 2014.

[6] C. Ma, Y. Zhang, and M. Zhang, "Tree Kernel-based protein-protein interaction extraction considering both modal verb phrases and appositive dependency features," in *Proceedings of the 24th International Conference on World Wide Web Companion*, pp. 655–660, International World Wide Web Conferences Steering Committee, Florence, Italy, May 2015.

[7] R. C. Bunescu, R. Ge, R. J. Kate et al., "Comparative experiments on learning information extractors for proteins and their interactions," *Artificial Intelligence in Medicine*, vol. 33, no. 2, pp. 139–155, 2005.

[8] G. Erkan, A. Özgür, and D. R. Radev, "Semi-supervised classification for extracting protein interaction sentences using dependency parsing," in *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL '07)*, vol. 7, pp. 228–237, Prague, Czech Republic, June 2007.

[9] R. C. Bunescu and R. J. Mooney, "A shortest path dependency kernel for relation extraction," in *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 724–731, Association for Computational Linguistics, Vancouver, Canada, October 2005.

[10] R. Sætrea, K. Sagae, and J. Tsujii, "Syntactic features for protein-protein interaction extraction," in *Proceedings of the 2nd International Symposium on Languages in Biology and Medicine (LBM '07)*, Singapore, January 2008.

[11] A. Airola, S. Pyysalo, J. Björne, T. Pahikkala, F. Ginter, and T. Salakoski, "A graph kernel for protein-protein interaction extraction," in *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing (BioNLP '08)*, pp. 1–9, Association for Computational Linguistics, 2008.

[12] A. Moschitti, "Making tree kernels practical for natural language learning," in *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL '06)*, pp. 113–120, 2006.

[13] A. Airola, S. Pyysalo, J. Björne, T. Pahikkala, F. Ginter, and T. Salakoski, "All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning," *BMC Bioinformatics*, vol. 9, article 52, 2008.

[14] M. Miwa, R. Sætre, Y. Miyao, and J. Tsujii, "Protein-protein interaction extraction by leveraging multiple kernels and parsers," *International Journal of Medical Informatics*, vol. 78, no. 12, pp. e39–e46, 2009.

[15] Md. F. M. Chowdhury and A. Lavelli, "Combining tree structures, flat features and patterns for biomedical relation extraction," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL '12)*, pp. 420–429, Association for Computational Linguistics, 2012.

[16] A. W. Muzaffar, F. Azam, and U. Qamar, "A relation extraction framework for biomedical text using hybrid feature set," *Computational and Mathematical Methods in Medicine*, vol. 2015, Article ID 910423, 12 pages, 2015.

[17] D. Zhou, D. Zhong, and Y. He, "Biomedical relation extraction: from binary to complex," *Computational and Mathematical Methods in Medicine*, vol. 2014, Article ID 298473, 18 pages, 2014.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proceedings of the 26th Annual Conference on Neural Information Processing Systems (NIPS '12)*, pp. 1097–1105, Lake Tahoe, Nev, USA, December 2012.

[19] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '13)*, pp. 6645–6649, IEEE, Vancouver, Canada, May 2013.

[20] Y. Bengio, H. Schwenk, J.-S. Senécal, F. Morin, and J.-L. Gauvain, "Neural probabilistic language models," in *Innovations in Machine Learning*, pp. 137–186, Springer, Berlin, Germany, 2006.

[21] T. Mikolov, I. Sutskever, K. Chen, and S. Greg, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, pp. 3111–3119, Curran Associates, 2013.

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

[23] D. Zeng, K. Liu, S. Lai, and G. Zhou, "Relation classification via convolutional deep neural network," in *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers (COLING '14)*, pp. 2335–2344, Dublin, Ireland, August 2014.

[24] C. N. dos Santos, B. Xiang, and B. Zhou, "Classifying relations by ranking with convolutional neural networks," http://arxiv.org/abs/1504.06580.

[25] L. Van Der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2625, 2008.

[26] K. Fundel, R. Küffner, and R. Zimmer, "RelEx-relation extraction using dependency parse trees," *Bioinformatics*, vol. 23, no. 3, pp. 365–371, 2007.

[27] M. D. Zeiler, "ADADELTA: an adaptive learning rate method," https://arxiv.org/abs/1212.5701.

[28] M. Miwa, R. Sætre, Y. Miyao, and J. Tsujii, "A rich feature vector for protein-protein interaction extraction from multiple corpora," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP '09)*, vol. 1, pp. 121–130, Association for Computational Linguistics, 2009.

[29] Y. Kim, "Convolutional neural networks for sentence classification," 2014, https://arxiv.org/abs/1408.5882.

[30] S. Kim, J. Yoon, J. Yang, and S. Park, "Walk-weighted subsequence kernels for protein-protein interaction extraction," *BMC Bioinformatics*, vol. 11, article 107, 2010.

[31] M. Miwa, R. Sætre, Y. Miyao, and J. Tsujii, "A rich feature vector for protein-protein interaction extraction from multiple corpora," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, vol. 1, pp. 121–130, Association for Computational Linguistics, August 2009.

[32] Z. Yang, N. Tang, X. Zhang, H. Lin, Y. Li, and Z. Yang, "Multiple kernel learning in protein-protein interaction extraction from biomedical literature," *Artificial Intelligence in Medicine*, vol. 51, no. 3, pp. 163–173, 2011.

[33] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2323, 1998.

Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 9140541, 12 pages
http://dx.doi.org/10.1155/2016/9140541

Research Article

Therapeutic Effects of CUR-Activated Human Umbilical Cord Mesenchymal Stem Cells on 1-Methyl-4-phenylpyridine-Induced Parkinson's Disease Cell Model

Li Jinfeng,[1] Wang Yunliang,[2] Liu Xinshan,[3] Wang Yutong,[4] Wang Shanshan,[2] Xue Peng,[5] Yang Xiaopeng,[5] Xu Zhixiu,[5] Lu Qingshan,[5] Yin Honglei,[2] Cao Xia,[1] Wang Hongwei,[6] and Cao Bingzhen[1]

[1]*Neurology Department of General Hospital of Jinan Military Region, Jinan, Shandong 250031, China*
[2]*The Neurology Department, The 148th Hospital, Zibo, Shandong 255300, China*
[3]*Sanbo Brain Hospital Capital Medical University, Haidian District, Beijing 100093, China*
[4]*Medical School of Henan University, Zhengzhou, Henan 475000, China*
[5]*Neurology Department, The Second Hospital Affiliated to Zhengzhou University, Zhengzhou, Henan 450014, China*
[6]*Department of Medicine, The University of Chicago, Chicago, IL 60637, USA*

Correspondence should be addressed to Wang Hongwei; hwang1@medicine.bsd.uchicago.edu
and Cao Bingzhen; cbzxia2011@163.com

Received 23 November 2015; Revised 2 March 2016; Accepted 27 March 2016

Academic Editor: Nicola Simola

The purpose of this study is to evaluate the therapeutic effects of human umbilical cord-derived mesenchymal stem cells (hUC-MSC) activated by curcumin (CUR) on PC12 cells induced by 1-methyl-4-phenylpyridinium ion (MPP+), a cell model of Parkinson's disease (PD). The supernatant of hUC-MSC and hUC-MSC activated by 5 μmol/L CUR (hUC-MSC-CUR) were collected in accordance with the same concentration. The cell proliferation and differentiation potential to dopaminergic neuronal cells and antioxidation were observed in PC12 cells after being treated with the above two supernatants and 5 μmol/L CUR. The results showed that the hUC-MSC-CUR could more obviously promote the proliferation and the expression of tyrosine hydroxylase (TH) and microtubule associated protein-2 (MAP2) and significantly decreased the expression of nitric oxide (NO) and inducible nitric oxide synthase (iNOS) in PC12 cells. Furtherly, cytokines detection gave a clue that the expression of IL-6, IL-10, and NGF was significantly higher in the group treated with the hUC-MSC-CUR compared to those of other two groups. Therefore, the hUC-MSC-CUR may be a potential strategy to promote the proliferation and differentiation of PD cell model, therefore providing new insights into a novel therapeutic approach in PD.

1. Introduction

PD is a neurodegenerative disorder being characterized by the progressive loss of dopaminergic neurons of the nigrostriatal pathway, with an accompanying neuroinflammation and Lewy body in the brain [1, 2]. The degeneration of dopaminergic neurons located in the substantia nigra characterizes PD and leads to a decline of dopamine (DA), as well as its biosynthetic enzyme, tyrosine hydroxylase (TH), and its high-affinity cellular transporter (dopamine transporter, DAT) [1, 2]. Current treatment for PD relies on medicine, such as levodopa, that alleviates early symptoms but fails to prevent disease progression [3, 4], and the risk of surgical operation therapy is relatively high [5, 6]. In recent years, cell therapies have gained traction in the treatment of PD with focus on the regeneration of DA producing neurons [7–9]. The mesenchymal stem cells derived from hUC-MSC have priority to repairing PD due to multiple advantages including ethical agreeableness, a less invasive procedure for isolation, low immunogenicity, high proliferation capacity, and multilineage differentiation capability [10–12]. Recent studies have shown that the hUC-MSC have more biological

activity after gene modification or activation, making it more conducive to repair PD. In this study, curcumin (CUR) was chosen to activate the hUC-MSC. CUR is a natural phenolic compound extracted from the plant *Curcuma longa* L. In previous studies, CUR has been shown to have anticancer, antioxidant, and anti-inflammatory effects [13–16]. In 2012, the researchers found that the CUR could bind to α-Synuclein (α-Syn), the main component of LB, so as to prevent its accumulation in the DA neurons [17]. The study also pointed out that CUR cannot pass through the blood-brain barrier (BBB). Data on CUR concentrations in human brains is indeed lacking, only with one exception reporting human serum CUR concentrations in a low micromolar range [18]; all other studies supplementing CUR in human subjects only resulted in serum CUR concentrations in a low nanomolar range or they were not detectable [19]. It is therefore plausible that the CUR levels in human brains are also very low even if CUR can penetrate the blood-brain barrier. Therefore the effect of CUR may be lessened in the brain, while hUC-MSC can pass through the BBB and also has the powerful differentiation potential [20]. Based on the advantages and disadvantages of hUC-MSC and CUR, we conducted the study on the effects of hUC-MSC activated by CUR on the treatment of PD. To our knowledge, there is no related report. Therefore, the purpose of the study is to observe the change of the PC12 PD cells after treatment by the concentrated supernatant from hUC-MSC activated by CUR and investigate its relevant mechanism.

2. Materials and Methods

2.1. Materials. This study was approved by the ethics committee of Jinan Military General Hospital and the 148 Central Hospital of PLA. Umbilical cord tissues were obtained from healthy patients admitted to our hospital, and all patients signed the written informed consent. PC12 cell strains were purchased from Shanghai Institute of Cell Biology, Chinese Academy of Science. F12 medium, CCK-8, CUR, MPP+, and iNOS antibody were purchased from Sigma (USA), while the rest of the antibodies were purchased from R&D (USA). Annexin V/PI Apoptosis Detection Kit was purchased from Shanghai Qcbio Science & Technologies Co., Ltd. DA ELISA Detection Kit was purchased from Cayman Chemical (USA). The nitric oxide (NO) and Griess Detection Kit were purchased from Shanghai Biyuntian Biological Co., Ltd.

2.2. Isolation and Identification of hUC-MSC. hUC-MSCs were isolated from human umbilical cords and cultured as previously described [11, 21].

2.3. Preparation of CUR Stock Solution. CUR powder was dissolved in DMSO to obtain a concentration of 100 μmol/L and then was stored at −20°C protected from light. Different concentrations (0, 1, 2.5, 5, 10, 15, 20, and 25 μmol/L) of CUR were prepared by diluting the stock solution with DMSO.

2.4. Preparation of Conditioned Medium. Firstly, hUC-MSCs were activated by CUR of different concentrations (0, 1, 2.5, 5, 10, 15, 20, and 25 μmol/L, resp.) and the cell proliferation was detected using CCK-8 assay. According to the OD value, we thought that the concentrations of 5 μmol/L are most appropriate and were therefore selected for the following experiments (Table 1).

Table 1: The OD value of hUC-MSC activated by CUR and DMSO.

Groups	Concentration (μmol/L)								
	0	0.1	1	2.5	5	10	15	20	25
CUR	2.89	2.82	2.97	3.32	3.55	3.22	2.48*	2.08**	0.99**
DMSO	2.54	2.78	2.65	2.79	2.98	2.98	2.67	2.89	2.84

*$p < 0.05$, **$p < 0.01$.

5 μmol/L CUR was added to hUC-MSC cell medium for 24 h, washed with sterile PBS 3 times, and applied with serum-free medium for 48 h. The supernatant was drawn and centrifuged at 4°C and 3000 r/min using an ultrafiltration tube for 1.5 h and was repeated 3 times and the ultimate concentration was confirmed to be 10 times the original supernatant. It was then cryopreserved at −80°C. The hUC-MSC cell supernatant was concentrated using the same method. The conditioned medium, by concentrating the CUR-activated hUC-MSC supernatant, was named CM-CUR, and the hUC-MSC supernatant was concentrated to obtain medium with the same concentration and was named as CM-MSC.

2.5. PC12 PD Cell Model. Formulation of MPP+: 2.9714 mg of MPP+ was weighted and dissolved in 1 mL of double-distilled water to obtain 10 mmol/L MPP+ solution, which was filtered and stored at −20°C in dark conditions.

PC12 cells were incubated in 96 well plates at a density of 1×10^5/mL and 100 μL/well for 12 h and then treated with MPP+ of various concentrations for 24 h (final concentrations of 250, 500, and 1000 μmol/L). Cell viability was assessed by adding 10 mL of CCK-8 to the culture and it was incubated for 2 h. Cell viability was measured by spectrometry with 450 nm wavelength. The concentration of MPP+ for desired cell damage was determined for the PD cell model based on cell viability.

2.6. CCK-8 Assay and Flow Cytometry. PC12 cells are divided into 5 groups: control group (without treatment), model group (cells were cultured for 12 h and treated with MPP+), CM-CUR group (cells were cultured for 12 h and treated with MPP+ and 24 h later treated with 10 μL CM-CUR), CM-MSC group (cells were cultured for 12 h and treated with MPP+ and 24 h later treated with 10 μL CM-MSC), and CUR group (cells were cultured for 12 h and treated with MPP+ and 24 h later treated with 5 μmol/L CUR). Cell viability was assessed by CCK-8 assay.

Cell apoptosis and necrosis were detected by flow cytometry (Annexin V/PI double staining). The identifications are as follows: on the scattered plots obtained by bivariate flow cytometry, the lower left quadrant refers to the living cells (FITC−/PI−), while the lower right quadrant refers to the dead cells or apoptotic cells (FITC+/PI−). In this study, due to

Table 2: Primers of target genes for amplified PCR.

Target gene	Oligonucleotide sequence	Product size
Bcl-2	F: TAAGCTGTCACAGAGGGGCT	250 bp
	R: GCGACGAGAGAAGTCATCCC	
Caspase-3	F: GCTTCTTCAGAGGCGACTAC	350 bp
	R: GTGGAAAGTGGAGTCCAGGG	
β-actin	F: TACCAACTGGGACGACATGG	120 bp
	R: CGGTTGGCCTTAGGGTTCAG	

F: forward primer; R: reward primer.

loose adherence of undifferentiated PC12 cells, larger damage from MPP+, and long observation time, the sum of apoptotic and necrotic rates were selected as the outcomes to assess the protective effect of CM-CUR, CM-MSC and CUR against MPP+ induced PC12 cell injury after comprehensive analysis.

2.7. Quantitative Real-Time PCR. Total RNA was isolated and purified using an RNA extraction kit (Tiangen Biochemical Technology Co., Ltd., Beijing, China) according to the manufacturer's instructions. 1 μg total RNA was used for reverse transcription in a final volume of 20 μL. Reverse transcription was performed according to the manufacturer's protocol (DRR047A, Takara, Otsu, Shiga, Japan). Then 2 μL cDNA was used for real-time PCR with the SYBR Premix Ex Taq (DRR041A, Takara, Japan). Quantitative real-time PCR was performed under the following conditions: 95°C for 30 seconds, 95°C for 5 seconds, 60°C for 34 seconds, and 40 cycles. All PCR reactions were performed in triplicate. The specific oligonucleotide primers for rat Bcl-2 (B-cell lymphoma 2), caspase-3, and β-actin are listed in Table 2. The level of expression for the target gene was calculated as the ratio of the copy number of the target gene to that of β-actin.

2.8. Western Blot. PC12 cells were treated with CM-CUR, CM-MSC, and CUR for 96 h, and the cells were collected into lysate A. The cells were rinsed with precooling PBS 3 times, the residual PBS was removed, and the precooling lysate A was added, wherein cells were scraped. Protein was quantified using a BCA-200 protein assay kit. 20 μg of each sample was collected and mixed with loading buffer and DTT in a proportion of 8 : 10 : 2, was boiled for degeneration for 5 min, underwent electrophoresis with 12% SDS-PAGE protein, and was transferred to a membrane. Then, the membrane was sealed with 5% defatted milk and hybridized overnight at 4°C with rabbit anti-TH, DAT, the neural specific marker microtubule associated protein-2 (MAP2), and iNOS (Abcam, Cambridgeshire, UK). The unbinding antibodies were fully washed, applied with anti-rabbit horse radish peroxidase (Abcam, Cambridgeshire, UK), and incubated at room temperature for 1 h, followed by color rendering with enhanced chemical fluorescein and image analysis using Quantity One software (BIO-RAD).

2.9. ELLSA. The above cell supernatants were mixed with coating buffer (0.5 mol/L $NaHCO_3$ buffer, pH 9.6) in a proportion of 1 : 1 and 100 μL (in each well) was added to an enzyme label plate coated with monoclonal antibodies containing dopamine (DA), interleukin-10 (IL-10), interleukin-6 (IL-6), interleukin-1β (IL-1β), tumor necrosis factor-alpha (TNF-α), interferon-γ (IFN-γ), and nerve growth factor (NGF) and kept overnight at 4°C according to the manufacturer's instructions (Cayman Chemical, USA).

2.10. Immunocytochemistry. The above 5 groups of cells were cultured for 96 h and fixed with 4% polysorbate for 10 min and then rinsed again with PBS for 5 min. Next, they were perforated with 0.5% Triton for 15 min and then sealed with 1% BSA for 30 min and treated with rabbit MAP-2 diluted with 1% BSA (Abcam, Cambridgeshire, UK). They were incubated overnight at 4°C, rinsed with PBS twice, each for 5 min, and then treated with the second antibody (37°C for 1 h, Abcam, Cambridgeshire, UK), followed by DAB color rendering and mounting.

2.11. Griess Method. Griess Reagents I and II were collected and placed at room temperature for 0.5 h (Shanghai Biyuntian Biological Co., Ltd., China). Then, 1 M of standard $NaNO_2$ was diluted using medium for culturing PC12 to obtain concentrations of 0, 1, 2, 5, 10, 20, 40, 60, and 100 μM, respectively. The standards and samples were placed into a 96-well plate, with 50 μL/well, and 50 μL Griess Reagent I and 50 μL Griess Reagent II were added to each well in sequence, followed by absorbance detection at 540 nm.

2.12. Statistical Analysis. Statistical analyses were performed using SPSS10. 0 software. Data presented as means ± SEM were subjected to one- or two-way ANOVA, followed by either Newman-Keuls or Bonferroni's multiple-comparisons test (as a post hoc test). $p < 0.05$ was considered to indicate statistical significance. The results of the immunocytochemistry and Western blot were analyzed by Image-Pro Plus 5.0 image analyzer (Media Cybernetics, USA). The integrated optical density (IOD) and gray values were assayed by statistical analysis.

3. Results

3.1. Mesenchymal Stem Cells Have Been Successfully Isolated from the Umbilical Cord Using Tissue Blocking Method Conveniently and Economically. We have previously shown that hUC-MSC can be successfully isolated from the human umbilical cord. They were positive for mesenchymal stem cell marker CD105 (90.03%) and integrin markers CD29 (94.20%) and CD44 (95.63%), but negative for endothelial cell marker CD31 (6.89%) and hematopoietic cell marker CD45 (5.07%), or lymphocyte surface markers HLA-DR (0.33%). After a stringent quality control procedure, the hUC-MSCs were clean and free of pollution and can be used in the subsequent experiment [11, 21].

3.2. CM-CUR Tends to Present the Strongest Effects on Promoting Proliferation and Inhibiting Apoptosis of PD Model Cells. According to the results of CCK-8 assay, the PC12

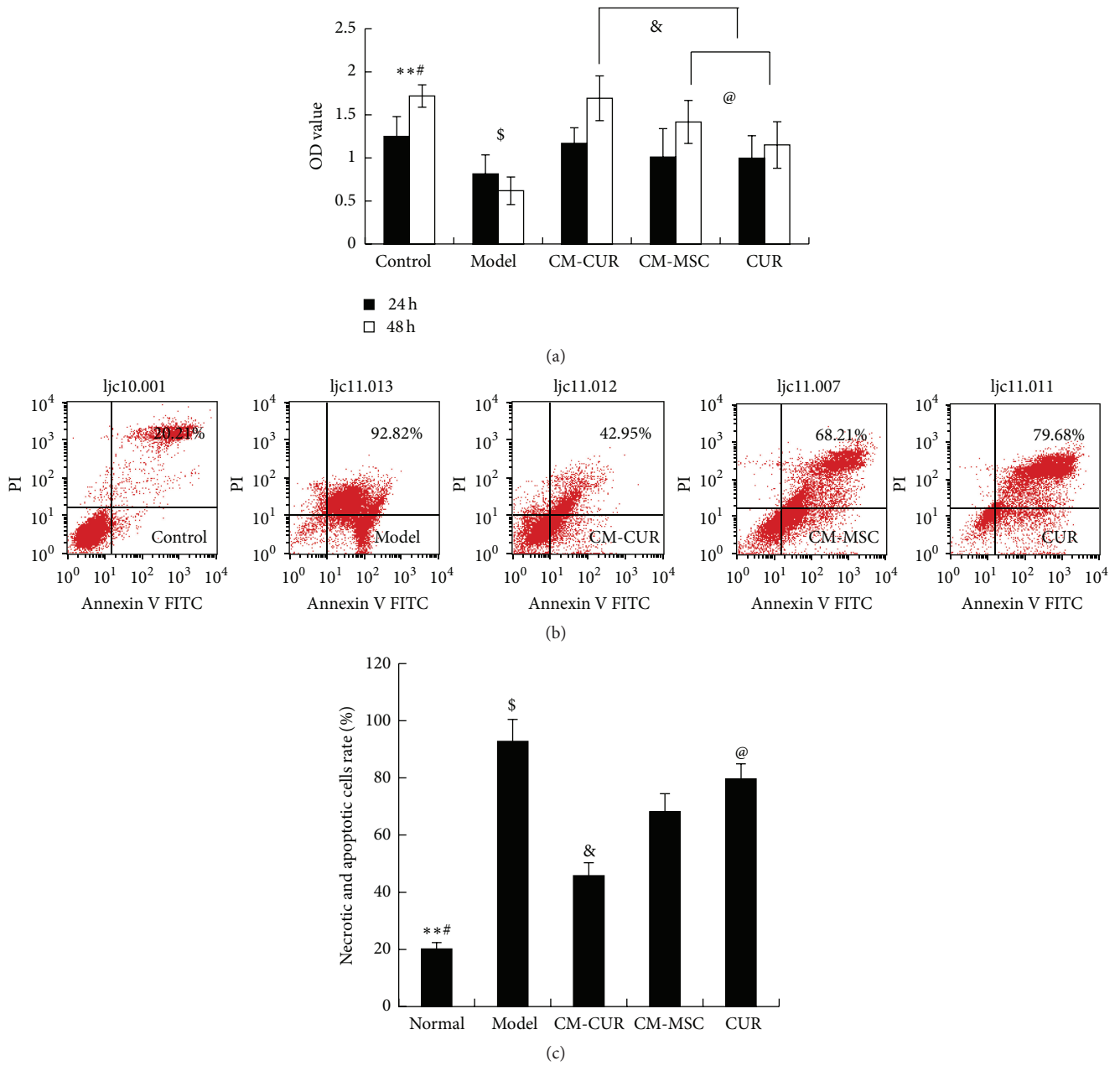


Figure 1: (a) The OD value of PD model cells was gradually increased after treatment with CM-CUR, CM-MSC, and CUR at 24 h and 48 h ($^{\$}p < 0.05$). There were obviously differences of promoting effects on proliferation between the CM-CUR and the other two groups at 48 h ($^{\&}p < 0.05$). At 48 h, the OD values were lower in the CM-MSC and the CUR groups compared with the control group, and the difference was statistically significant ($^{\#}p < 0.05$). The OD value of CUR group was lower than that of the CM-MSC group ($^{@}p < 0.05$). ($^{**}p < 0.01$: the control group versus the model group.) (b) The flow cytometry results showed that the sum of the necrotic rate and apoptotic rate was 20.21% in the normal cells, 92.82% in the model group, and 45.95%, 68.21%, and 79.68% in the CM-CUR, CM-MSC, and CUR groups, respectively. (c) Statistical analyses showed that the cell necrotic and apoptotic rate were lowest in the CM-CUR group and showed statistically significant differences compared with the model group ($^{**}p < 0.01$), as well as the other two groups ($^{@}p < 0.05$). ($^{\$}p < 0.01$ the normal control group versus the three groups, $^{\#}p < 0.05$ CM-MSC and CUR groups versus the model group.)

cells were incubated with 500 μmol/L MPP+ for 24 h and the OD value was gradually increased after treatment with CM-CUR, CM-MSC, and CUR ($^{\$}p < 0.05$), respectively. At 24 h, the three groups did not display significant differences compared with the control group, while at 48 h, only the OD value in the CM-CUR group exceeded that of the control group without a statistically significant difference. The OD values were lower in the CM-MSC and the CUR groups and exhibited a statistically significant difference ($^{\#}p < 0.05$) (Figure 1(a)).

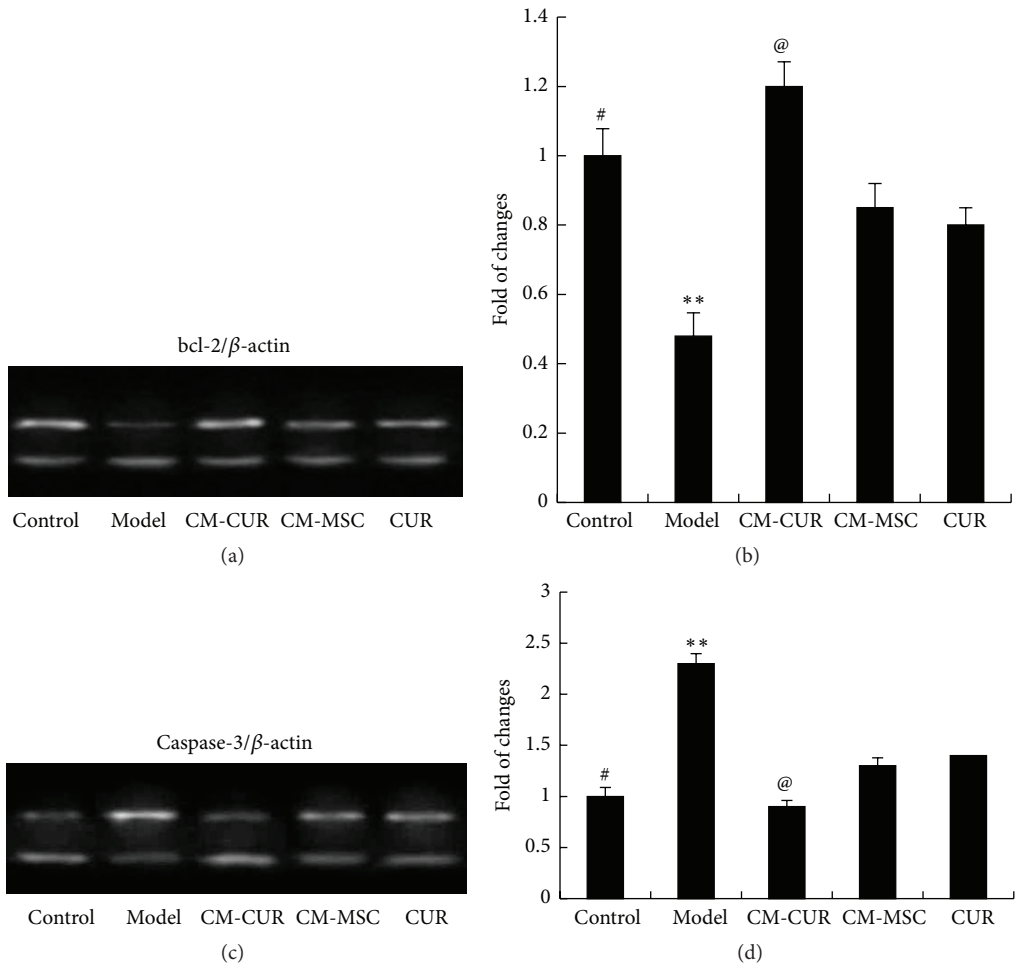


Figure 2: Expressions of bcl-2 and caspase detected by RT-PCR: the bcl-2 mRNA expression was elevated (a, b) and caspase-3 mRNA expression was reduced (c, d) after the PD cell model was treated with CM-CUR, CM-MSC, and CUR for 48 h and showed statistically significant differences compared with the model group ($^{**}p < 0.01$). The effect was still the strongest in the CM-CUR group ($^{@}p < 0.05$), which did not show any significant difference compared with the control group. The mRNA expressions in the CM-MSC and CUR groups were lower than the control group ($^{\#}p < 0.05$), while the difference between the CM-MSC and CUR groups was not significant.

The flow cytometry results showed that the sum of the necrotic rate and apoptotic rate was 20.21% in the normal cells, 92.82% in the model group, and 45.95%, 68.21%, and 79.68% in the CM-CUR, CM-MSC, and CUR groups, respectively (Figure 1(b)). Compared with the control group, the model group was very seriously injured ($^{**}p < 0.01$). Among the three groups, the cell necrotic rate and apoptotic rate were lowest in the CM-CUR group ($^{\&}p < 0.05$), followed by the CM-MSC group and CUR group and they appeared significantly different compared with the model group ($^{\$}p < 0.05$, Figure 1(c)).

Then we detected the apoptosis related factors bcl-2 and caspase-3 using RT-PCR. The bcl-2 mRNA expression was elevated (Figures 2(a) and 2(b)) and caspase-3 mRNA expression was decreased (Figures 2(c) and 2(d)) after the PD cell model was processed with CM-CUR, CM-MSC, and CUR for 48 h and showed statistically significant difference compared with the model group ($^{**}p < 0.01$). The effect was still the strongest in the CM-CUR group ($^{@}p < 0.05$), which did not show significant difference compared with the control group. The mRNA expressions in the CM-MSC and CUR groups were lower than the control group with statistically significant differences ($^{\#}p < 0.05$), while the difference between the CM-MSC and CUR groups was not significant.

3.3. CM-CUR Significantly Elevated the Expressions of TH, DAT, and DA in PC12 Cells. TH, DAT, and DA are critical for DA neuron cells and can be considered as the markers of the DA neurons. Western blot results showed that the expressions of TH and DAT were elevated in the PC12 PD model cells after treatment with CM-CUR and CM-MSC for 48 h ($^{**}p < 0.01$), with no significant differences in the CUR group compared with the model group. Moreover, the CM-CUR group presented a most significant effect compared with the CM-MSC and CUR groups ($^{@}p < 0.05$). Compared with the control group, the expressions of TH and DAT in the CM-CUR group did not show a statistically significant difference, while those in the CM-MSC and CUR groups

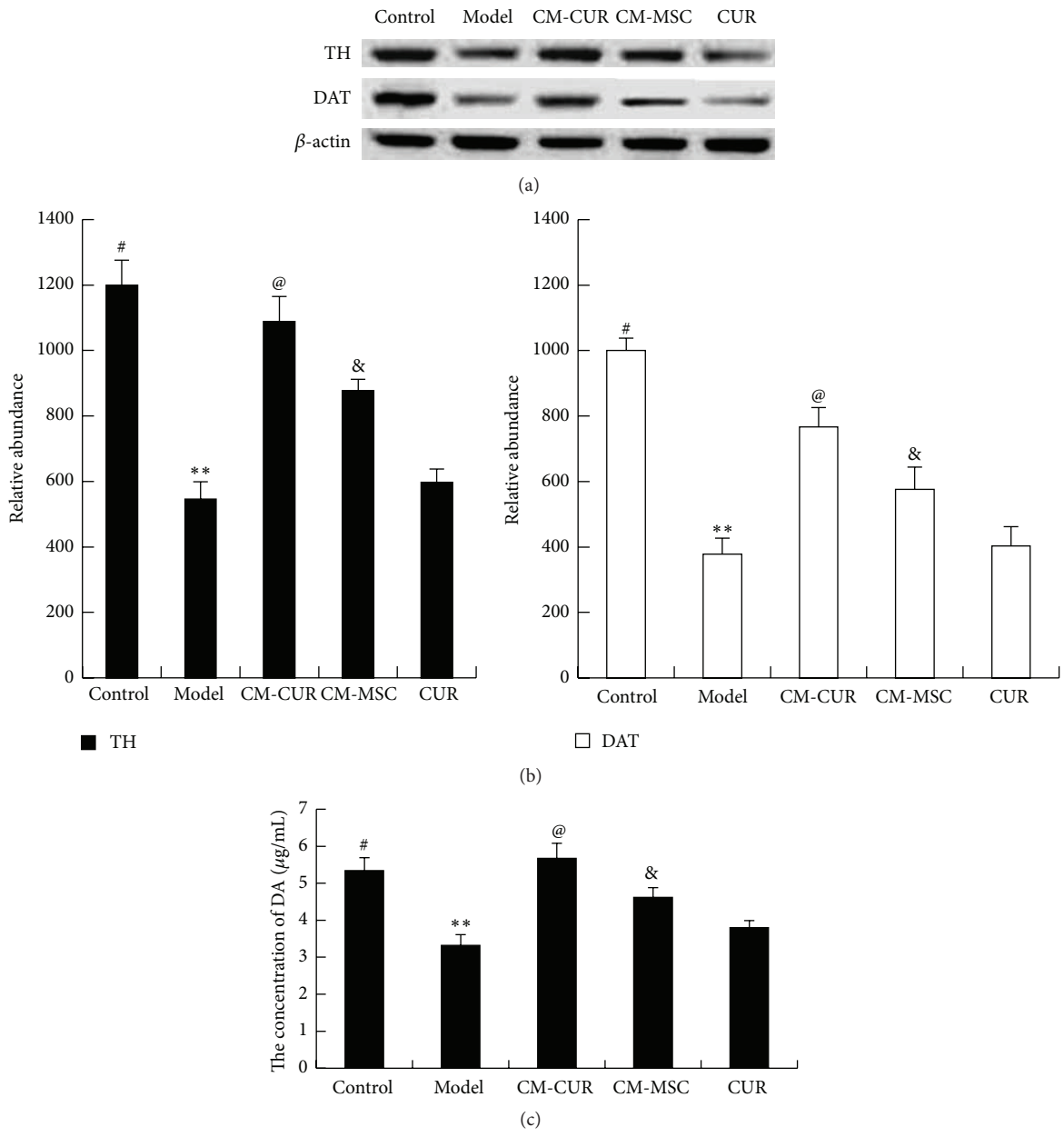


Figure 3: The expression of TH, DAT, and DA in PD model cells. (a), (b) The Western blot assay results showed that compared with the model group, the expressions of TH and DAT were elevated in the three groups ($^{**}p < 0.01$). The CM-CUR presented a more significant effect compared with the CM-MSC and CUR ($^{@}p < 0.05$), and the effect was more significant in CM-MSC group than the CUR group ($^{\&}p < 0.05$). Compared with the control group, the expressions of TH and DAT in the CM-CUR group did not show statistically significant differences, while those in the CM-MSC and CUR groups were significantly lower than those in the control group ($^{\#}p < 0.05$). (c) ELISA results showed that CM-CUR, CM-MSC, and CUR could promote the DA secretion of PC12 PD model cells, showing a tendency consistent with the expressions of TH and DAT.

were significantly lower ($^{\#}p < 0.05$) (Figures 3(a) and 3(b)). According to ELISA results, the DA concentration secreted by cells in the control group was 5.34 μg/mL, while those in the model and CM-CUR, CM-MSC, and CUR groups were 3.32 μg/mL, 5.67 μg/mL, 4.62 μg/mL, and 3.8 μg/mL, respectively. This suggests that the expression tendency of DA was roughly consistent in that of TH and DAT. Therefore, these results indicated that the PC12 cells tend to differentiate into DA neurons in a certain degree, with the most significant effect in the CM-CUR group, followed by CM-MSC and CUR groups.

3.4. CM-CUR Promoted the Differentiation of PC12 Cells into Neurons. After treatment with CM-CUR, CM-MSC, and CUR for 96 h, the MAP2 in the PC12 cells were stained using immunohistochemistry. The results showed that in the control group, the PC12 cells displayed round, short

fusiform or triangle shapes, with a diameter of 6–8 μm, and some cells had short processes in their poles with a length no longer than 2 μm. In the model group, the PC12 cells presented nuclear condensation, with their cell bodies sharply shrunk and rounded and brown particles in the cytoplasm were significantly decreased. However, after treatment with CM-CUR, CM-MSC, and CUR, the brown particles were significantly increased. In addition, the cells appeared to have processes with different sizes and numbers, as some were longer and thicker with small processes up to more than 10 millimeters, resembling axons (as shown by the arrows). These phenomena were obvious in the CM-CUR and CM-MSC groups, particularly the CM-CUR group, while almost no similar protrusions were observed in the CUR group (Figure 4(a)).

Furthermore, the expression of MAP2 was detected by a Western blot assay. The results suggested that the MAP2 expression was gradually enhanced in the CM-CUR and CM-MSC groups and showed statistically significant differences compared with the model group ($^{**}p < 0.01$). Among the three groups, the MAP2 expression was strongest in the CM-CUR group ($^{@}p < 0.05$), while the expression in the CM-MSC group was significantly higher than that of the CUR group ($^{\&}p < 0.05$). Compared with the control group, the MAP2 expression was higher in the CM-CUR and CM-MSC groups ($^{\#}p < 0.05$), with the relative abundance of the CM-CUR being highest (Figures 4(b) and 4(c)). These phenomena suggested that MSC supernatant can effectively promote the differentiation of the PC12 cells, and this outcome was even more significant with CUR treatment. In order to investigate the mechanism, we further detected the changes of NGF concentration in the supernatant and the results are discussed as follows.

3.5. The Changes of Various Cytokines in PC12 Supernatant. In recent years, evidence suggested that the inflammatory reaction in the brain is involved in the degeneration process of DA neurons. Therefore, we also detected the variations of various cytokines in the PD cell model supernatant (Figure 5). The results showed that changes were observed in IL-6, IL-10, and NGF while the other three cytokines, IL-1β, TNF-α, and IFN-γ, did not show significant differences compared with the model and normal control groups. In this study, compared with the model group, the IL-6 and IL-10 expressions were elevated after the PC12 cells were applied with CM-CUR, CM-MSC, and CUR ($^{*}p < 0.05$). Among these, the expression was highest in the CM-CUR group ($^{@}p < 0.05$), followed by the CM-MSC and CUR groups, with the expression neglecting to show significant differences between the latter two groups. NGF is not expressed in the normal PC12 cells ($^{\#}p < 0.05$) and the model group ($^{\#}p < 0.05$) but gradually increased after the PC12 cells were treated with CM-CUR and CM-MSC, not CUR, and were accompanied by the above morphological changes. The NGF expression was significantly higher in the CM-CUR group compared with the CM-MSC group ($^{@}p < 0.05$).

3.6. Expressions of NO and iNOS Show Greatest Decline in the CM-CUR Group. As shown in Figures 6(a) and 6(b), the expression of iNOS was detected using a Western blot assay. The results showed that the expressions of iNOS were low in the control group but dramatically increased in the model group and gradually decreased after treatment with CM-CUR, CM-MSC, and CUR, all showing significant differences compared with the model group ($^{**}p < 0.0$). Among these, the decrease was most significant in the CM-CUR group ($^{@}p < 0.05$), followed by the CUR group and CM-MSC group, respectively, while the differences between the latter two groups were not obvious. However, the iNOS expressions in the three treatment groups were still very high compared with the control group ($^{\#}p < 0.01$). The NO content in the supernatant was detected using Griess method, and the results indicated that its expression tendency was consistent with that of the iNOS (Figure 6(c)).

4. Discussion

In recent years, more and more evidences have proved that hUC-MSC is suitable for the treatment of PD [10, 11, 21]. Weiss transplanted the undifferentiated hUC-MSC into the striatum of PD rats and found that the clinical symptoms of rats were significantly improved with the number of dopaminergic neurons in the injured site increased [22]. After transplantation, brain tumor, immune rejection, and any rotational behavior were not observed in normal rats. Some researchers performed genetic modification in hUC-MSC and transplanted it into the striatum and nigra of PC rhesus and the results showed that the rhesus performances were significantly improved [10]. However, treatment of PD with CUR-modified hUC-MSC has not been reported yet. CUR is a kind of phenolic pigment extracted from turmeric rhizome and is an important active ingredient of the CUR. Previous studies have shown that PD treatment with CUR is associated with its antioxidation and antiapoptosis effects [23–25]. In 2012, researchers discovered that the CUR tends to bind to α-Syn so as to prevent its aggregation in neurons, whose results are of significance in treating PD with pathological feature of Lewy bodies. But there are still many issues, as the CUR is not able to penetrate the BBB. Lapidus believed that the actual drug effects of the CUR may be very limited [17]. In contrast, hUC-MSC are able to pass through the BBB [7–10].

Based on their merits and demerits, we hypothesized that the combination of hUC-MSC and CUR may be a better treatment of PD. CCK-8 and flow cytometry revealed that CM-CUR strongly promotes the proliferation of apoptotic PC12 PD model cells compared with the hUC-MSC supernatant and CUR. Therefore, it suggests that the CUR-modified hUC-MSC has a meaningful effect in repairing PD model cells. The results of RT-PCR gave a clue that the proliferation effect was related to the apoptosis related factors bcl-2 and caspase-3. bcl-2 protein is the most important member of the bcl-2 family and is always considered to be the apoptosis-inhibiting ingredient [26]. The caspase family plays a very important role in mediating apoptosis of cells. Of these, the caspase-3 is the key execution molecule, which functions in many apoptosis signaling pathways. In the process of transmitting

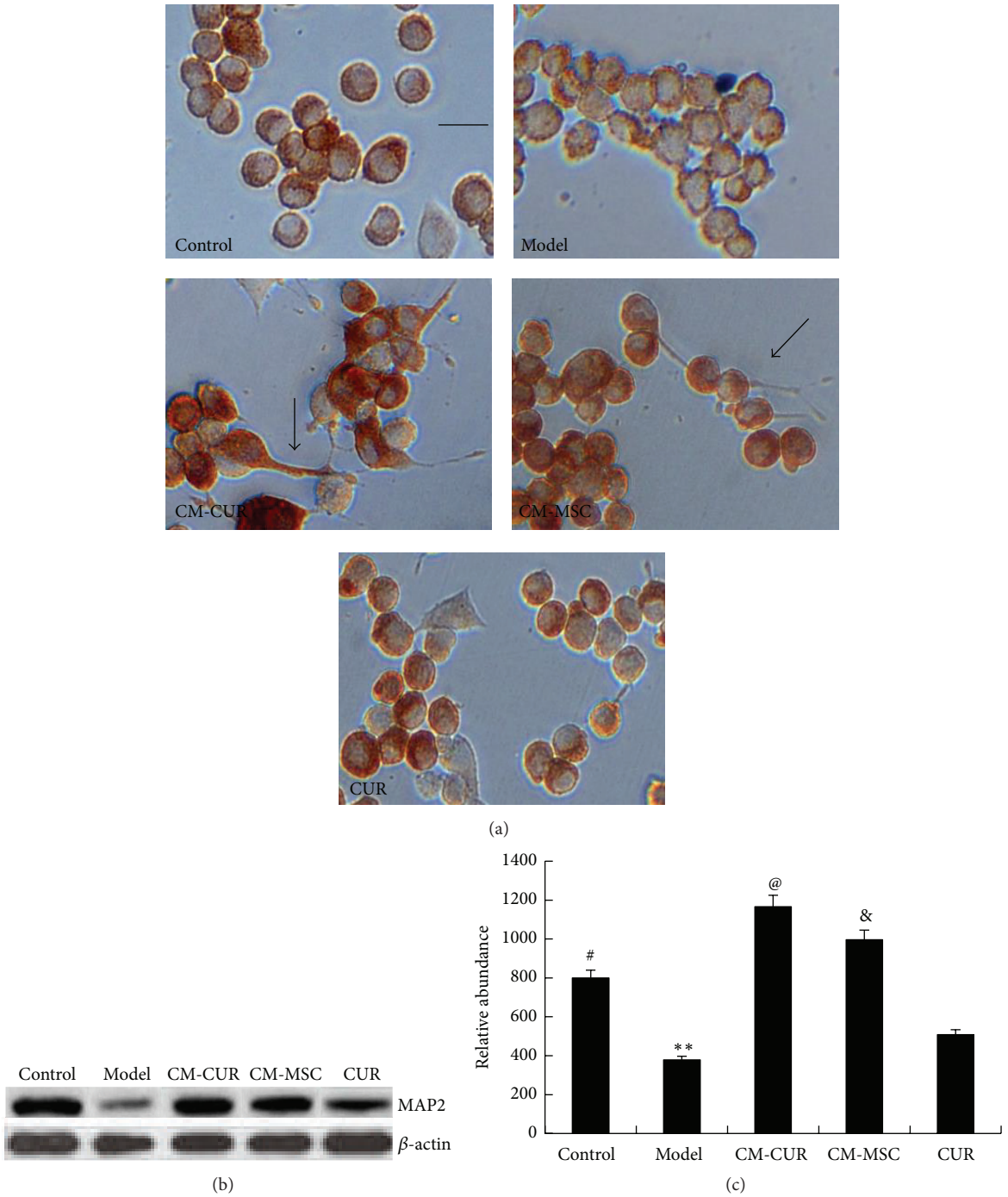


Figure 4: Differentiation of PC12 PD model cells into neurons after treatment with CM-CUR, CM-MSC, and CUR. (a) The immunohistochemistry results showed the PD PC12 cells presented nuclear condensation, cell bodies sharply shrunk and rounded, and significantly decreased brown particles in the cytoplasm. However, after treatment with CM-CUR, CM-MSC, and CUR, the shrinkage of the cell bodies was significantly improved, the brown particles were significantly increased, and the cells appeared with different sizes and numbers (indicated with the arrows, bar is 10 μm). These phenomena were more obvious in the CM-CUR group. (b) The Western blot assay data showed that the MAP2 expression was gradually enhanced in the CM-CUR and CM-MSC groups and showed obvious significant differences compared with the model group ($^{**}p < 0.01$). ($^{@}p < 0.05$ the CM-CUR group versus the other two groups, $^{\#}p < 0.05$ CM-MSC and CUR groups versus the model group, $^{\&}p < 0.05$ CM-MSC groups versus CUR groups.)

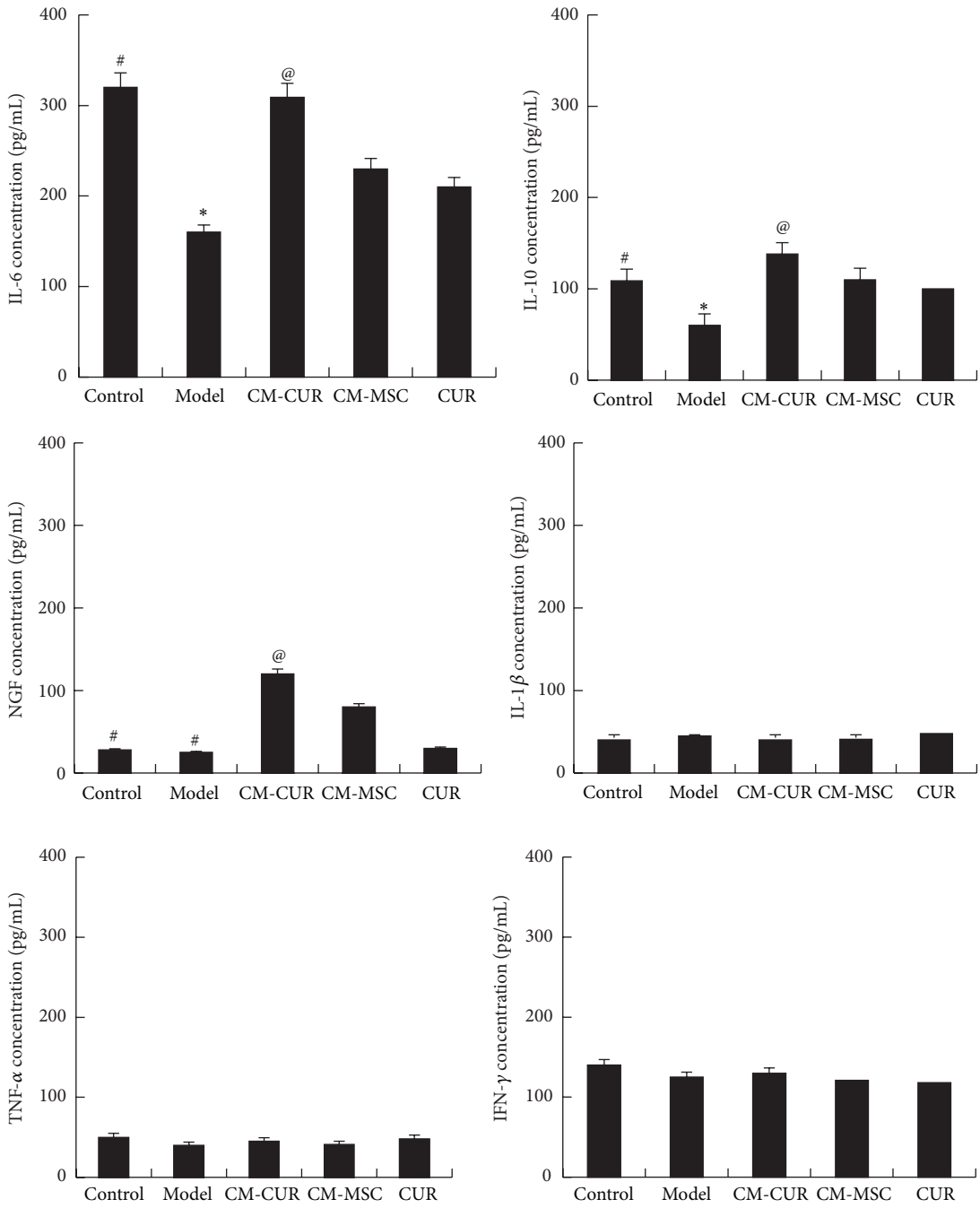


Figure 5: Variations of cytokines in the PC12 cells supernatant. Compared with the model group, the IL-6 and IL-10 expressions were elevated in CM-CUR, CM-MSC, and CUR groups ($^{*}p < 0.05$) and were most significant in the CM-CUR group ($^{@}p < 0.05$), followed by the CM-MSC and CUR groups. NGF was not expressed in the normal PC12 cells ($^{\#}p < 0.05$) or the model group ($^{\#}p < 0.05$). However, the NGF expression gradually increased after treatment with CM-CUR and CM-MSC ($^{@}p < 0.05$ CM-CUR group versus CM-MSC group). The other cytokines, IL-1β, TNF-α, and IFN-γ, did not show significant differences compared with the control and model groups.

apoptosis, it is generally believed that the bcl-2 plays a role in the upstream of caspase-3 through suppressing its activation [27].

Subsequently, several cell markers of DA neurons were detected to assess the efficacies of promoting differentiation of PC12 cells into DA neurons of CM-CUR, CM-MSC, and CUR. The results revealed that the CM-CUR presented a stronger effect in elevating the expressions of TH, DAT, and DA in comparison with the CM-MSC and CUR. PD is a kind of neurodegenerative disease due to a serious shortage of nigrostriatal DA. TH is a key enzyme in the pathway of DA biosynthesis, wherein its increase and decrease may directly affect the DA contents and possibly induce a series of abnormal changes as a secondary factor in the pathophysiology

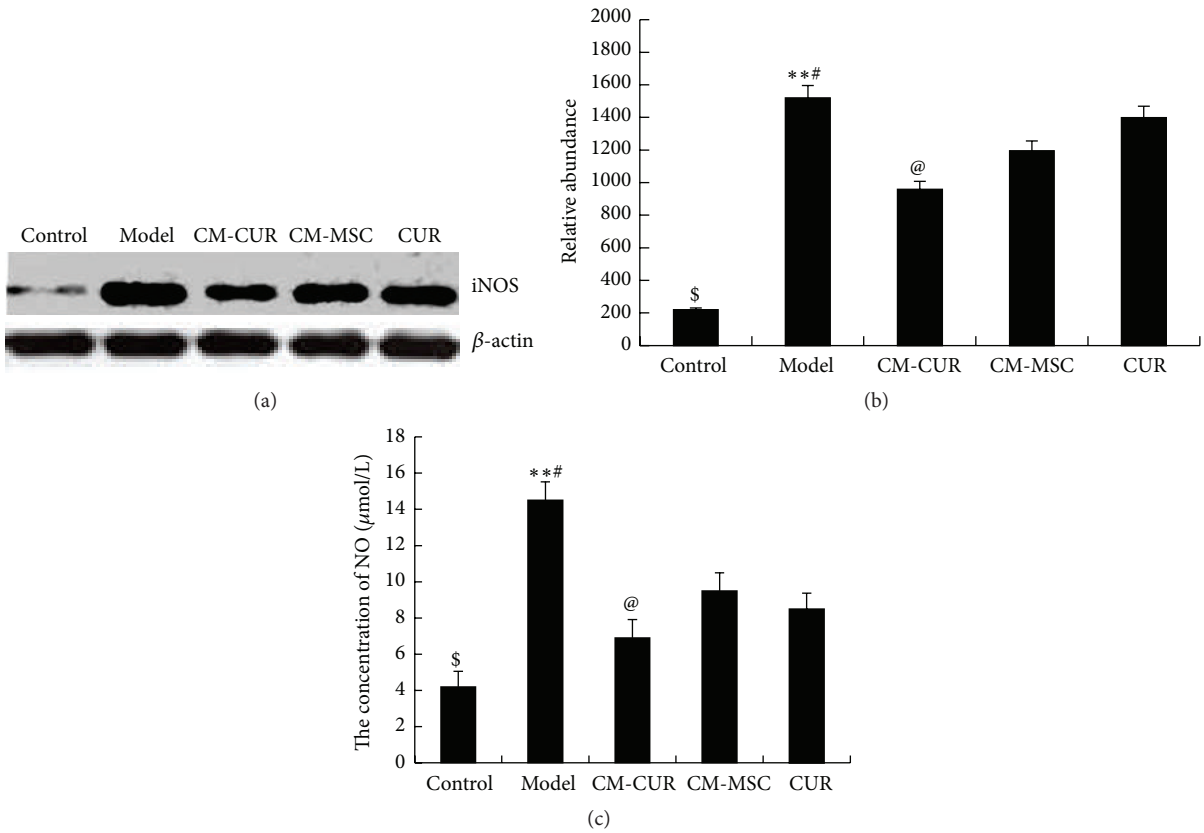


Figure 6: The expressions of NO and iNOS in PD model cells. (a) Western blot assay revealed that iNOS was rarely expressed in the control group, with its expression suddenly increased in the model group and gradually decreased after treatment with CM-CUR, CM-MSC, and CUR ($^{**}p < 0.05$). Among these, the effect was most obvious in the CM-CUR group ($^{@}p < 0.05$), followed by the CUR group and CM-MSC groups, with the difference between the latter two groups not being significant. However, the iNOS expressions in the three groups were still very high compared with the control group ($^{\#}p < 0.01$). (b) The tendency of NO content in the supernatant was consistent with that of the iNOS detected by Griess method.

of PD [5, 6, 28]. DAT is a glycoprotein molecule located in the presynaptic membrane of dopamine neurons, mainly obtained by the synthesis of nerve cell bodies, dendrites, and axons of the nigrostriatal dopamine, and plays an important role in the recovery of the dopamine [5, 6, 29].

Furtherly, we focus on the differentiation of PC12 PD model cells into neuron-like cells after treatment with CM-CUR and hUC-MSC. We detected the expression of MAP2, which is a specific marker of neurons and is present in both the cell bodies and dendrites of neurons but is more prevalent in the dendrites. It can be considered as a labeling protein of the neurons and plays an important role in the development, differentiation, shaping of neurons, and acquisition of neuronal polarity [30, 31]. In our results, PC12 cells tended to express MAP2, while the MAP2 protein was significantly decreased in the model group. After receiving three treatments, the MAP2 expression was restored, while the expression was still highest in the CM-CUR group, followed by CM-MSC group, and was not significant in the CUR group. We hypothesized that the NGF expression would be increased in the supernatants of the CM-CUR and CM-MSC groups, which was confirmed by ELISA. Meanwhile, we also detected the expressions of other cytokines, among which the expressions of IL-6 and IL-10 presented changes. Inflammation plays an important role in the pathogenesis of PD. The activation of glial cells and damage of cytokines may lead to degeneration and even death of DA neurons, which means that the inflammation of the central nervous system tends to aggravate the occurrence and development of PD [1, 2]. Three common proinflammatory cytokines, IL-1β, TNF-α, and IFN-γ, play important roles in PD [32, 33]. Their expressions were relatively low in the normal control, model, and treatment groups, potentially because the PC12 cells themselves do not secrete the above three cytokines. IL-6 plays different roles in PD, one of which being having a strong proinflammatory effect [33, 34]. Müller et al. and Beharka et al. believed that the IL-6 can promote the repair and regeneration of neurons in PD patients [35, 36], while Gadient and Otten believed that the IL-6 could protect the injured neurons but also induce the degeneration and necrosis of the neurons [37]. Our results showed that the IL-6 expression in the supernatant of PC12 PD model cells was increased after treatment with CM-CUR, CM-MSC, and CUR and its expression was highest in the CM-CUR group. Therefore, further studies are needed to investigate the role of IL-6 in PD. IL-10 is a kind of single chain glycoprotein

produced by Th2 cells, and it usually inhibits the syntheses and expressions of monocyte-macrophage inflammatory mediators IL-1, IL-8, and TNF-α [38, 39]. ELISA results showed that the IL-10 content in the PC12 supernatant was elevated and was highest in the CM-CUR group. There are systemic mitochondrial dysfunctions, oxygen free radicals, and oxidative stress reactions in PD patients [1, 2]. MPP+ is the active metabolite of the neurotoxin 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP) and it can promote the generation of free radicals and oxidative stress reactions after entering the cells while stimulating in vivo environments of PD patients. NO is a kind of oxygen free radical that can stimulate cells to produce excitatory glutamate and cause direct damage [40]. High-level NO can penetrate the mitochondrial membrane and suppresses the vitality of various complexes in the mitochondrial respiratory chain, causing irreversible oxidative damage to the cells [2, 40]. Therefore, we selected NO as one of the outcomes after the PC12 cells were damaged by MPP+. Nitric oxide synthase (NOS) is responsible for the synthesis of NO. Currently, three kinds NOS have been discovered in the human body and we selected synthase II type as another outcome since it is expressed only after the cells are stimulated, so called iNOS [2, 40]. The iNOS has been found to extensively participate in the expression of chemokines and generation of reactive oxygen products. Numerous experiments have proved that the CUR is able to clear the oxygen free radicals and plays a role in antioxidation. Very few studies have been reported in which the stem cells have the effect of inhibiting iNOS expression, and whether the hUC-MSC has the above effects or not is still unknown. However, our experiments revealed that the CUR-modified hUC-MSC displayed the strongest effect of antioxidative stress.

In summary, all the results indicated that CUR-activated hUC-MSC tends to display significant efficacy in proliferation and apoptosis, differentiation into neurons, and antioxidative ability compared with the hUC-MSC and CUR. This presents a powerful combination of the effects of two ingredients. Therefore, a perfect combination of hUC-MSC and CUR is going to be a new type of biological therapy for repairing PD in the future.

Ethical Approval

All experiments were reviewed by the Ethics Committee of General Hospital of Jinan Military Region and the 148th Hospital.

Competing Interests

The authors declare that there are no competing interests regarding the publication of this paper.

Authors' Contributions

Li Jinfeng, Wang Yunliang, and Liu Xinshan contributed equally to this work.

Acknowledgments

Thanks are due to Dr. Jian-min Xing for coming from Peking Union Medical College; he made great contributions to the statistics in these studies.

References

[1] G. C. Pluck and R. G. Brown, "Apathy in Parkinson's disease," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 73, no. 6, pp. 636–642, 2002.

[2] A. J. Yarnall, L. Rochester, and D. J. Burn, "Mild cognitive impairment in parkinson's disease," *Age and Ageing*, vol. 42, no. 5, pp. 567–576, 2013.

[3] L. Kuramoto, J. Cragg, R. Nandhagopal et al., "The nature of progression in parkinson's disease: an application of non-linear, multivariate, longitudinal random effects modelling," *PLoS ONE*, vol. 8, no. 10, Article ID e76595, 2013.

[4] A. C. Nóbrega, P. Pinho, M. Deiró, and N. Argolo, "Levodopa treatment in Parkinson's disease: how does it affect dysphagia management?" *Parkinsonism and Related Disorders*, vol. 20, no. 3, pp. 340–341, 2014.

[5] K. Yamada, S. Goto, J.-I. Kuratsu et al., "Stereotactic surgery for subthalamic nucleus stimulation under general anesthesia: a retrospective evaluation of Japanese patients with Parkinson's disease," *Parkinsonism and Related Disorders*, vol. 13, no. 2, pp. 101–107, 2007.

[6] G. Deuschl, J. Herzog, G. Kleiner-Fisman et al., "Deep brain stimulation: postoperative issues," *Movement Disorders*, vol. 21, no. 14, pp. S219–S237, 2006.

[7] A. Bjorklund and J. H. Kordower, "Cell therapy for Parkinson's disease: what next?" *Movement Disorders*, vol. 28, no. 1, pp. 110–115, 2013.

[8] R. Sharma, C. R. McMillan, and L. P. Niles, "Neural stem cell transplantation and melatonin treatment in a 6-hydroxydopamine model of Parkinson's disease," *Journal of Pineal Research*, vol. 43, no. 3, pp. 245–254, 2007.

[9] D. F. Emerich, "Cell transplantation for Parkinson's disease," *Cell Transplantation*, vol. 11, no. 1, pp. 1–3, 2002.

[10] M. Yan, M. Sun, Y. Zhou et al., "Conversion of human umbilical cord mesenchymal stem cells in Wharton's jelly to dopamine neurons mediated by the Lmx1a and neurturin in vitro: potential therapeutic application for Parkinson's disease in a rhesus monkey model," *PLoS ONE*, vol. 8, no. 5, Article ID e64000, 2013.

[11] J.-F. Li, H.-L. Yin, A. Shuboy et al., "Differentiation of hUC-MSC into dopaminergic-like cells after transduction with hepatocyte growth factor," *Molecular and Cellular Biochemistry*, vol. 381, no. 1-2, pp. 183–190, 2013.

[12] X.-S. Liu, J.-F. Li, S.-S. Wang et al., "Human umbilical cord mesenchymal stem cells infected with adenovirus expressing *HGF* promote regeneration of damaged neuron cells in a Parkinson's disease model," *BioMed Research International*, vol. 2014, Article ID 909657, 7 pages, 2014.

[13] Y. Tizabi, L. L. Hurley, Z. Qualls, and L. Akinfiresoye, "Relevance of the anti-inflammatory properties of curcumin in neurodegenerative diseases and depression," *Molecules*, vol. 19, no. 12, pp. 20864–80879, 2014.

[14] Y. Jaisin, A. Thampithak, B. Meesarapee et al., "Curcumin I protects the dopaminergic cell line SH-SY5Y from 6-hydroxydopamine-induced neurotoxicity through attenuation

of p53-mediated apoptosis," *Neuroscience Letters*, vol. 489, no. 3, pp. 192–196, 2011.

[15] Z. Qualls, D. Brown, C. Ramlochansingh, L. L. Hurley, and Y. Tizabi, "Protective effects of curcumin against rotenone and salsolinol-induced toxicity: implications for parkinson's disease," *Neurotoxicity Research*, vol. 25, no. 1, pp. 81–89, 2014.

[16] Y. H. Siddique, F. Naz, and S. Jyoti, "Effect of curcumin on lifespan, activity pattern, oxidative stress, and apoptosis in the brains of transgenic drosophila model of Parkinson's disease," *BioMed Research International*, vol. 2014, Article ID 606928, 6 pages, 2014.

[17] B. Ahmad and L. J. Lapidus, "CUR prevents aggregation in α-Synuclein by increasing reconfiguration rate," *The Journal of Biological Chemistry*, vol. 287, no. 12, pp. 9193–9199, 2012.

[18] A.-L. Chen, C.-H. Hsu, J.-K. Lin et al., "Phase I clinical trial of curcumin, a chemopreventive agent, in patients with high-risk or pre-malignant lesions," *Anticancer Research B*, vol. 21, no. 4, pp. 2895–2900, 2001.

[19] T. Esatbeyoglu, P. Huebbe, I. M. A. Ernst, D. Chin, A. E. Wagner, and G. Rimbach, "Curcumin—from molecule to biological function," *Angewandte Chemie—International Edition*, vol. 51, no. 22, pp. 5308–5332, 2012.

[20] L.-H. Zhu, X. Bai, N. Zhang, S.-Y. Wang, W. Li, and L. Jiang, "Improvement of human umbilical cord mesenchymal stem cell transplantation on glial cell and behavioral function in a neonatal model of periventricular white matter damage," *Brain Research*, vol. 1563, pp. 13–21, 2014.

[21] J. F. Li, D. J. Zhang, T. Geng et al., "The potential of human umbilical cord-derived mesenchymal stem cells as a novel cellular therapy for multiple sclerosis,," *Cell Transplantation*, vol. 23, supplement 1, pp. S113–S122, 2014.

[22] M. L. Weiss, S. Medicetty, A. R. Bledsoe et al., "Human umbilical cord matrix stem cells: preliminary characterization and effect of transplantation in a rodent model of Parkinson's disease," *Stem Cells*, vol. 24, no. 3, pp. 781–792, 2006.

[23] J. Chen, X. Q. Tang, J. L. Zhi et al., "Curcumin protects PC12 cells against 1-methyl-4-phenylpyridinium ion-induced apoptosis by bcl-2-mitochondria-ROS-iNOS pathway," *Apoptosis*, vol. 11, no. 6, pp. 943–953, 2006.

[24] J. Yang, S. Song, J. Li, and T. Liang, "Neuroprotective effect of curcumin on hippocampal injury in 6-OHDA-induced Parkinson's disease rat," *Pathology Research and Practice*, vol. 210, no. 6, pp. 357–362, 2014.

[25] H. Lv, J. Liu, L. Wang et al., "Ameliorating effects of combined curcumin and desferrioxamine on 6-OHDA-induced rat mode of Parkinson's disease," *Cell Biochemistry and Biophysics*, vol. 70, no. 2, pp. 1433–1438, 2014.

[26] Y. Wang, J. Gao, Y. Miao et al., "Pinocembrin protects SH-SY5Y cells against MPP+-induced neurotoxicity through the mitochondrial apoptotic pathway," *Journal of Molecular Neuroscience*, vol. 53, no. 4, pp. 537–545, 2014.

[27] W. Wu, O. W. Wan, and K. K. K. Chung, "S-nitrosylation of XIAP at Cys 213 of BIR2 domain impairs XIAP's anti-caspase 3 activity and anti-apoptotic function," *Apoptosis*, vol. 20, no. 4, pp. 491–499, 2015.

[28] J. Corbitt, T. Hagerty, E. Fernandez, W. W. Morgan, and R. Strong, "Transcriptional and post-transcriptional regulation of tyrosine hydroxylase messenger RNA in PC12 cells during persistent stimulation by VIP and PACAP38: differential regulation by protein kinase A and protein kinase C-dependent pathways," *Neuropeptides*, vol. 36, no. 1, pp. 34–45, 2002.

[29] A. Storch, A. C. Ludolph, and J. Schwarz, "Dopamine transporter: involvement in selective dopaminergic neurotoxicity and degeneration," *Journal of Neural Transmission*, vol. 111, no. 10, pp. 1267–1286, 2004.

[30] H. Cho, Y.-K. Seo, S. Jeon, H.-H. Yoon, Y.-K. Choi, and J.-K. Park, "Neural differentiation of umbilical cord mesenchymal stem cells by sub-sonic vibration," *Life Sciences*, vol. 90, no. 15-16, pp. 591–599, 2012.

[31] T. Yan, K. O. Skaftnesmo, L. Leiss et al., "Neuronal markers are expressed in human gliomas and NSE knockdown sensitizes glioblastoma cells to radiotherapy and temozolomide," *BMC Cancer*, vol. 11, article 524, 2011.

[32] C. Barcia, C. M. Ros, V. Annese et al., "IFN-γ signaling, with the synergistic contribution of TNF-α, mediates cell specific microglial and astroglial activation in experimental models of Parkinson's disease," *Cell Death and Disease*, vol. 3, no. 8, article e382, 2012.

[33] D. D. Lofrumento, G. Nicolardi, A. Cianciulli et al., "Neuroprotective effects of resveratrol in an MPTP mouse model of Parkinson's-like disease: possible role of SOCS-1 in reducing pro-inflammatory responses," *Innate Immunity*, vol. 20, no. 3, pp. 249–260, 2014.

[34] R. Haddadi, A. Mohajjel Nayebi, and S. E. Brooshghalan, "Pretreatment with silymarin reduces brain myeloperoxidase activity and inflammatory cytokines in 6-OHDA hemi-parkinsonian rats," *Neuroscience Letters*, vol. 555, pp. 106–111, 2013.

[35] A. A. Beharka, M. Meydani, D. Wu, L. S. Leka, A. Meydani, and S. N. Meydani, "Interleukin-6 production does not increase with age," *Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, vol. 56, no. 2, pp. B81–B88, 2001.

[36] T. Müller, D. Blum-Degen, H. Przuntek, and W. Kuhn, "Interleukin-6 levels in cerebrospinal fluid inversely correlate to severity of Parkinson's disease," *Acta Neurologica Scandinavica*, vol. 98, no. 2, pp. 142–144, 1998.

[37] R. A. Gadient and U. H. Otten, "Interleukin-6 (IL-6)—a molecule with both beneficial and destructive potentials," *Progress in Neurobiology*, vol. 52, no. 5, pp. 379–390, 1997.

[38] J. Infante, I. García-Gorostiaga, P. Sánchez-Juan et al., "Inflammation-related genes and the risk of Parkinson's disease: a multilocus approach," *European Journal of Neurology*, vol. 15, no. 4, pp. 431–433, 2008.

[39] P. Vasseur, I. Devaure, J. Sellier et al., "High plasma levels of the pro-inflammatory cytokine IL-22 and the anti-inflammatory cytokines IL-10 and IL-1ra in acute pancreatitis," *Pancreatology*, vol. 14, no. 6, pp. 465–469, 2014.

[40] Z.-K. Xiong, J. Lang, G. Xu et al., "Excessive levels of nitric oxide in rat model of parkinson's disease induced by rotenone," *Experimental and Therapeutic Medicine*, vol. 9, no. 2, pp. 553–558, 2015.

Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 9142190, 12 pages
http://dx.doi.org/10.1155/2016/9142190

Research Article

Impacts of Nonsynonymous Single Nucleotide Polymorphisms of Adiponectin Receptor 1 Gene on Corresponding Protein Stability: A Computational Approach

Md. Abu Saleh,[1] Md. Solayman,[1] Sudip Paul,[1] Moumoni Saha,[1] Md. Ibrahim Khalil,[1,2] and Siew Hua Gan[2]

[1]*Department of Biochemistry and Molecular Biology, Jahangirnagar University, Savar, Dhaka 1342, Bangladesh*
[2]*Human Genome Centre, School of Medical Sciences, Universiti Sains Malaysia, 16150 Kubang Kerian, Kelantan, Malaysia*

Correspondence should be addressed to Sudip Paul; sudippaul.bcmb@gmail.com

Received 15 February 2016; Accepted 11 April 2016

Academic Editor: Ryuji Hamamoto

Despite the reported association of adiponectin receptor 1 (*ADIPOR1)* gene mutations with vulnerability to several human metabolic diseases, there is lack of computational analysis on the functional and structural impacts of single nucleotide polymorphisms (SNPs) of the human *ADIPOR1* at protein level. Therefore, sequence- and structure-based computational tools were employed in this study to functionally and structurally characterize the coding nsSNPs of *ADIPOR1* gene listed in the dbSNP database. Our *in silico* analysis by SIFT, nsSNPAnalyzer, PolyPhen-2, Fathmm, I-Mutant 2.0, SNPs&GO, PhD-SNP, PANTHER, and SNPeffect tools identified the nsSNPs with distorting functional impacts, namely, rs765425383 (A348G), rs752071352 (H341Y), rs759555652 (R324L), rs200326086 (L224F), and rs766267373 (L143P) from 74 nsSNPs of *ADIPOR1* gene. Finally the aforementioned five deleterious nsSNPs were introduced using Swiss-PDB Viewer package within the X-ray crystal structure of ADIPOR1 protein, and changes in free energy for these mutations were computed. Although increased free energy was observed for all the mutants, the nsSNP H341Y caused the highest energy increase amongst all. RMSD and TM scores predicted that mutants were structurally similar to wild type protein. Our analyses suggested that the aforementioned variants especially H341Y could directly or indirectly destabilize the amino acid interactions and hydrogen bonding networks of ADIPOR1.

1. Introduction

In recent years, the number of obese individuals has been dramatically increased throughout the world which leads to the acceleration of obesity related health problems [1, 2]. Decreased insulin sensitivity, the most common arena of obesity, predisposes the affected persons to a variety of pathological abnormalities including type 2 diabetes, hypertension, and cardiovascular diseases [3–5]. The concomitance of these diseases has been considered as metabolic syndrome. In multiple studies, it has been reported that genetic variations in the adiponectin gene are associated with these types of diseases [6]. Adiponectin is an adipokine or adipocytokine specially secreted by adipocytes [7] and placenta during pregnancy [8] that circulates at relatively high (2–20 mg/mL) concentrations in the blood stream. Biologically active adiponectin hormone is a collagen-like circulating protein which acts as a principle antidiabetic and antiatherogenic adipokine [9–12]. Reduced adiponectin level in plasma has been observed in obesity, insulin resistance, and type 2 diabetes [9–12]. Adiponectin exerts its insulin sensitizing effects by increasing fatty-acid oxidation via activation of AMP-activated protein kinase (AMPK) peroxisome proliferator-activated receptor-alpha (PPAR-α) [13, 14]. Therefore, adiponectin is anticipated to be a novel therapeutic target for diabetes and the metabolic syndrome.

To employ proper functions, adiponectin binds to a number of receptors. Different studies have identified two receptors named adiponectin receptor-1 (ADIPOR1) and adiponectin receptor-2 (ADIPOR2) (those are homologous to G protein-coupled receptors) as well as one receptor similar to the cadherin family [15, 16]. In human beings,

ADIPOR1 and *ADIPOR2* genes are located at chromosomal locations 1p36.13-q41 and 12p13.31, respectively [17]. The expression of *ADIPOR1* gene is found principally in skeletal muscles but may be presented ubiquitously also, while the expression of *ADIPOR2* is the most abundant in liver [17]. Among these receptors, ADIPOR1 plays crucial roles in regulation of energy homeostasis as well as glucose and lipid metabolism [18]. According to several studies conducted, single nucleotide polymorphisms (SNPs) in *ADIPOR1* gene can hamper the physiological functions exerted by the ADIPOR1 protein. A recent comprehensive investigation on a European-Australian population has established the association of genetic variation in adiponectin receptors with type 2 diabetes [19, 20]. Moreover, some SNPs of *ADIPOR1* gene have been found to exert significant effects on the risk of prostate cancer (rs12733285) [21], insulin resistance (rs1342387) [22], and even liver fat deposition (−1927 T/C) [23]. Although there are several *in vivo* studies describing the association of SNPs in the *ADIPOR1* gene with metabolic disorders [24, 25], computational analysis has not yet been undertaken on the functional and structural consequences of nsSNPs in this gene.

In current years, computational tools are being widely used to characterize the impacts of deleterious nsSNPs in candidate genes by utilizing the information obtained from physicochemical properties of polypeptides [26, 27], conserved sequences across the species [28], and their structural attributes [29]. With the help of computational algorithms, several *in silico* studies have effectively filtrated functional SNPs out of large pool of diseases sensitive SNPs of *BRCA1, ATM* [30], and *PON1* [31] genes based on their functional consequences and structural stabilities. In spite of the availability of undoubted data referring the extensive involvement of *ADIPOR1* gene mutations in human diseases, the computational analysis of nsSNPs is still unveiled.

In this study, the clinical variants of *ADIPOR1* were collected for *in silico* analysis. By utilizing these data, we employed different publicly available bioinformatics tools and databases for a comprehensive analysis of nsSNPs in *ADIPOR1* gene. We also calculated the free energy changes for mutants and wild type ADIPOR1 protein in order to evaluate their stability. This study might be helpful for further investigation in order to discover new therapeutic drugs related to adiponectin receptor 1 associated diseases.

2. Methods and Materials

2.1. SNP Data Mining. The data on the human *ADIPOR1* gene were obtained from web-based data sources such as Online Mendelian Inheritance in Man (OMIM; http://www.ncbi.nlm.nih.gov/omim/) and the National Center for Biological Information (http://www.ncbi.nlm.nih.gov/). The information about SNPs of *ADIPOR1* gene of *Homo sapiens* was collected from the dbSNP-NCBI (http://www.ncbi.nlm.nih.gov/SNP/) [32] for further computational analysis. The protein sequence of *ADIPOR1* gene was obtained from UniProtKB database (http://www.uniprot.org/uniprot/).

2.2. Analysis of the Functional Consequence of nsSNPs by Sorting Intolerant from Tolerant (SIFT). SIFT (http://sift.jcvi.org/) predicts the deleterious and tolerated SNPs in order to characterize the consequences of amino acid substitutions on phenotypic and functional changes of protein molecules. By using sequence homology based method, SIFT assumes that significant positions in a protein sequence have been conserved throughout evolution and, therefore, substitutions at these positions may affect protein function. The identification numbers (rsIDs) of each nsSNP of *ADIPOR1* gene were submitted as an input to SIFT server for homology searching. SIFT calculates the SIFT score or tolerance index (TI) score for each nsSNP. The SIFT value ≤ 0.05 indicates the deleterious effect of nonsynonymous variants on protein function [33].

2.3. Investigation of Functional Impacts of nsSNPs by nsSNPAnalyzer. nsSNPAnalyzer (http://snpanalyzer.uthsc.edu/) server was used to predict whether a nsSNP of ADIPOR1 protein affects its phenotypic effect. The input options for nsSNPAnalyzer are protein sequences in FASTA format and detailed information on amino acid substitutions. This server usually uses information contained in the multiple sequence alignment and the 3D structure in order to make a prediction. The prediction of this tool is based on a machine learning method known as Random Forest. The results of this server depict whether an nsSNP is associated with disease or neutral [34].

2.4. Analysis of the Functional Impacts of nsSNPs by Screening for Nonacceptable Polymorphisms (SNAP2). To find the functional effects of nsSNP, SNAP2 (https://rostlab.org/services/snap2web/) server was used. The prediction done by SNAP2 is based on a learning device method known as neural network. In order to make a prediction, SNAP2 utilizes the information of automatically created multiple sequence alignment and also some structural features such as predicted secondary structure and solvent accessibility. FASTA format of protein sequences is only the input option for SNAP2. The output of this server consists of prediction (Effect or neutral), score (ranges from −100 strong neutral prediction to +100 strong effect prediction), and expected accuracy [35].

2.5. Characterization of Functional Consequence of nsSNPs by PolyPhen-2. PolyPhen-2 (http://genetics.bwh.harvard.edu/pph2/) is an advanced version of the PolyPhen tool that was used to find out the possible effect of an amino acid substitution on the structure and function of ADIPOR1 protein. UniProtKB accession number/FASTA sequence and details of amino acid substitutions are required for the input options of PolyPhen-2 server. This tool calculates Naïve Bayes posterior probability that this mutation is damaging and reports estimation of corresponding false positive and true positive rate. A mutation is estimated qualitatively as probably damaging (probabilistic score >0.85), possibly damaging (probabilistic score >0.15), and benign (remaining) with specificity and sensitivity values [36].

2.6. Prediction of Disease Related nsSNPs by SNPs&GO. SNPs&GO (http://snps.biofold.org/snps-and-go/snps-and-go.html) is a support vector machine (SVM) based classifier [37]. This server accurately predicts the mutation related to disease from protein sequence. The probability score greater than 0.5 indicates that the disease related effect is caused by nsSNPs on the function of parent protein. The whole protein sequence in FASTA format is the input for this server. The server also provides the output display for additional two servers such as PHD-SNP [38] and PANTHER [39] algorithms.

2.7. Functional Analysis of nsSNP through Hidden Markov Models (Fathmm). Fathmm (http://fathmm.biocompute.org.uk/inherited.html) not only predicts the potentially deleterious nature of protein variants but also the skill of annotating the molecular and phenotypic consequences of these mutations [40]. This server is composed of two algorithms: sequence/conservation based (unweighted) and other combined sequence conservation with pathogenicity weights (weighted). In this study, we used weighted algorithm because this algorithm is capable of adjusting conservation-based predictions to account for the tolerance of related sequences to mutations.

2.8. Investigation of the Molecular Phenotypic Effects of nsSNPs by SNPeffect. The SNPeffect database 4.0 (http://snpeffect.switchlab.org/) utilizes sequence- and structure-based bioinformatics tools in order to make prediction of molecular phenotypic impacts of nsSNP on *ADIPOR1* gene. This server mainly integrates three different tools such as TANGO, WALTZ, and LIMBO and also uses FoldX server to find out a decision whether the mutation is stabilizing or destabilizing the structure of native proteins. TANGO algorithm identifies the aggregation prone regions in a protein sequence by calculating the hydrophobicity and beta-sheet forming propensity. WALTZ algorithm predicts amyloid forming regions in protein sequences with accuracy and specificity, while LIMBO algorithm predicts a chaperone binding site for the Hsp70 chaperones. The input options are usually composed of FASTA sequence/PDB ID/PDB file/UniProt ID and details of nsSNP [41].

2.9. Prediction of Protein Stability Changes upon nsSNPs by I-Mutant 2.0. I-Mutant 2.0 (http://folding.biofold.org/i-mutant/i-mutant2.0.html) is a support vector machine-(SVM-) based tool which was used to predict the protein stability changes upon nsSNPs. In this study, sequence of protein, temperature (25°C), pH (7), and details of nsSNPs were used as input parameters to this server. The output is a free energy change value ($\Delta\Delta G$) of protein after and before mutation. Positive $\Delta\Delta G$ value concludes that the protein being mutated is of higher stability and vice versa is also true [42].

2.10. Identification of Functional Regions in Proteins by ConSurf. ConSurf (http://consurf.tau.ac.il/) is a web-based tool that automatically analyzes evolutionary conservation of amino acid substitutions in protein by using an empirical Bayesian inference. This server is composed of combining two self-governing servers (ConSeq and ConSurf). After providing the FASTA sequence of ADIPOR1 protein to ConSurf tool, the conserved regions were predicted with conservation grades color-coded onto its surface that can finally be pictured online using the Protein Explorer engine [43].

2.11. Analysis of Impacts of nsSNPs on Surface and Solvent Accessibility of Protein by NetSurfP. The active site of a protein in its three-dimensional conformation can be traced by surface and solvent accessibility region of amino acids of that protein. The FASTA sequence of ADIPOR1 protein was submitted to NetSurfP (http://www.cbs.dtu.dk/services/NetSurfP/) server in order to predict its secondary structure, surface, and solvent accessibility of amino acids [44]. The output of this server provides 3 subclasses defined for solvent accessibility of amino acids: low accessibility (buried), moderate accessibility (partially buried), and high accessibility (exposed).

2.12. Modeling the Molecular Effects of nsSNPs on Protein Structure and Evaluating Their Difference of RMSD Value and TM Score. Structural analysis was done in order to explore the structural deviations and stability differences between native and mutant forms of ADIPOR1 proteins. The crystal structure of ADIPOR1 protein available in Protein Data Bank (PDB) [45] has an ID 3WXV. The ADIPOR1 protein contains 375 amino acids from which 287 amino acids have been resolved in crystal structure with a resolution of 2.90 Å [46]. The Swiss-PDB viewer [47] was utilized in order to carry out amino acid substitutions, followed by the energy minimization of the modeled 3D structure of variants using a version of the GROMOS 43B1 force field in GROMOS96 software package embraced in the Swiss-PDB viewer. TM-Align was used to calculate the TM scores and root mean square deviations (RMSDs) [48].

2.13. Identification of Ligand Binding Sites on Unbound Protein Structure by FTSite. Detection of ligand binding sites on unbound proteins is essential to elucidate the protein structure-function relation and for protein engineering. FTSite (http://ftsite.bu.edu/) predicts ligands or small molecule binding sites of proteins based on experimental evidence with 94% accuracy [49]. The input options of this server generally consist of job name, Protein Data Bank ID (PDB ID) or file, and also PDB chain ID if proteins contain multiple subunits.

2.14. Investigation of Protein-Protein Interactions. Protein-protein interaction networks are important to investigate the functions of the interactions of a particular protein with other proteins at cellular level. Online database resource Search Tool for the Retrieval of Interacting Genes (STRING) was applied to identify the interactions of ADIPOR1 protein with other corresponding proteins [50]. This server provided a unique coverage and ease of access to both experimental and predicted interaction information of ADIPOR1. In this study,

we operated KEGG (http://www.genome.jp/kegg/) PATHWAY and LIGAND to make prediction of the functional networking of ADIPOR1 protein.

3. Results and Discussion

3.1. Retrieval of SNPs. The dbSNP-NCBI database was searched for retrieving the SNPs in the human *ADIPOR1* gene (Gene ID: 51094). A total of 138 SNPs were found in the exonic region, among them 62 (44.93%) were synonymous, 74 (53.62%) nonsynonymous and missense, 1 (0.72%) nonsynonymous and nonsense, and 1 (0.72%) frame-shift mutations. However, only nonsynonymous SNPs were selected from coding region for this computational analysis.

3.2. Detection of Functional nsSNPs in Exonic Regions. The searching of functionally significant nsSNPs was done by predicting those which substitute the amino acids that are critical for *ADIPOR1* gene function. This computational study was accomplished and authenticated using different *in silico* tools, namely, SIFT, nsSNPAnalyzer, SNAP2, PolyPhen-2, SNPs&GO, Fathmm, SNPeffect, and I-Mutant 2.0.

3.2.1. Analysis of Phenotypic Impacts by SIFT. SIFT tools filtrated that a total of 13 variants (17.568%) were damaging (score of 0.00–0.04) and the remaining 61 variants (82.432%) became tolerated (score of 0.08–0.55). It was noted that, among 13 variants, 2 nsSNPs (rs765487840, rs775780092) were predicted as damaging with low confidence. Therefore, SIFT suggested that these 11 nsSNPs might disrupt both the protein function and structure. The detailed results are provided in supporting information (see Table S1 of the Supplementary Material available online at http://dx.doi.org/10.1155/2016/9142190).

3.2.2. Functionally Significant nsSNPs by nsSNPAnalyzer and SNAP2. The results obtained from the nsSNPAnalyzer (Table S2) predicted that a total of 27 nsSNPs (36.486%) might be disease causal. In contrast, 47 nsSNPs (63.513%) have no effect on protein function and, hence, are considered as neutral. In addition, the results from SNAP2 server (Table S2) indicated 19 variants (25.675%) as significant and the remaining nsSNPs (74.324%) as neutral. Among the three computational tools, the highest number of significant nsSNP (27 variants) was detected by the nsSNPAnalyzer. The results obtained from SIFT, nsSNPAnalyzer, and SNAP2 concluded that the 7 nsSNPs with rsIDs of rs764078304, rs765425383, rs752071352, rs759555652, rs764912508, rs200326086, and rs766267373 are found as significant among three servers and thereby the result has one step refined and validated (Table 1).

3.2.3. Simulation of Functional Consequences by PolyPhen-2. The results (Table S3) obtained from PolyPhen-2 server indicated that 12 (16.216%) out of 74 nsSNPs were predicted as probably damaging (score of 0.96–1.00; more confident prediction) and 12 (16.216%; less confident prediction) nsSNPs were ranked as possibly damaging (score of 0.531–0.874) as well. Meanwhile, 50 (67.567%) nsSNPs were also classified as benign (score of 0.411–0.000). The classification of SNPs on the basis of PolyPhen-2 scores permits us to assess the potential quantitative effect of SNPs on wild type protein. Moreover, 7 nsSNPs (rs765425383, rs752071352, rs759555652, rs200326086, rs772408783, rs766267373, and rs749789403), predicted as damaging by SIFT, are also found as damaging using PolyPhen-2. This result gives clear indication that there is a strong correlation exists between evolutionary based approaches SIFT and the structural based approach PolyPhen-2 tools.

3.2.4. Functional Characterization by PhD-SNP, PANTHER, SNPs&GO, and Fathmm. We performed PhD-SNP, PANTHER, SNPs&GO, and Fathmm analyses of human *ADIPOR1* nsSNPs in order to add another layer of refinement in nsSNPs characterization. The predicted results by these servers are shown in Tables S4 and S5.

The predictions gained form PhD-SNP server offer the fact that 29 nsSNPs cause disease with probability score greater than 0.5 and the remaining nsSNPs are marked as neutral. The number of disease causing variants has been decreased in case of the prediction of PANTHER and SNPs&GO tools. The disease causing variants predicted by PANTHER and SNPs&GO are 15 and 13 nsSNPs, respectively. The results from PANTHER server showed that 17 nsSNPs remain unclassified.

From Fathmm, the nsSNPs in amino acids positions 4 to 122 in human ADIPOR1 protein are found to be damaging with score of −3.97 to −4.20.

The efficacy of functional SNP prediction can be increased more reliably by integrating the results of SVM based approaches. By combining the predictions of SIFT, nsSNPAnalyzer, SNAP2, PolyPhen-2, PhD-SNP, PANTHER, SNPs&GO, and Fathmm, five nsSNPs (A348G, H341Y, R324L, L224F, and L143P) are found to br more deleterious and disease associated (Tables 1 and 2).

3.2.5. Functional Investigation by SNPeffect. Biological macromolecules including proteins undergo self-assembly into functional complex in a tightly regulated manner to conduct the defined function [51]. Failure of correct aggregation of proteins may result in some conditions including type 2 diabetes, Alzheimer's disease, and other neurological diseases [52]. The results from TANGO investigation presented that only two variants, namely, A348G (dTANGO score is −39.32) and R324L (dTANGO score is −1.01), were found to be not affected among 5 selected variants in the aggregation prone regions of ADIPOR1 protein. In addition, the aggregation tendency of the other two variants, H341Y (dTANGO score is 222.09) and L224F (dTANGO score is 51.62), was increased and only one variant (L143P) with dTANGO score of −230.61 was decreased. On the other hand, WALTZ analysis screened that H341Y mutant (dWALTZ score −241.53) was found to be decreased to protein amyloid forming propensity and the rest of mutants were not affected. LIMBO prediction revealed that no variants were detected to modify the chaperone binding sites for Hsp70 chaperones. In this study, we analyzed the variants by SNPeffect tools at 90% homology searching of protein structures. SNPeffect could

Table 1: Refined SNPs obtained from SIFT, SNAP2, and nsSNPAnalyzer based classifications system.

rsIDs	Amino acid change	SIFT		SNAP2		nsSNPAnalyzer
		Score	Prediction	Score	Prediction	Prediction
rs764078304	*G367R*	*0.04*	*Damaging*	*72*	*Effect*	*Disease*
rs139371614	G364S	0.79	Tolerated	−74	Neutral	Disease
rs765425383	*A348G*	*0*	*Damaging*	*34*	*Effect*	*Disease*
rs752071352	*H341Y*	*0*	*Damaging*	*80*	*Effect*	*Disease*
rs759555652	*R324L*	*0.02*	*Damaging*	*56*	*Effect*	*Disease*
rs778848411	V279G	0.02	Damaging	36	Effect	Neutral
rs759593783	G275A	0.24	Tolerated	45	Effect	Disease
rs764912508	*R264W*	*0.02*	*Damaging*	*49*	*Effect*	*Disease*
rs369530077	I251N	0.01	Damaging	19	Effect	Neutral
rs200326086	*L224F*	*0.01*	*Damaging*	*2*	*Effect*	*Disease*
rs772408783	L215V	0.02	Damaging	1	Effect	Neutral
rs756988796	R202W	0.08	Tolerated	−35	Neutral	Disease
rs772165061	V200L	0.23	Tolerated	8	Effect	Neutral
rs760115326	F173L	0.1	Tolerated	48	Effect	Disease
rs770463342	K170N	0.09	Tolerated	25	Effect	Neutral
rs767286210	L149F	0.32	Tolerated	−50	Neutral	Disease
rs780018580	F145L	0.75	Tolerated	−45	Neutral	Disease
rs766267373	*L143P*	*0.01*	*Damaging*	*75*	*Effect*	*Disease*
rs764226232	R122W	0.17	Tolerated	−36	Neutral	Disease
rs751626519	M118K	0.2	Tolerated	60	Effect	Disease
rs781585434	P116S	0.18	Tolerated	6	Effect	Neutral
rs749789403	D108N	0.02	Damaging	−10	Neutral	Neutral
rs141511034	P96L	0.33	Tolerated	−75	Neutral	Disease
rs769729230	E78K	0.15	Tolerated	31	Effect	Disease
rs751028180	G31E	1	Tolerated	−81	Neutral	Disease
rs149582032	A28T	0.61	Tolerated	−88	Neutral	Disease
rs780838176	E26Q	0.34	Tolerated	−89	Neutral	Disease
rs749145406	A15P	0.35	Tolerated	−34	Neutral	Disease
rs200868442	G14F	0.91	Tolerated	−4	Neutral	Disease
rs774465119	N13K	0.93	Tolerated	18	Effect	Disease
rs372656012	G12E	1	Tolerated	−63	Neutral	Disease
rs759643470	V9E	0.74	Tolerated	61	Effect	Disease
rs765487840	H4L	0.03	Damaging	−9	Neutral	Disease
rs775780092	H4Y	0.11	Damaging	−27	Neutral	Disease

not find any reliable structural information for protein to carry out a FoldX stability analysis. Detailed results for selected 5 variants are supplied in Table 3.

3.2.6. Protein Stability Changes Found by I-Mutant 2.0. The prediction of stability changes of selected 5 nsSNPs by I-Mutant 2.0 is given in Table 3. The results are predicted to be either increase or decrease of the free energy change upon amino acid substitutions. Four out of five selective mutants were found to be decreased in protein stability and the remaining one mutant (H341Y) was predicted as increased in protein stability with reliability index (RI) 3.

3.2.7. Visualization of Evolutionary Conserved Amino Acid Residues by ConSurf. ConSurf server is able to discriminate appropriately between the conservation caused by a short evolutionary time and genuine sequence conservation using Empirical Bayesian method. Our findings indicated that human *ADIPOR1* is highly conserved (Figure 2). The sequence alignment from different species revealed that residues A348 and H341 were located in highly conserved

Table 2: Refined SNPs obtained from PhD-SNP, PANTHER, SNPs&GO, Fathmm, and PolyPhen-2 based classification systems.

Amino acid change	PhD-SNP	PANTHER	SNPs&GO	Fathmm	PolyPhen-2
G367R	Disease	Neutral	Neutral	Tolerated	Benign
A348G	*Disease*	*Disease*	*Disease*	*Tolerated*	*Probably damaging*
H341Y	*Disease*	*Disease*	*Disease*	*Tolerated*	*Probably damaging*
R324L	*Disease*	*Disease*	*Disease*	*Tolerated*	*Probably damaging*
A307V	Disease	Neutral	Disease	Tolerated	Benign
T296R	Disease	Neutral	Neutral	Tolerated	Benign
V279G	Disease	Disease	Disease	Tolerated	Benign
G275A	Neutral	Disease	Disease	Tolerated	Probably damaging
V270M	Neutral	Disease	Neutral	Tolerated	Possibly damaging
R264W	Disease	Disease	Disease	Tolerated	Possibly damaging
I251N	Disease	Disease	Disease	Tolerated	Possibly damaging
S231P	Disease	Neutral	Neutral	Tolerated	Possibly damaging
L224F	*Disease*	*Neutral*	*Neutral*	*Tolerated*	*Probably damaging*
L251V	Disease	Neutral	Neutral	Tolerated	Probably damaging
R202W	Disease	Disease	Disease	Tolerated	Benign
F173L	Disease	Neutral	Neutral	Tolerated	Possibly damaging
K170N	Neutral	Disease	Neutral	Tolerated	Probably damaging
F145L	Disease	Neutral	Neutral	Tolerated	Benign
L143P	*Disease*	*Disease*	*Disease*	*Tolerated*	*Probably damaging*
R130C	Disease	Disease	Disease	Tolerated	Benign
R122W	Disease	Disease	Neutral	Damaging	Benign
M118K	Disease	Disease	Disease	Damaging	Benign
P116S	Neutral	Disease	Neutral	Tolerated	Probably damaging
D108N	Disease	Neutral	Neutral	Tolerated	Probably damaging
R91H	Disease	Neutral	Neutral	Tolerated	Benign
E89D	Disease	Neutral	Disease	Tolerated	Benign
E78K	Disease	Neutral	Neutral	Tolerated	Benign
P70S	Neutral	Neutral	Neutral	Tolerated	Probably damaging
R40P	Disease	Unknown	Neutral	Tolerated	Benign
G31E	Neutral	Unknown	Neutral	Tolerated	Probably damaging
A15P	Disease	Unknown	Neutral	Tolerated	Benign
G14F	Disease	Unknown	Neutral	Tolerated	Benign
N13K	Disease	Unknown	Neutral	Tolerated	Benign
G12E	Disease	Unknown	Neutral	Tolerated	Benign
V9E	Disease	Unknown	Neutral	Tolerated	Benign

regions and predicted to cause structural and functional impacts, respectively, on ADIPOR1 protein. On the other hand, the residues L224 and L143 had average conserved scores and the remaining one residue (R324) was located in conserved region of the protein.

3.3. Structural Analysis of Mutant Structures. The five predicted deleterious and disease causing variants were mapped to the PDB ID 3WXV native structure and substitution of amino acid residues was carried out using Swiss-PDB Viewer individually in order to generate five mutant modeled structures. After that, we calculated the total energy before and after energy minimization for both mutant model and wild type structures (Table 5). The values of total energy for five mutant modeled structures exhibit deviation from native structure considered before and after energy minimization. Five modeled structures (A348G, H341Y, R324L, L224F, and L143P) revealed an increase in energy (less favorable change) after energy minimization in comparing native structure. Among five screened mutations, H341Y showed the highest increase in energy which may be explained by the energetically unfavorable substitution of His to Tyr amino acids.

Table 3: A list of selected variants for analyzing SNPeffect and I-Mutant tools.

Amino acid change	SNPeffect				I-Mutant	
	TANGO Aggregation tendency (dTANGO score)	WALTZ Amyloid propensity (dWALTZ score)	LIMBO Chaperone binding tendency (dLIMBO score)	FoldX Protein stability	Prediction	RI
A348G	Not affected (−39.32)	Not affected (46.87)	Not affected (0.00)	NP	Decrease	8
H341Y	Increased (222.09)	Decreased (−241.53)	Not affected (0.00)	NP	Increase	3
R324L	Not affected (−1.01)	Not affected (0.49)	Not affected (0.00)	NP	Decrease	7
L224F	Increased (51.62)	Not affected (−46.88)	Not affected (0.00)	NP	Decrease	8
L143P	Decreased (−230.61)	Not affected (2.58)	Not affected (0.00)	NP	Decrease	4

Table 4: Surface accessibility of wild type and mutants of ADIPOR1 protein.

Amino acid change	Class assignment	Relative surface accessibility (RSA)	Absolute surface accessibility	Z-fit score for RSA prediction
A348*G*	Buried	0.094	10.403	−0.412
	Buried	*0.093*	*7.288*	*−0.405*
H341*Y*	Buried	0.070	12.769	0.848
	Buried	*0.063*	*13.463*	*0.712*
R324*L*	Exposed	0.34	77.860	−0.843
	Exposed	*0.352*	*64.470*	*−0.734*
L224*F*	Buried	0.029	5.273	−0.077
	Buried	*0.027*	*5.519*	*0.018*
L143*P*	Exposed	0.305	55.937	−0.824
	Exposed	*0.329*	*46.628*	*−0.942*

The zinc-binding domain is found in the intracellular layer of the membrane and zinc ion is coordinated by three His residues, His191, His337, and His341, of ADIPOR1 protein. In H341Y variants, His is replaced by Tyr. Due to the presence of aromatic amino acid Tyr in 341 position of ADIPOR1 protein there may be a good chance to disrupt zinc coordination. Adiponectin stimulated AMPK phosphorylation and UCP2 upregulation are mediated by zinc-binding domain [46].

The results from TM score are delivered in Table 6. TM score was utilized in order to evaluate the topological similarity of two protein structures and RMSD measured the average distance between the backbones of two superimposed proteins [53]. The TM score for five variants reveals that structurally there are no differences between native and mutant modeled structures. It might be concluded that mutants and wild type structures are matched perfectly. We also considered another parameter (RMSD) in order to predict the structural similarity between native and mutant structures of ADIPOR1 protein. The higher is the RMSD value, the more is the deviation between the two structures which in turn fluctuates their functional activities. It can be seen from Table 6 that the RMSD values between the native structure and the mutant modeled structures are all similar. By considering the above two values of TM score and RMSD, it could be suggested that these mutations do not bring a significant alteration in the mutant structures with regard to the native protein structure.

Nonbonding interactions such as H-bond has significant role in stabilizing the secondary structure of proteins [54]. Therefore, we have utilized the Swiss-PDB Viewer to visualize the hydrogen bonding pattern of five selected substituted amino acids with their surrounding amino acid residues in mutant proteins with regard to wild type (Figure 1). The hydrogen bonding pattern of variants A348G, L224F, and L143P has remained similar in comparison with wild type structure (PDB ID 3WXV). In variant H341Y, His341 indicates six hydrogen bonding interactions with Thr140, His141, Val344, Val345, His337, and Gln338, whereas mutant aromatic Tyr341 indicates five hydrogen bonding interactions. This has occurred due to the differences in the charge density and hydrophobicity between wild type and mutant residues. In variant R324L, one H-bond disappeared due to the substitution of Arg324 by Leu324. Additionally, A348G, H341Y, R324L, L224F, and L143P variants were analyzed for

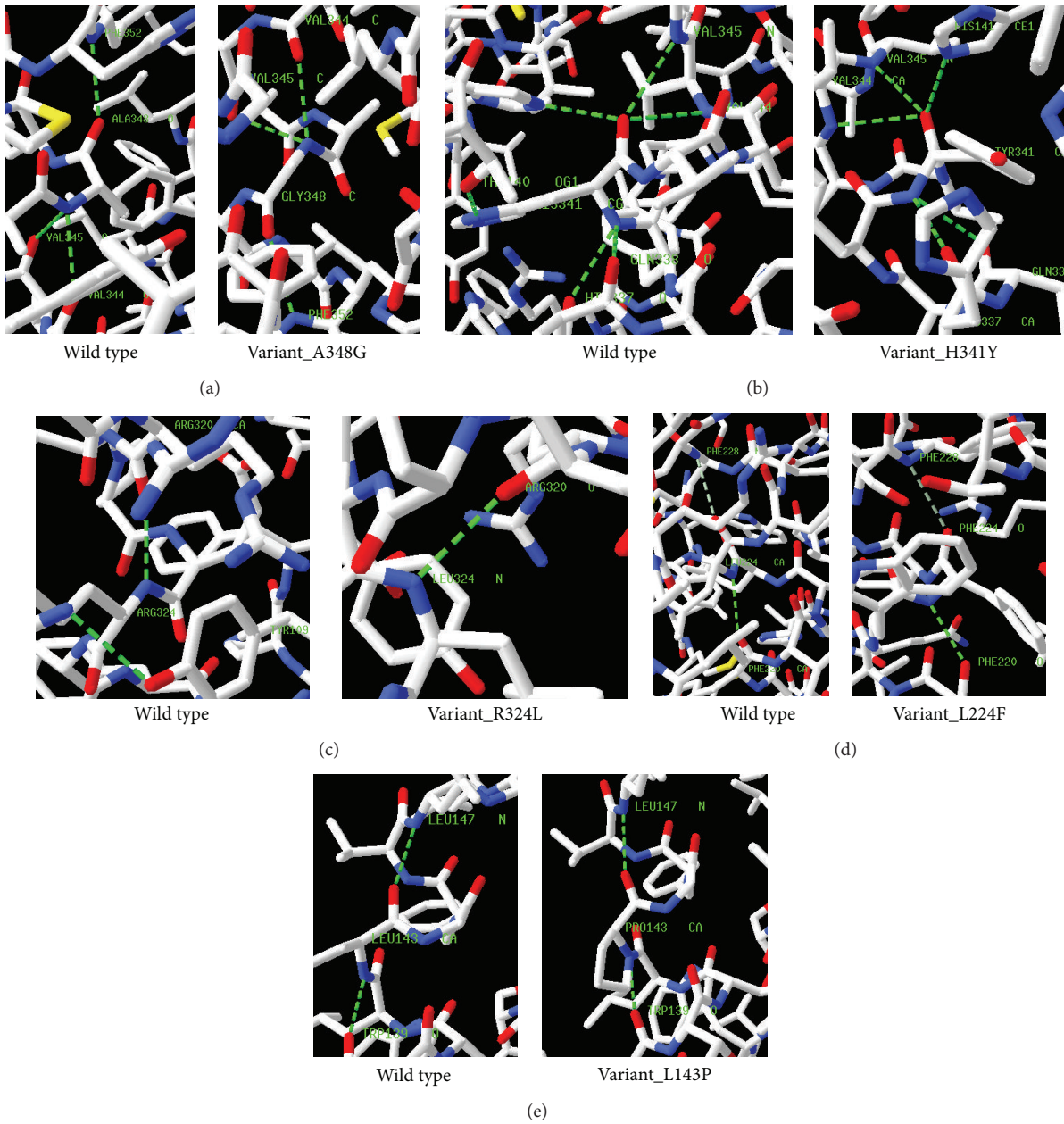


Figure 1: H-bond (green discontinuous line) and weak hydrogen bond (dark white discontinuous line) of wild type and mutant analogues with the adjacent amino acids residues. (a) At 348 position, three hydrogen bonds (H-bond) are observed with Val344, Val345, and Phe352 in both wild type (Ala348) and mutant (Gly348) structures. (b) His341 is visualized with six H-bonding interactions for Thr140, His141, His337, Gln338, Val344, and Val34, and one H-bond is abolished due to the replacement of mutant Tyr34 at the same position. (c) At 324 position, two H-bonds are observed with Tyr109 and Arg320 in native (Arg324) structure, but only one H-bond is found with Arg320 in mutant (Leu324) structure at the same position. (d) Phe228 is examined with single weak H-bonding in both native (Leu224) and substituted (Phe224) structures. In addition, single H-bond is also pictured with Phe220 in both structures. (e) Trp139 and Leu147 participated in forming two H-bonds at same position (143) in both mutant and wild type structures.

solvent accessibility and stability and significant changes in both parameters were seen for all five variants (Table 4).

3.4. Analysis of Ligand Binding Sites and Protein-Protein Interactions. FTSite identifies 3 ligand binding sites on ADIPOR1 protein (Figure 3). The amino acids found in these 3 sites of ADIPOR1 protein are given in Table 7. By the results of FTSite, it is observed that our 5 selected variants are not involved among these sites.

STRING database predicted the functional interaction pattern of ADIPOR1 protein to other proteins in a cell. Strong functional associations of ADIPOR1 protein have been observed with *ADIPOQ, APPL1, LEP,* and *INS* partners (Figure 4). Besides, weak interactions with less confidence

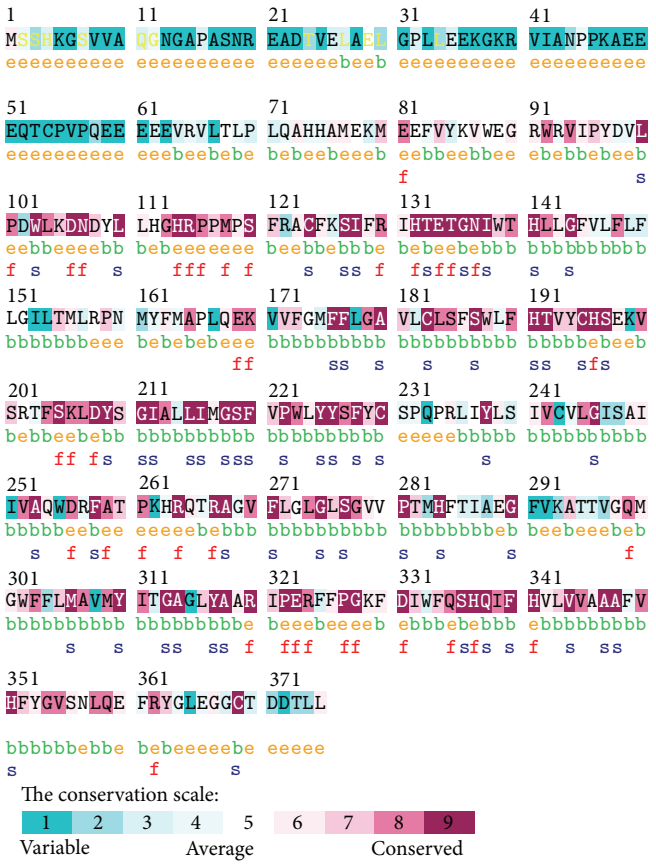


Figure 2: Unique and conserved amino acids in ADIPOR1 protein were predicted by ConSurf. Amino acids were ordered based on a conservation scale of 1–9 and highlighted as follows: blue residues (1–4) are variable, white residues (5) are average, and purple residues (6–9) are conserved. (e) Exposed residues are colored via an orange letter. (b) Buried residues are marked via a green letter. (f) Putative functional highly conserved and exposed residues are revealed with a red letter. (s) Predicted structural residues which are highly conserved and buried are indicated via blue letter.

Table 5: Total energy of native and mutant ADIPOR1 structures before and after energy minimization.

Amino acid variants	Total energy before energy minimization (kj/mol)	Total energy after energy minimization (kj/mol)
Native	−6555.888	−11406.533
A348G	−6468.591	−11319.678
H341Y	86183.258	−10520.564
R324L	−6078.778	−11107.321
L224F	372036.875	−11199.636
L143P	179867.000	−11088.141

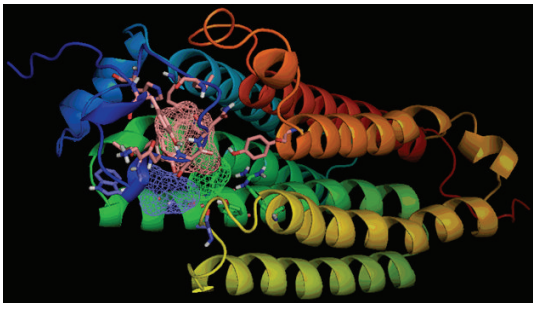
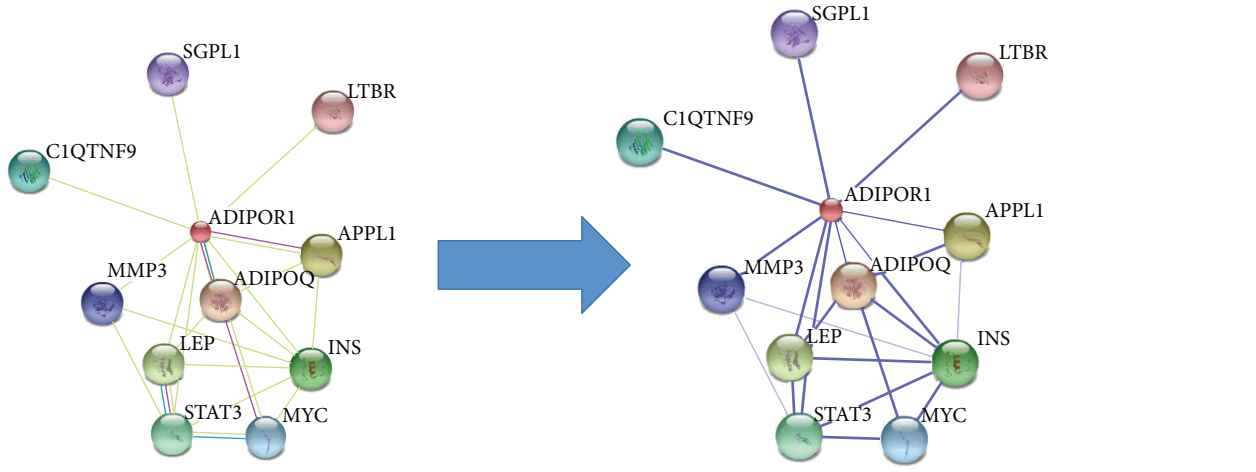


Figure 3: Binding of ligands in ADIPOR1 proteins ligand binding pocket 1-3 predicted by FTSite. No mutants were observed in binding site 1-3.



ADIPOR1 functional partners:

		Neighborhood	Gene fusion	Cooccurrence	Coexpression	Experiments	Databases	Text-mining	[Homology]	Score
ADIPOQ	Adiponectin, C1Q and collagen domain containing (244 aa)						●	●		0.994
APPL1	Adaptor protein, phosphotyrosine interaction, PH domain, and leucine zipper containing 1 (709 aa)					●		●		0.985
LEP	Leptin; part of a signaling pathway that acts to regulate the size of the body fat depot (167 aa)							●		0.942
INS	Insulin; insulin decreases blood glucose concentration (110 aa)							●		0.941
STAT3	Signal transducer and activator of transcription 3 (770 aa)							●		0.865
C1QTNF9	C1q and tumor necrosis factor related protein 9 (333 aa)							●		0.862
MYC	V-myc myelocytomatosis viral oncogene homolog (avian) (454 aa)							●		0.833
MMP3	Matrix metallopeptidase 3 (stromelysin 1, progelatinase) (477 aa)							●		0.829
SGPL1	Sphingosine-1-phosphate lyase 1 (568 aa)							●		0.806
LTBR	Lymphotoxin beta receptor (TNFR superfamily, member 3) (435 aa)							●		0.804

Figure 4: ADIPOR1 protein-protein interactions with 10 partners. One color is specified for each type of evidence in the predicted functional links (edges) among eight colored lines. (a) Only *APPL1* with score of 0.985 is observed for interaction with *ADIPOR1* by experimental and text-mining basis. From text-mining data, *ADIPOR1* interactions are detected for *ADIPOQ, APPL1, LEP, INS, STAT3, C1QTNF9, MYC, MMP3, SGPL1,* and *LTBR* proteins with 0.994, 0.985, 0.942, 0.941, 0.865, 0.862, 0.833, 0.828, 0.806, and 0.804 scores, respectively. (b) Strong association pattern (thick blue lines) of *ADIPOR1* is predicted for *ADIPOQ, APPL1, LEP,* and *INS* partners with high confidence. The remaining partners have weak association and shown in the form of thin blue lines.

Table 6: RMSD value and TM score of mutant modeled structures of ADIPOR1 protein.

Variants	RMSD value	TM score
A348G	0.00	1.00
H341Y	0.00	1.00
R324L	0.00	1.00
L224F	0.00	1.00
L143P	0.00	1.00

Table 7: Residues at ligand binding sites of ADIPOR1 protein.

Site 1	Site 2	Site 3
TYR A 97	PHE A 190	TRP A 103
LEU A 104	SER A 205	LEU A 104
LYS A 105	ASP A 208	ASP A 106
ASP A 106	TYR A 209	PHE A 190
ASN A 107	ARG A 267	TYR A 194
LEU A 110	TYR A 317	SER A 201
HIS A 114		SER A 205
GLU A 134		ALA A 259
PHE A 190		
HIS A 191		
TYR A 194		
TYR A 317		

have been observed for *STAT3, C1QTNF9, MYC, MMP3, SGPL1,* and *LTBR* proteins.

The associations between polymorphisms of *ADIPOR1* gene (such as rs12733285 and rs1342387) and metabolic diseases such as diabetes, obesity, and insulin resistance have been reported [19, 20, 22, 23]. However, no such study has established the association between damaging nsSNPs (rs765425383, A348G; rs752071352, H341Y; rs759555652, R324L; rs200326086, L224F; and rs766267373, L143P) and diseases. Hence, the confirmation of these nsSNPs in any disease is required to complement the existing limited body of knowledge. The combination of the analysis of human genetic variations of the *ADIPOR1* gene together with the

computational method to predict their possible functional impact may help in the analysis of *ADIPOR1* gene variants and establish their effects on protein functional characteristics. Specifically, this approach permits the estimation of the probability of amino acid changes which can be detrimental for ADIPOR1 protein functions.

4. Conclusion

In this comprehensive computational study, we have identified five deleterious mutations (A348G, H341Y, R324L, L224F, and L143P) among the coding region of *ADIPOR1* gene with the help of different bioinformatics tools. The variants were predicted to be similar to wild type ADIPOR1 protein structurally. However, decreased stability of mutant proteins has been observed with classical molecular dynamics study compared to wild type. Among the potential five nsSNPs, H341Y mutant has been found to cause considerable changes in amyloid forming propensity and aggregation tendency of ADIPOR1 protein. Additionally, there might be chance to disrupt the zinc coordination domain which is responsible for adiponectin stimulated MPK phosphorylation and UCP2 upregulation. The deleterious mutations of *ADIPOR1* should be further investigated to establish their roles in the pathogenesis of related diseases.

Competing Interests

The authors declare that there are no competing interests regarding this paper.

Acknowledgments

The authors would like to greatly acknowledge the technical support provided by the Department of Biochemistry and Molecular Biology, Jahangirnagar University, Savar, Dhaka, Bangladesh, during the study.

References

[1] J. M. Friedman, "Obesity in the new millennium," *Nature*, vol. 404, no. 6778, pp. 632–634, 2000.

[2] J. S. Flier, "Obesity wars: molecular progress confronts an expanding epidemic," *Cell*, vol. 116, no. 2, pp. 337–350, 2004.

[3] J. M. Olefsky and C. K. Glass, "Macrophages, inflammation, and insulin resistance," *Annual Review of Physiology*, vol. 72, no. 1, pp. 219–246, 2010.

[4] D. LeRoith and D. Accili, "Mechanisms of disease: using genetically altered mice to study concepts of type 2 diabetes," *Nature Clinical Practice Endocrinology and Metabolism*, vol. 4, no. 3, pp. 164–172, 2008.

[5] S. Gesta, Y.-H. Tseng, and C. R. Kahn, "Developmental origin of fat: tracking obesity to its source," *Cell*, vol. 131, no. 2, pp. 242–256, 2007.

[6] J. Breitfeld, M. Stumvoll, and P. Kovacs, "Genetics of adiponectin," *Biochimie*, vol. 94, no. 10, pp. 2157–2163, 2012.

[7] J. Bełtowski, "Adiponectin and resistin—new hormones of white adipose tissue," *Medical Science Monitor*, vol. 9, no. 2, pp. RA55–RA61, 2003.

[8] J. Chen, B. Tan, E. Karteris et al., "Secretion of adiponectin by human placenta: differential modulation of adiponectin and its receptors by cytokines," *Diabetologia*, vol. 49, no. 6, pp. 1292–1302, 2006.

[9] C. Hug and H. F. Lodish, "The role of the adipocyte hormone adiponectin in cardiovascular disease," *Current Opinion in Pharmacology*, vol. 5, no. 2, pp. 129–134, 2005.

[10] Y. Matsuzawa, "The metabolic syndrome and adipocytokines," *FEBS Letters*, vol. 580, no. 12, pp. 2917–2921, 2006.

[11] P. E. Scherer, "Adipose tissue: from lipid storage compartment to endocrine organ," *Diabetes*, vol. 55, no. 6, pp. 1537–1545, 2006.

[12] T. Kadowaki, T. Yamauchi, N. Kubota, K. Hara, K. Ueki, and K. Tobe, "Adiponectin and adiponectin receptors in insulin resistance, diabetes, and the metabolic syndrome," *Journal of Clinical Investigation*, vol. 116, no. 7, pp. 1784–1792, 2006.

[13] T. Yamauchi, J. Kamon, Y. Minokoshi et al., "Adiponectin stimulates glucose utilization and fatty-acid oxidation by activating AMP-activated protein kinase," *Nature Medicine*, vol. 8, no. 11, pp. 1288–1295, 2002.

[14] E. Tomas, T.-S. Tsao, A. K. Saha et al., "Enhanced muscle fat oxidation and glucose transport by ACRP30 globular domain: acetyl-CoA carboxylase inhibition and AMP-activated protein kinase activation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 25, pp. 16309–16313, 2002.

[15] X. Fang and G. Sweeney, "Mechanisms regulating energy metabolism by adiponectin in obesity and diabetes," *Biochemical Society Transactions*, vol. 34, no. 5, pp. 798–801, 2006.

[16] C. Hug, J. Wang, N. S. Ahmad, J. S. Bogan, T.-S. Tsao, and H. F. Lodish, "T-cadherin is a receptor for hexameric and high-molecular-weight forms of Acrp30/adiponectin," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 28, pp. 10308–10313, 2004.

[17] T. Yamauchi, J. Kamon, Y. Ito et al., "Cloning of adiponectin receptors that mediate antidiabetic metabolic effects," *Nature*, vol. 423, no. 6941, pp. 762–769, 2003.

[18] T. Yamauchi, Y. Nio, T. Maki et al., "Targeted disruption of AdipoR1 and AdipoR2 causes abrogation of adiponectin binding and metabolic actions," *Nature Medicine*, vol. 13, no. 3, pp. 332–339, 2007.

[19] K. J. Mather, C. A. Christophi, K. A. Jablonski et al., "Common variants in genes encoding adiponectin (ADIPOQ) and its receptors (ADIPOR1/2), adiponectin concentrations, and diabetes incidence in the Diabetes Prevention Program," *Diabetic Medicine*, vol. 29, no. 12, pp. 1579–1588, 2012.

[20] K. E. Peters, J. Beilby, G. Cadby et al., "A comprehensive investigation of variants in genes encoding adiponectin (ADIPOQ) and its receptors (ADIPOR1/R2), and their association with serum adiponectin, type 2 diabetes, insulin resistance and the metabolic syndrome," *BMC Medical Genetics*, vol. 14, article 15, 2013.

[21] V. Kaklamani, N. Yi, K. Zhang et al., "Polymorphisms of ADIPOQ and ADIPOR1 and prostate cancer risk," *Metabolism: Clinical and Experimental*, vol. 60, no. 9, pp. 1234–1243, 2011.

[22] N. A. Crimmins and L. J. Martin, "Polymorphisms in adiponectin receptor genes ADIPOR1 and ADIPOR2 and insulin resistance," *Obesity Reviews*, vol. 8, no. 5, pp. 419–423, 2007.

[23] N. Stefan, F. Machicao, H. Staiger et al., "Polymorphisms in the gene encoding adiponectin receptor 1 are associated with insulin resistance and high liver fat," *Diabetologia*, vol. 48, no. 11, pp. 2282–2291, 2005.

[24] S. C. Collins, J. Luan, A. J. Thompson et al., "Adiponectin receptor genes: mutation screening in syndromes of insulin resistance and association studies for type 2 diabetes and metabolic traits in UK populations," *Diabetologia*, vol. 50, no. 3, pp. 555–562, 2007.

[25] M. Vaxillaire, A. Dechaume, V. Vasseur-Delannoy et al., "Genetic analysis of ADIPOR1 and ADIPOR2 candidate polymorphisms for type 2 diabetes in the Caucasian population," *Diabetes*, vol. 55, no. 3, pp. 856–861, 2006.

[26] S. Sunyaev, V. Ramensky, I. Koch, W. Lathe III, A. S. Kondrashov, and P. Bork, "Prediction of deleterious human alleles," *Human Molecular Genetics*, vol. 10, no. 6, pp. 591–597, 2001.

[27] Z. Wang and J. Moult, "SNPs, protein structure, and disease," *Human Mutation*, vol. 17, no. 4, pp. 263–270, 2001.

[28] R. Grantham, "Amino acid difference formula to help explain protein evolution," *Science*, vol. 185, no. 4154, pp. 862–864, 1974.

[29] D. Chasman and R. M. Adams, "Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation," *Journal of Molecular Biology*, vol. 307, no. 2, pp. 683–706, 2001.

[30] C. Baynes, C. S. Healey, K. A. Pooley et al., "Common variants in the ATM, BRCA1, BRCA2, CHEK2 and TP53 cancer susceptibility genes are unlikely to increase breast cancer risk," *Breast Cancer Research*, vol. 9, no. 2, article R27, 2007.

[31] S. Paul, M. Solayman, M. Saha, and M. S. Hossain, "In silico analysis of the functional and structural impacts of non-synonymous single nucleotide polymorphisms in the human paraxonase 1 gene," *International Journal Bioautomation*, vol. 19, no. 3, pp. 275–286, 2015.

[32] S. T. Sherry, M.-H. Ward, M. Kholodov et al., "DbSNP: the NCBI database of genetic variation," *Nucleic Acids Research*, vol. 29, no. 1, pp. 308–311, 2001.

[33] P. Kumar, S. Henikoff, and P. C. Ng, "Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm," *Nature Protocols*, vol. 4, no. 7, pp. 1073–1082, 2009.

[34] L. Bao, M. Zhou, and Y. Cui, "nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms," *Nucleic Acids Research*, vol. 33, no. 2, pp. W480–W482, 2005.

[35] Y. Bromberg, G. Yachdav, and B. Rost, "SNAP predicts effect of mutations on protein function," *Bioinformatics*, vol. 24, no. 20, pp. 2397–2398, 2008.

[36] I. A. Adzhubei, S. Schmidt, L. Peshkin et al., "A method and server for predicting damaging missense mutations," *Nature Methods*, vol. 7, no. 4, pp. 248–249, 2010.

[37] R. Calabrese, E. Capriotti, P. Fariselli, P. L. Martelli, and R. Casadio, "Functional annotations improve the predictive score of human disease-related mutations in proteins," *Human Mutation*, vol. 30, no. 8, pp. 1237–1244, 2009.

[38] E. Capriotti, R. Calabrese, and R. Casadio, "Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information," *Bioinformatics*, vol. 22, no. 22, pp. 2729–2734, 2006.

[39] D. J. Moore, L. Zhang, T. M. Dawson, and V. L. Dawson, "A missense mutation (L166P) in DJ-1, linked to familial Parkinson's disease, confers reduced protein stability and impairs homo-oligomerization," *Journal of Neurochemistry*, vol. 87, no. 6, pp. 1558–1567, 2003.

[40] H. A. Shihab, J. Gough, D. N. Cooper et al., "Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models," *Human Mutation*, vol. 34, no. 1, pp. 57–65, 2013.

[41] G. De Baets, J. Van Durme, J. Reumers et al., "SNPeffect 4.0: online prediction of molecular and structural effects of protein-coding variants," *Nucleic Acids Research*, vol. 40, no. 1, pp. D935–D939, 2012.

[42] E. Capriotti, P. Fariselli, and R. Casadio, "I-mutant2.0: predicting stability changes upon mutation from the protein sequence or structure," *Nucleic Acids Research*, vol. 33, no. 2, pp. W306–W310, 2005.

[43] G. Celniker, G. Nimrod, H. Ashkenazy et al., "ConSurf: using evolutionary data to raise testable hypotheses about protein function," *Israel Journal of Chemistry*, vol. 53, no. 3-4, pp. 199–206, 2013.

[44] B. Petersen, T. N. Petersen, P. Andersen, M. Nielsen, and C. Lundegaard, "A generic method for assignment of reliability scores applied to solvent accessibility predictions," *BMC Structural Biology*, vol. 9, no. 1, article 51, 2009.

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

[46] H. Tanabe, Y. Fujii, M. Okada-Iwabu et al., "Crystal structures of the human adiponectin receptors," *Nature*, vol. 520, no. 7547, pp. 312–316, 2015.

[47] M. U. Johansson, V. Zoete, O. Michielin, and N. Guex, "Defining and searching for structural motifs using DeepView/Swiss-PdbViewer," *BMC Bioinformatics*, vol. 13, no. 1, article 173, 2012.

[48] Y. Zhang and J. Skolnick, "TM-align: a protein structure alignment algorithm based on the TM-score," *Nucleic Acids Research*, vol. 33, no. 7, pp. 2302–2309, 2005.

[49] C.-H. Ngan, D. R. Hall, B. Zerbe, L. E. Grove, D. Kozakov, and S. Vajda, "FTSite: high accuracy detection of ligand binding sites on unbound protein structures," *Bioinformatics*, vol. 28, no. 2, pp. 286–287, 2012.

[50] D. Szklarczyk, A. Franceschini, M. Kuhn et al., "The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored," *Nucleic Acids Research*, vol. 39, no. 1, pp. D561–D568, 2011.

[51] R. J. Ellis, "Macromolecular crowding: obvious but underappreciated," *Trends in Biochemical Sciences*, vol. 26, no. 10, pp. 597–604, 2001.

[52] C. M. Dobson, "The structural basis of protein folding and its links with human disease," *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, vol. 356, no. 1406, pp. 133–145, 2001.

[53] J. C. Jimenez-Lopez, E. W. Gachomo, M. J. Seufferheld, and S. O. Kotchoni, "The maize ALDH protein superfamily: linking structural features to functional specificities," *BMC Structural Biology*, vol. 10, article 43, 2010.

[54] J. A. Ippolito, R. S. Alexander, and D. W. Christianson, "Hydrogen bond stereochemistry in protein structure and function," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 457–471, 1990.

Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 4814702, 6 pages
http://dx.doi.org/10.1155/2016/4814702

Research Article

Differential Proteomics Analysis of Colonic Tissues in Patients of Slow Transit Constipation

Songlin Wan, Weicheng Liu, Cuiping Tian, Xianghai Ren, Zhao Ding, Qun Qian, Congqing Jiang, and Yunhua Wu

Zhongnan Hospital of Wuhan University, Department of Colorectal & Anal Surgery, Clinical Center of Intestinal and Colorectal Diseases of Hubei Province, Key Laboratory of Intestinal & Colorectal Diseases of Hubei Province, Wuhan University, No. 169, Donghu Road, Wuchang District, Wuhan, Hubei 430071, China

Correspondence should be addressed to Congqing Jiang; wb002554@whu.edu.cn and Yunhua Wu; wb002564@whu.edu.cn

Received 2 March 2016; Accepted 11 April 2016

Academic Editor: Terry K. Smith

Objective. To investigate and screen the different expression of proteins in STC and normal group with a comparative proteomic approach. *Methods*. Two-dimensional electrophoresis was applied to separate the proteins in specimens from both 5 STC patients and 5 normal controls. The proteins with statistically significant differential expression between two groups were identified by computer aided image analysis and matrix assisted laser desorption ionization tandem time of flight mass spectrometry (MALDI-TOF-MS). *Results*. A total of 239 protein spots were identified in the average gel of the normal control and 215 in patients with STC. A total of 197 protein spots were matched and the mean matching rate was 82%. There were 14 protein spots which were expressed with statistically significant differences from others. Of those 14 protein spots, the expression of 12 spots increased markedly, while that of 2 spots decreased significantly. *Conclusion*. The proteomics expression in colonic specimens of STC patients is statistically significantly different from that of normal control, which may be associated with the pathogenesis of STC.

1. Introduction

Constipation is a very common functional gastrointestinal disorder that affects patients' quality of life [1, 2]. The prevalence of constipation in the worldwide general population is very variable, ranging from 2% to 35% in adults and ranging from 0.7% to 29.6% in children [3]. Slow transit constipation (STC) is common type of chronic constipation, which affects 13–37% of patients who have chronic, treatment-resistant constipation [4, 5]. The epidemiological data show that there is a higher incidence of STC in young females than in males [6, 7].

STC is characterized by slow proximal colonic transit and prolongation of transit time through the colon, which can be demonstrated with radiopaque marker transit tests [8, 9]. Despite its high prevalence, its etiology and precise mechanism(s) remain unknown [4, 10], and STC have become the main subject of many colorectal and anal surgeries. Although several morphological changes have been reported, the exact mechanism(s) of STC pathogenesis remains poorly understood and there are no effective clinical treatments. Even if some patients' clinical symptoms may improve after conservative management, the small subset of STC patients who do not respond to conservative management are considered candidates for surgery [11].

At present, quantitative proteomic analysis has become an important approach to screen the key protein of the pathogenesis of diseases. "Proteome" first referred to the total protein complement encoded by a given genome. Now it comprises any isoforms, posttranslational modifications, interactions, and actually everything "postgenomic" [12, 13]. The research about proteomics involves large scale detection, identification, and characterization of proteins, which is highly promising for biomarker detection over many diseases [14, 15]. The most common method applied is a combination of two-dimensional electrophoresis (2DE) and mass spectrometry.

In our study, we separate protein in test group and control group through two-dimensional electrophoresis (2DE), ImageMaster 2D Elite computer aided image analysis, and

matrix assisted laser desorption ionization tandem time of flight mass spectrometry (MALDI-TOF-MS), to find out the key protein acting in the pathogenesis of STC.

2. Materials and Methods

2.1. General Information. Patients were divided into a test group and a control group; each group included 5 cases. All patients had undergone a rigorous selection. First, all candidates should have prior screening through repeated gastrointestinal transit time (GITT) study, using 20 radiopaque markers, and abdominal imaging tests performed on 6 h, days 1, 2, and 3. And the abdominal imaging test performed on 6 h after taking 20 radiopaque markers was used to identify whether there was slow transit in the terminal intestine. All these five STC patients showed slow transit mainly on the left colon through GITT study and all samples were collected from the left hemicolon, and the same regions were available for both STC and control groups. Second, autoimmune diseases such as diabetes mellitus type I, autoimmune enteric leiomyositis, rheumatoid arthritis, systemic sclerosis, and systemic lupus erythematosus can prolong colonic transit time by damage of colonic smooth muscle cells. And all candidates with any of these autoimmune diseases were excluded and all these five STC patients selected were without any autoimmune diseases at all. Other evaluations including colonoscopy, barium enema, defecography, and anorectal manometry were used to rule out colorectal neoplasms, megacolon, outlet obstruction, and pelvic floor dysfunction. The conventional medical therapy had failed in all STC patients. Conservative treatment did not improve their bowel habits, and defecation frequently had to be achieved by the application of stimulus laxatives and enemas. All STC patients ($n = 5$, mean age 50 years, range 43–81 years; 5 women and 0 men) underwent subtotal colectomy with antiperistaltic cecoproctostomy or total colectomy with ileorectal anastomosis. Patients in the control group ($n = 5$, mean age 52 years, range 41–79 years; 5 women and 0 men) underwent partial colectomy for incomplete intestinal obstruction caused by colorectal neoplasms and the control tissue was taken from the proximal end. All control patients were reported with normal bowel habits. All specimens were obtained immediately after resection and snap-frozen in liquid nitrogen until being used.

The high-abundance proteins were removed with the steps recommended by Calbiochem Company and the improved Bradford method was used for sample protein quantitation [16].

2.2. Two-Dimensional Electrophoresis (2D)

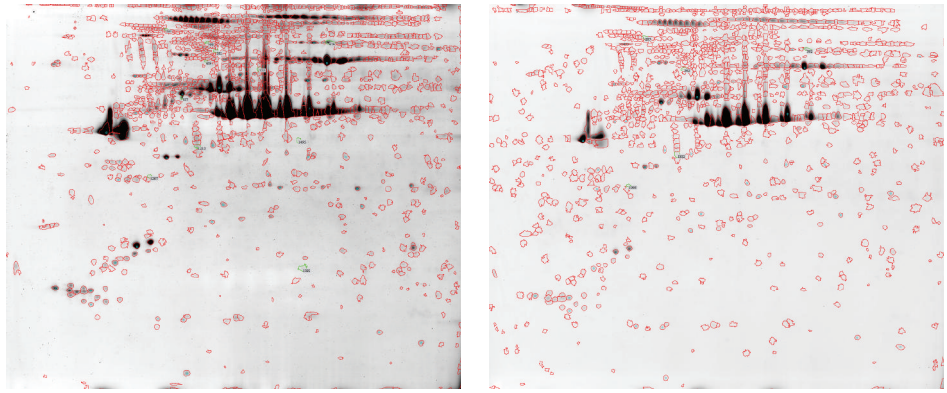
2.2.1. Extraction and Determination of the Concentration of Protein. A 2.0 mL sterile EP tube was prepared. About 200 mg small piece of tissue was rapidly clipped from the specimen and weighed on electronic scale, and then the tissue was placed in the precooling Petri dish with normal saline (6 mL, 0.9% NaCl). The tissue was shredded in the dish and then transferred to 1.5 mL EP tube. 800 μL cold protein lysate (containing a protease inhibitor) was added into the EP tube which was to be quickly transferred to an ultrasonic homogenizer. The EP tube was coated with ice. The ultrasonic homogenizer was triggered 5 times, 3 seconds each time. The tube was statically placed for 30 minutes; the supernatant was collected by centrifugation (4000 r/min 4°C, 60 min) and then stored in the −80°C refrigerator.

2.2.2. Protein Purification. The protein sample was transferred to a labeled EP tube (2 mL). After adding protein cracking fluid with 3 times volume of samples, the mixture was shock-cracked and subsequently placed on ice for 15 min. Then, the mixture was centrifuged for 8 min with a high speed of 12000 r/min at 10°C. The supernatant was discarded and the precipitation was resuspended and was shock-cracked for 30 s every 10 min for 3 cycles. Finally, the precipitation was resuspended again and used for next steps.

2.2.3. Concentration Determination. Protein concentrations were determined by the instruction of the Bio-Rad RCDC kit manual. The specific steps were as follows: standard concentration and gradient BSA reagent were prepared and the concentrations were, respectively, 0, 0.25, 0.5, 0.75, 1, 1.25, and 1.5 (g/mL). Reagent A working fluid resulted from the mix of 20 μL DC Reagent S and 1 mL DC Reagent A. Eight 1.5 mL EP tubes were, respectively, added with 25 μL standard BCA solution and 5-fold dilution of the sample, and then in each tube 125 μL RC Reagent I and RC Reagent II were added. The mix was collected by centrifugation (15000 r/min, 5 min) after the supernatant was discarded. 127 μL Reagent A was added to the tubes. The tubes were statically placed for 5 minutes after blending. Then, 1 mL DC Reagent B was added to each tube and blended. Absorbance values (750 nm) were measured after statically placed at room temperature for 15 min.

2.2.4. Isoelectrofocusing (IEF). Rehydration stock solution with IPG buffer was dissolved at room temperature (DTT and Bio-Lyte were added before using). The proper amount of rehydration stock solution with IPG buffer was added to samples till a final volume of 480 μL according to the concentration of the samples. The sample was linearly added along with the extension of the focusing plate. The protective layer on the surface of the gel was stripped with tweezers so as to distinguish positive and negative electrodes, and then the gel was placed downside onto the sample solution in the focusing plate. 2 mL mineral oil was covered by each gel. The mineral oil was slowly added onto the support film drop by drop. IEF was carried out at 50 V (14 h), 250 V linear (30 min), 1000 V fast (60 min), 10000 V linear (1 h), 10000 V fast 80000 V, and 500 V (arbitrary time).

2.2.5. SDS-PAGE Electrophoresis. Two 10% acrylamide gels were prepared. 1 cm was reserved on the upper side and covered with water-saturated n-butanol and then polymerized for 30 min. The n-butanol was then discarded. The gels were washed three times with ultrapure water. Water on the glass plate was sucked with filter paper. The gels were taken to the hydration plates, and 6 mL equilibrium liquid I (0.2 g DTT per 10 mL equilibrium mother liquid before using) and II



(a) (b)

Figure 1: Comparison of two-dimensional (2D) gel electrophoresis results using colonic tissues between STC and control groups. STC group (a) and control group (b) 2D gel electrophoresis.

(0.25 g iodoacetamide per 10 mL equilibrium mother liquid before using) were then added into each plate. The plate was placed on a shaker (15 min). The strip was carefully placed on the surface of the gel. Air bubble between them was removed. The heated sealant was cooled to 37°C and then applied to the surface of gel. The gel was transferred to the electrophoresis tank after the sealant was solidified. The upper electrophoresis tank was filled with electrophoresis solution after checking the positive and negative charges. The electrophoresis was first carried out at Low Voltage (5 mA/gel) within 2 cm downside the sample, and then the electrophoresis was carried out to the bottom of the sample with a current of 30 mA/gel.

2.2.6. Coomassie Blue Staining. Gels were dyed with Coomassie Blue for 16 h after being fixed with 12% TCA for 2 h. Then, the gels were washed with 0.1 mol/L Tris-H_3PO_4 (pH 6.5) for 2 min and then with 25% alcohol for 1 min at most. At last, the gels were fixed with protein-dye compounds again.

2.2.7. Gel analysis. After scanned by Image Scanner II transmission scanner and marked with Scan scanning software, the gels were analyzed by image analysis software (ImageMaster 2D Elite5.0). And plots with expression differences more than 10 times ($P < 0.05$) were chosen as targeted proteins.

2.3. Mass Spectrum Identification

2.3.1. Enzymatic Hydrolysis. Targeted parts of the gels were removed and washed with ultrapure water and then decolorized with 200 μL mixed liquid (25 mmol/LNH_4HCO_3, 50% ACN) for 20 min at 37°C. An alternative procedure was ultrasonic decoloring. Then, 100 μL ACN was added in the mixture and discarded when the color of the gels turns white. Subsequently, the gels were enzymatically hydrolyzed with Trypsin (diluted with 25 mmol/LNH_4HCO_3 at a concentration of 12.5 mg/mL) for 45 min at 4°C. Finally, gels were continuously hydrolyzed with 25 mmol/LNH_4HCO_3 10 μL at 37°C for 16 h.

2.3.2. Mass Spectrometry Analysis. Substrate (4 mg HCCA in 1 mL solution of (ACN: 0.1% TFA (70 : 30))) was prepared; the enzymatically hydrolyzed gels were dried and covered with 0.1 μL substrate. Then the mixture was diluted 25 times its original volume and salts were removed using 0.1% TFA.

2.3.3. Database Analysis. Data from mass spectrometry were analyzed and screened with Swiss-Prot Database in Mascot.

2.4. Statistical Analysis. Statistical analyses were performed using SPSS 19.0 software for Windows (SPSS, Chicago, IL, USA). All groups of variables were tested for normal distribution using the Kolmogorov-Smirnov test and normally distributed data were expressed as mean ± SD, followed by a two-tailed Student's t test to determine P values. All results with $P < 0.05$ were considered statistically significant.

All research involving human participants was approved by the Zhongnan Hospital of Wuhan University Ethics Committee, and we obtained written informed consent from all participants before they were enrolled in the study (ethical considerations: ethics number: 2010010; ethical approval starting date: March 8, 2010; ethical approval expiration date: March 8, 2013).

3. Results

(1) Two-dimensional electrophoresis in proteomics research find that there were 14 differential protein spots between the two groups and the expression difference were more than 10 times (Figure 1).

(2) Mass spectrum (MS) identification and database search: after the protein spots with more than 10-fold difference between two groups were removed for the gels, they were subsequently dealt with using the following procedures including enzyme digestion, extraction of the peptides, and sample desalination. Finally, peptide mass fingerprint (PMF) and MS/MS spectrum were achieved through MALDI-TOF-MS analysis of proteins.

The retrieval for mass spectrometric data in Swiss-Prot successfully identified anterior 14 spots, wherein the

Table 1: Elevated expression proteins in colonic tissues of STC patients compared with control group.

Name	Score	pI	Formula weight	Fold change
ACTG	82	5.31	42108	17.91
ENPL	80	4.76	92696	29.30
F102B	80	6.62	39911	11.90
GBB1	76	5.60	38151	15.98
GELS	80	5.90	86043	13.72
K1C9	63	5.14	62255	20.01
MTMR7	62	5.94	76754	15.92
PDIA1	91	4.76	57480	22.71
PDIA6	68	4.95	48490	30.79
PFD6	78	8.83	14574	10.21
PRVA	60	4.98	12051	15.31
TRFE	79	6.81	79294	17.31

Table 2: Decreased expression proteins in colonic tissues of STC patients compared with control group.

Name	Score	pI	Formula weight	Fold change
MRP4	99	8.41	150344	−32.86
VIME	81	5.06	53676	−12.39

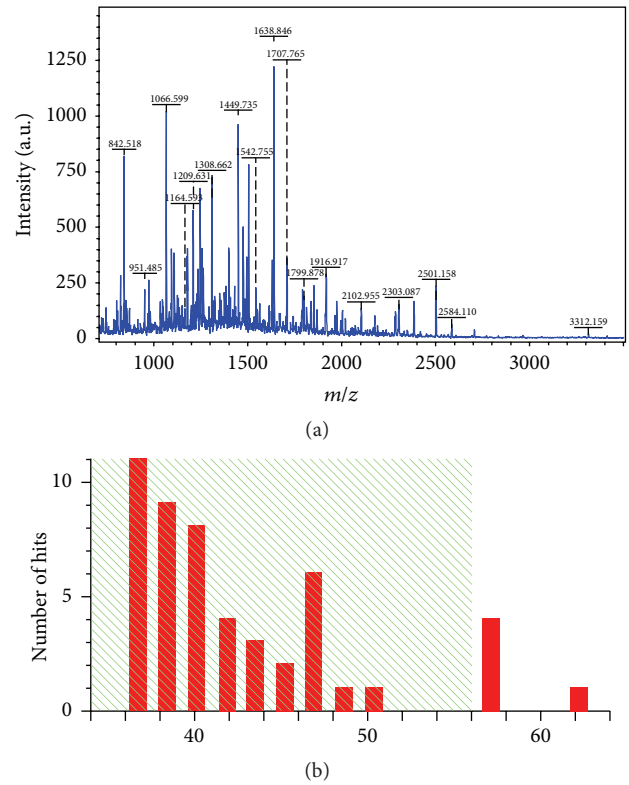


Figure 2: Mass spectrum identification figure of protein MRP4 and its matched analysis diagram. Mass spectrum identification figure of protein MRP4 (a) and its matched analysis diagram (b).

expression of 12 spots is up and that of 2 spots is down in STC group (Tables 1 and 2). Our further studies show that multidrug resistance-associated protein 4 (MRP4) is of clinical significance (Figure 2). MRP4 is playing a critical role in the process of maintaining the stability of smooth muscle cells (SMC) [16, 17]. The downregulation of MRP4 expression in colon tissue of STC patients may be related to the decrease of SMC and may play a role in the pathogenesis of STC by affecting colon transmission.

4. Discussion

STC pathogenesis and mechanism not yet completely expounded that, at present clinical STC, patients mainly depend on purgative and gastrointestinal prokinetic agents for conservative management. A few clinical studies have confirmed that pharmacotherapy is effective in treating STC, but most patients with STC to accept drug treatment do not quite approve of the effect. The effectiveness of current STC drugs is fairly limited. The abuse of laxatives to assist defecation of patients with STC in clinical is very common; some patients who do not respond to conservative management are considered candidates for surgery [11].

Proteomics has been used to study the specific protein group that is functioning on different space and time and the study included three dimensions: expression proteomics, functional proteomics, and structural proteomics. Through the former three dimensions of study, researchers explore the function mechanism, regulatory mechanism, and interaction mechanism within group of corresponding proteins at the level of protein, which provide a theoretical reference for clinical diagnosis, pathologic study, drug screening, and pathogenesis of disease. STC currently faces many challenges, ranging from the elucidation of its pathophysiology to the effective treatment in patients. Proteomics has been widely employed in many diseases in the search of biomarkers, particularly cancer proteins. It has great potential to improve both our understanding and clinical management of STC. In the field of surgical research, proteomics studies mainly focus on tumor-related diseases at present; the proteomics studies for functional diseases are quite few and the proteomics study for STC is not reported so far. In our proteomics studies, we find that, compared with control group, the expression of 12 protein spots increased markedly and that of 2 protein spots decreased significantly in STC group. Among the low expression of proteins of STC, multidrug resistance-associated protein 4 (MRP4) is of clinical significance.

MRP4 belongs to transmembrane protein family; its main function is to transport intracellular cyclic nucleotides that include cyclic adenosine monophosphate and cyclic guanosine monophosphate. Some studies have confirmed that MRP4 participated in cellular excretion of resveratrol 3-O-glucuronide and resveratrol 4′-O-glucuronide [18], the inhibition of urinary excretion of methotrexate [19], cellular excretion of the raloxifene sulfates in breast cancer patients [20], the process of human obstructive cholestasis [21], and so on. Recently studies have revealed that MRP4 can be involved in maintaining the homeostasis of smooth muscle cell (SMC) by the regulation of intracellular cyclic nucleotide levels on the intracellular cyclic nucleotide signaling pathways (PKC,

PKA, etc.). When the level of intracellular MRP4 is decreasing or inhibited, the significant reduction of the stability of SMC results in the decrease of proliferation cells and increase of apoptosis, which may be involved in the pathogenesis of idiopathic pulmonary arterial hypertension (IPAH) and coronary heart disease (CHD) [16, 17]. Our previous studies about STC have shown that, compared with normal colonic tissue, the number of SMC decreases and the apoptosis of SMC increases in colonic tissue with STC, which may be involved in the pathogenesis of STC. In this study, the homeostasis of SMC in colonic specimens of STC patients is changed and the decrease of proliferation cells and increase of apoptosis may result from the reduction of MRP4 expression, which may play a role in the pathogenesis of STC.

The results suggest that the different expression of proteins between STC and normal group may be associated with the pathogenesis of STC. The pathogenesis mechanism of STC remains largely obscure, which may result from multifactor; the change of the expression of MRP4 may play a certain role. Further studies for the relationship between MRP4 and the pathogenesis mechanism of STC and targeted drugs to the change of MRP4 are required.

Competing Interests

The authors declare that they have no competing interests.

Authors' Contributions

Songlin Wan and Weicheng Liu contributed equally to this work.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (nos. 81500505 and 81570492); the Young and Middle-Aged Medical Talent Support Program of the Health and Family Planning Commission of Wuhan Municipality; the Natural Science Foundation of Hubei Province of China (nos. 2015CFB636 and 2011CDB521); Central Universities and Scientific Research Foundation of Wuhan University (no. 303410500160); the Natural Science Foundation of Zhongnan Hospital of Wuhan University (no. 22#).

References

[1] D. Polymeros, I. Beintaris, A. Gaglia et al., "Partially hydrolyzed guar gum accelerates colonic transit time and improves symptoms in adults with chronic constipation," *Digestive Diseases and Sciences*, vol. 59, no. 9, pp. 2207–2214, 2014.

[2] L. Leung, T. Riutta, J. Kotecha, and W. Rosser, "Chronic constipation: an evidence-based review," *Journal of the American Board of Family Medicine*, vol. 24, no. 4, pp. 436–451, 2011.

[3] S. M. Mugie, M. A. Benninga, and C. Di Lorenzo, "Epidemiology of constipation in children and adults: a systematic review," *Best Practice and Research: Clinical Gastroenterology*, vol. 25, no. 1, pp. 3–18, 2011.

[4] A. Lembo and M. Camilleri, "Current concepts: chronic constipation," *The New England Journal of Medicine*, vol. 349, no. 14, pp. 1360–1368, 2003.

[5] S. Singh, S. Heady, E. Coss-Adame, and S. S. C. Rao, "Clinical utility of colonic manometry in slow transit constipation," *Neurogastroenterology and Motility*, vol. 25, no. 6, pp. 487–495, 2013.

[6] C. H. Knowles, S. M. Scott, C. Rayner et al., "Idiopathic slow-transit constipation: an almost exclusively female disorder," *Diseases of the Colon and Rectum*, vol. 46, no. 12, pp. 1716–1717, 2003.

[7] D. M. Preston and J. E. Lennard-Jones, "Severe chronic constipation of young women: 'idiopathic slow transit constipation'," *Gut*, vol. 27, no. 1, pp. 41–48, 1986.

[8] M. Camilleri, W. Grant Thompson, J. W. Fleshman, and J. H. Pemberton, "Clinical management of intractable constipation," *Annals of Internal Medicine*, vol. 121, no. 7, pp. 520–528, 1994.

[9] M. Bouchoucha, G. Devroede, P. Arhan et al., "What is the meaning of colorectal transit time measurement?" *Diseases of the Colon & Rectum*, vol. 35, no. 8, pp. 773–782, 1992.

[10] A. Emmanuel, M. Cools, L. Vandeplassche, and R. Kerstens, "Prucalopride improves bowel function and colonic transit time in patients with chronic constipation: an integrated analysis," *The American Journal of Gastroenterology*, vol. 109, no. 6, pp. 887–894, 2014.

[11] M. A. Levitt, K. L. Mathis, and J. H. Pemberton, "Surgical treatment for constipation in children and adults," *Best Practice and Research: Clinical Gastroenterology*, vol. 25, no. 1, pp. 167–179, 2011.

[12] V. C. Wasinger, S. J. Cordwell, A. Cerpa-Poljak et al., "Progress with gene-product mapping of the Mollicutes: *Mycoplasma genitalium*," *Electrophoresis*, vol. 16, no. 7, pp. 1090–1094, 1995.

[13] A. J. Ferreri, G. Illerhaus, E. Zucca, F. Cavalli, and International Extranodal Lymphoma Study Group, "Flows and flaws in primary central nervous system lymphoma," *Nature Reviews Clinical Oncology*, vol. 7, no. 8, 2010.

[14] A. Vaiopoulou, M. Gazouli, G. Theodoropoulos, and G. Zografos, "Current advantages in the application of proteomics in inflammatory bowel disease," *Digestive Diseases and Sciences*, vol. 57, no. 11, pp. 2755–2764, 2012.

[15] P. P. Chan, V. C. Wasinger, and R. W. Leong, "Current application of proteomics in biomarker discovery for inflammatory bowel disease," *World Journal of Gastrointestinal Pathophysiology*, vol. 7, no. 1, pp. 27–37, 2016.

[16] Y. Hara, Y. Sassi, C. Guibert et al., "Inhibition of MRP4 prevents and reverses pulmonary hypertension in mice," *The Journal of Clinical Investigation*, vol. 121, no. 7, pp. 2888–2897, 2011.

[17] Y. Sassi, L. Lipskaia, G. Vandecasteele et al., "Multidrug resistance-associated protein 4 regulates cAMP-dependent signaling pathways and controls human and rat SMC proliferation," *The Journal of Clinical Investigation*, vol. 118, no. 8, pp. 2747–2757, 2008.

[18] S. Wang, F. Li, E. Quan, D. Dong, and B. Wu, "Efflux transport characterization of resveratrol glucuronides in UDP-glucuronosyltransferase 1A1 transfected hela cells: application of a cellular pharmacokinetic model to decipher the contribution of multidrug resistance-associated protein 4," *Drug Metabolism and Disposition*, vol. 44, no. 4, pp. 485–488, 2016.

[19] A. Kawase, T. Yamamoto, S. Egashira, and M. Iwaki, "Stereoselective inhibition of methotrexate excretion by glucuronides of nonsteroidal anti-inflammatory drugs via multidrug resistance

proteins 2 and 4," *The Journal of Pharmacology and Experimental Therapeutics*, vol. 356, no. 2, pp. 366–374, 2016.

[20] X. Zhou, S. Wang, H. Sun, and B. Wu, "Sulfonation of raloxifene in HEK293 cells overexpressing SULT1A3: involvement of breast cancer resistance protein (BCRP/ABCG2) and multidrug resistance-associated protein 4 (MRP4/ABCC4) in excretion of sulfate metabolites," *Drug Metabolism and Pharmacokinetics*, vol. 30, no. 6, pp. 425–433, 2015.

[21] W. Lian, X. Liu, L. Yang et al., "The role of TNFalpha in promoting hepatic MRP4 expression via the p38-Rb-E2F1 pathway in human obstructive cholestasis," *Biochemical and Biophysical Research Communications*, 2016.

Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 7237053, 7 pages
http://dx.doi.org/10.1155/2016/7237053

Research Article

A Comprehensive Curation Shows the Dynamic Evolutionary Patterns of Prokaryotic CRISPRs

Guoqin Mai,[1] Ruiquan Ge,[1,2] Guoquan Sun,[1] Qinghan Meng,[1,2] and Fengfeng Zhou[3,4]

[1] *Shenzhen Institutes of Advanced Technology and Key Lab for Health Informatics, Chinese Academy of Sciences, Shenzhen, Guangdong 518055, China*
[2] *Shenzhen College of Advanced Technology, University of Chinese Academy of Sciences, Shenzhen, Guangdong 518055, China*
[3] *College of Computer Science and Technology, Jilin University, Changchun, Jilin 130012, China*
[4] *Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, Jilin 130012, China*

Correspondence should be addressed to Fengfeng Zhou; fengfengzhou@gmail.com

Received 24 January 2016; Revised 24 March 2016; Accepted 28 March 2016

Academic Editor: Hongwei Wang

Motivation. Clustered regularly interspaced short palindromic repeat (CRISPR) is a genetic element with active regulation roles for foreign invasive genes in the prokaryotic genomes and has been engineered to work with the CRISPR-associated sequence (Cas) gene Cas9 as one of the modern genome editing technologies. Due to inconsistent definitions, the existing CRISPR detection programs seem to have missed some weak CRISPR signals. *Results.* This study manually curates all the currently annotated CRISPR elements in the prokaryotic genomes and proposes 95 updates to the annotations. A new definition is proposed to cover all the CRISPRs. The comprehensive comparison of CRISPR numbers on the taxonomic levels of both domains and genus shows high variations for closely related species even in the same genus. The detailed investigation of how CRISPRs are evolutionarily manipulated in the 8 completely sequenced species in the genus *Thermoanaerobacter* demonstrates that transposons act as a frequent tool for splitting long CRISPRs into shorter ones along a long evolutionary history.

1. Introduction

A CRISPR is an array of repeat copies (DR, direct repeat) connected by fixed-length linker sequences [1]. The linker sequences are called spacers and are usually acquired from the genetic elements invading the host microbial cells [2]. A CRISPR may be activated by its neighboring CRISPR-associated (Cas) genes, and the spacers will be processed into RNA molecular. The RNA form of spacers will repress the activities of foreign elements with reverse-complementary regions that reinvade the host cells [1–3]. Although CRISPRs are only detected in microbial genomes in the nature, it has been engineered as one of the major genomic editing technologies for both animal and plant genomes [4, 5]. So it is essential to study CRISPR's evolutionary dynamic patterns.

Only a few computational tools were released to automatically detect CRISPRs from a given genome, but they have different default parameter settings for a CRISPR. PILER-CR [6] screens for a repeat array using a local genomic self-alignment and has $O(L^3)$ for the complexities of both time and memory space, where L is the genome length. PILER-CR requires the DR length to be between 20 and 40 bps. CRT [7] starts with the scanning for local repetitive k-mers, which is a nucleotide sequence with length k. Due to its nature of local scanning, CRT runs for linear time and within linear memory space. Its default setting for DR lengths is between 21 and 37 bps. The latest tool CRISPRFinder [8] uses an existing tool Vmatch to find the DR array in a given genome and will discard the tandem repeats as false positives. CRISPRFinder has a slightly longer assumption for DRs between 20 and 47 bps. A comprehensive database DbCRISPR was also published to provide the CRISPR annotations for 2,762 microbial genomes [9].

Due to the different default settings of existing tools for a CRISPR structure, we hypothesize that a comprehensive manual curation may refine the current CRISPR annotations

and facilitate the discovery of CRISPR evolutionary mechanisms. This study proposed 95 updated CRISPR annotations, the majorities (59/95~62.11%) of which are transposon-broken CRISPRs. A new CRISPR definition is proposed and all the curated data are available as easy-to-use FASTA/GFF3 formats. The CRISPR variations within all the prokaryotic genus are summarized based on the curated annotations, and the dynamic patterns of CRISPRs in the genus *Thermoanaerobacter* are investigated in detail.

2. Material and Methods

2.1. Initial CRISPR Annotations. The complete annotation of CRISPRs in microbial genomes was downloaded from the latest version of the database DbCRISPR [9], which was updated on April 14, 2014. 4,065 CRISPRs are annotated in the 2,762 genomes of bacteria and archaea. The questionable structures in DbCRISPR are omitted. If a genome harbors CRISPRs, it has 3.11 CRISPRs on average.

SpacerDB consists of all the annotated CRISPR spacer sequences in the database DbCRISPR and was downloaded from the website of DbCRISPR [9] on April 14, 2014. CRISPR spacers are not random nucleotide sequences and are supposed to originate from the foreign invasive elements like phages [2]. So a DR flanking sequence matching a known spacer may also be a spacer.

2.2. Analysis Techniques and Tools. 2,762 genomes of bacteria and archaea are identified CRISPR by running CRT, CRISPRFinder, and PILER-CR, respectively. However, lots of CRISPR results are not common on these three software programs. Thus, the CRISPRs in the DbCRISPR are considered the gold standard. The ratios, which are produced by the number of the spacers in CRT, CRISPRFinder, or PILER-CR results to the number of the spacers in the DbCRISPR, are statistically analyzed. If the ratio is greater than 1 or less than 1, the corresponding CRISPRs are manually analyzing, checking, modifying, and correcting the above CRISPR results based on the database DbCRISPR. A comprehensive manual curation was conducted to screen for candidate DRs in the CRISPR flanking regions. For an annotated CRISPR, the homologous copies of DRs were screened by the local copy of NCBI BLAST version 2.2.25 [10]. NCBI BLAST is also used to screen the homologous matches of a given spacer sequence.

A CRISPR is usually activated by the closest CRISPR-associated (Cas) genes [11], and multiple CRISPRs may share the same group of Cas genes, if there is only one such group neighboring to these CRISPRs.

3. Results and Discussion

In summary, this study conducts a comprehensive curation of the current CRISPR annotation and proposes three types of revisions based on the observations that some annotated CRISPRs (1) have undetected DRs in the flanking regions, (2) are broken into two CRISPRs due to the nonstandard DRs or transposons in between, or (3) are annotated as two CRISPRs at the beginning of circular chromosomes. The following sections elaborate in detail the three types of annotation errors and demonstrate some interesting observations.

3.1. Detection of New DRs and Spacers. A CRISPR is a few copies of a short repeat (DR, direct repeat) gapped by unique linking sequences (spacer) [1], as shown in Figure S1 in Supplementary Material available online at http://dx.doi.org/10.1155/2016/7237053. But different computational annotation programs do not agree on the default values for the lengths of DRs and spacers. The CRISPR annotation database DbCRISPR restricts a DR within 21 to 47 bps in length and a spacer within 25 to 60 bps in length [9]. The same group of authors released a CRISPR detection program, CRISPRFinder [8], which has a different requirement for the DR length (within 23 to 55 bps). And CRISPRFinder also requires the length of a spacer to be within 0.6 to 2.5 times of the DR length. Two other existing programs CRT [7] and PILER-CR [6] require the DR lengths to be within 19 to 38 bps and 16 to 64 bps, respectively.

Lots of CRISPR results are not common on these three software programs CRISPRFinder, CRT, and PILER-CR. Based on the database DbCRISPR, we made novel discoveries by manually analyzing, modifying, and correcting the above CRISPR results and investigated the lengths of CRISPR DRs and spacers. After the corrections of CRISPR annotations in the following sections, we will give a revised CRISPR definition.

A few DRs were not detected in the flanking regions of CRISPRs, as demonstrated in Figures 1(a) and 1(b). Six CRISPRs may have one missing DR in the flanking region, as in Figure 1(a). For example, by screening for more DRs in the CRISPR flanking regions, we propose 10 spacers for the CRISPR NC_010125_2181482_2182111 in *Gluconacetobacter diazotrophicus* PAl 5, as in Figure S2. But DbCRISPR only detected 9 spacers for this CRISPR. The new DR is also confirmed using the tool CRISPRFinder [8]. Four other CRISPRs (NC_010125_62935_64899, NC_010125_2253748_2255112, NC_011365_388303_388536, and NC_011365_460172_461964) in the same bacterial strain *Gluconacetobacter diazotrophicus* PAl 5 missed one complete DR in their flanking regions too, as in Figure S2.

Two DRs were added to each of 4 CRISPRs, as in Figure 1(b). These DRs were missed by the database DbCRISPR mainly due to the fact that one of the two DRs is only partially identical to the other DRs, as in Figure S3. One of the example CRISPRs is NC_014152_2078344_2080300 in the bacterial genome *Thermincola* sp. JR, with 26 spacers. We propose two more DRs for this CRISPR, although one of the two new DRs is identical to the other DRs in half of its region. The mismatched region may be introduced by the gene conversion [12] or homologous recombination [13] mechanism. Another piece of supporting evidence for the two new spacers comes from their BLAST matches to two known spacers in the other genomes in the SpacerDB with 91.3% and 94.4% in matching identity percentages, respectively. A spacer is supposed to originate from the foreign invasive elements. Since it is low in probability to have such almost identical sequences just by the random single nucleotide

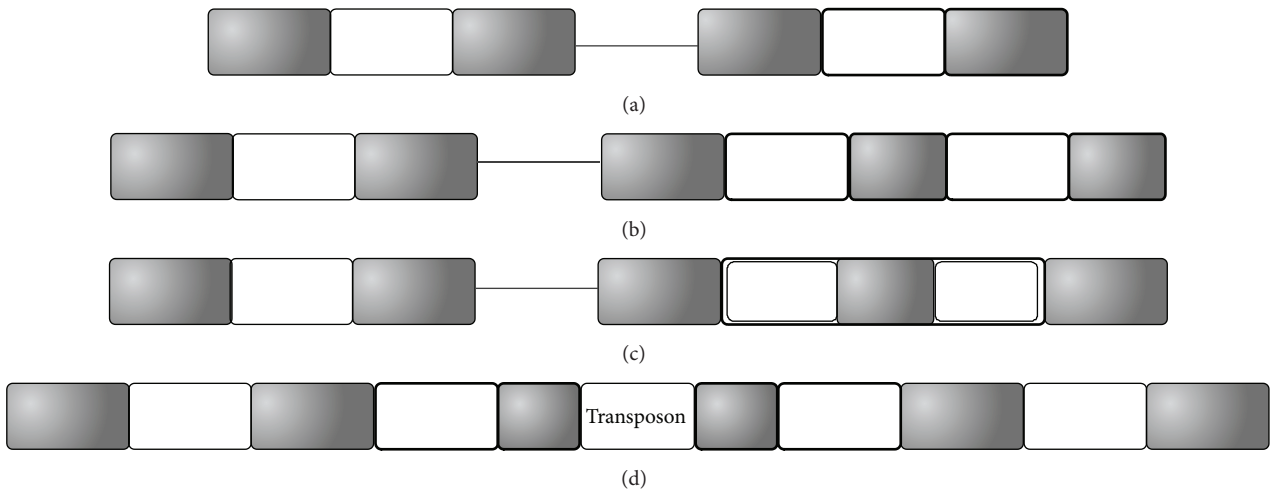


Figure 1: More DRs remain to be detected in the database DbCRISPR. A CRISPR annotation may miss (a) one complete DR or (b) two imperfect DRs in the flanking regions. (c) Some CRISPR spacers may harbor truncated DRs, which are partially identical to the hosting CRISPR's DRs. And (d) many CRISPRs have DRs broken by transposon insertions. DRs and spacers are represented by gray and white boxes, respectively. The new regions added to the DbCRISPR annotations are boxed in bold.

mutations, the two new candidate spacers are suggested to be real spacers originated from the same foreign invasive elements as the two homologous copies in the other genomes. Three more CRISPRs, that is, NC_015865_1907425_1908328 in *Thermococcus* sp. 4557, NC_015738_2085666_2087297 in *Eggerthella* sp. YY7918, and NC_014209_791663_793738 in *Thermoanaerobacter mathranii* subsp. *mathranii* str. A3, are expanded with two more DRs for the same reason, as in Figure S3.

Each of 3 CRISPRs has an extraordinarily long spacer with a truncated DR inside, as demonstrated in Figure 1(c). The representative example is the CRISPR NC_019693_6234891_6235861 with 12 spacers in the cyanobacterial strain *Oscillatoria acuminata* PCC 6304. Figure 2(a) illustrates that this CRISPR's ninth spacer harbors a partial DR copy with 70% length of the other DRs. And the two flanking sequences in this long spacer have reasonable lengths as spacers. So we propose that this CRISPR has 13 spacers, as in Figure 2(a). Similar cases are detected in two other CRISPRs, that is, NC_008639_1625359_1633049 in *Chlorobium phaeobacteroides* DSM 266 and NC_007777_3904715_3905896 in *Frankia* sp. CcI3, as in Figure S4.

Quite a number of CRISPRs acquired transposon insertions and were broken into two CRISPRs in the DbCRISPR annotations, as in Figure 2(b). All the 59 cases are demonstrated in Figure S5. Our curation shows that there are 59 CRISPRs with flanking DRs inserted by transposons, for example, insertion sequence (IS) elements [14, 15] or miniature inverted-repeat transposable elements (MITEs) [16]. Figure 2(b) illustrates that the 1,221 bp IS element is inserted into the DR sequence of the CRISPR in the genome *Thermoanaerobacter italicus* Ab9. The 4 bp tandem duplication ATAG in the DR sequence also supports that this IS copy was recently translocated here. A 180 bp MITE element is also observed to be within the DR sequence of a CRISPR in *Microcystis aeruginosa* NIES-843, and the 5 bp tandem duplication CTATT flanking the MITE should be produced during its recent translocation, as in Figure S5. Summary of all the 59 transposon insertions in CRISPRs may be found in Figure S5.

Some DRs were not detected in the database DbCRISPR, so that a long CRISPR may be annotated as two neighboring ones with almost identical DRs. 4 CRISPRs have a full DR copy that were not detected in the database DbCRISPR. The representative example is found in the Deltaproteobacteria *Myxococcus fulvus* HW-1. DbCRISPR annotates two neighboring CRISPRs NC_015711_2680594_2682129 and NC_015711_2682223_2687985, with 21 and 80 spacers, respectively. These two CRISPRs have the same DR sequence (GTCGCTCCCCGTGAACGCGGGGAGCGTGGGTTG-AAAC) and a 94 bp gap in between, as demonstrated in Figure 2(c). But there is an identical DR copy in the 94 bp gap, which is not detected in the database DbCRISPR due to an unknown reason. The sequence between this DR copy and the annotated CRISPR NC_015711_2680594_2682129 identically matches three spacers in the same genome. A CRISPR spacer is supposed to be acquired from foreign invasive elements [1], and the data suggests that the microbial defense system CRISPR has generated four spacers to respond to this foreign element. The shared Cas genes further support that NC_015711_2680594_2682129 and NC_015711_2682223_2687985 may be joined by the 94 bp gap as one longer CRISPR. Three other similar cases were detected in the bacteria *Caldicellulosiruptor obsidiansis* OB47, *Thermosipho africanus* TCF52B, and *Herpetosiphon aurantiacus* ATCC 23779, as shown in Figure S6. These DRs were missed mainly due to their short lengths slightly below the threshold of spacer/DR $\in [0.6, 2.5]$. 11 other CRISPRs were broken mainly due to an internal partial DR copy that was

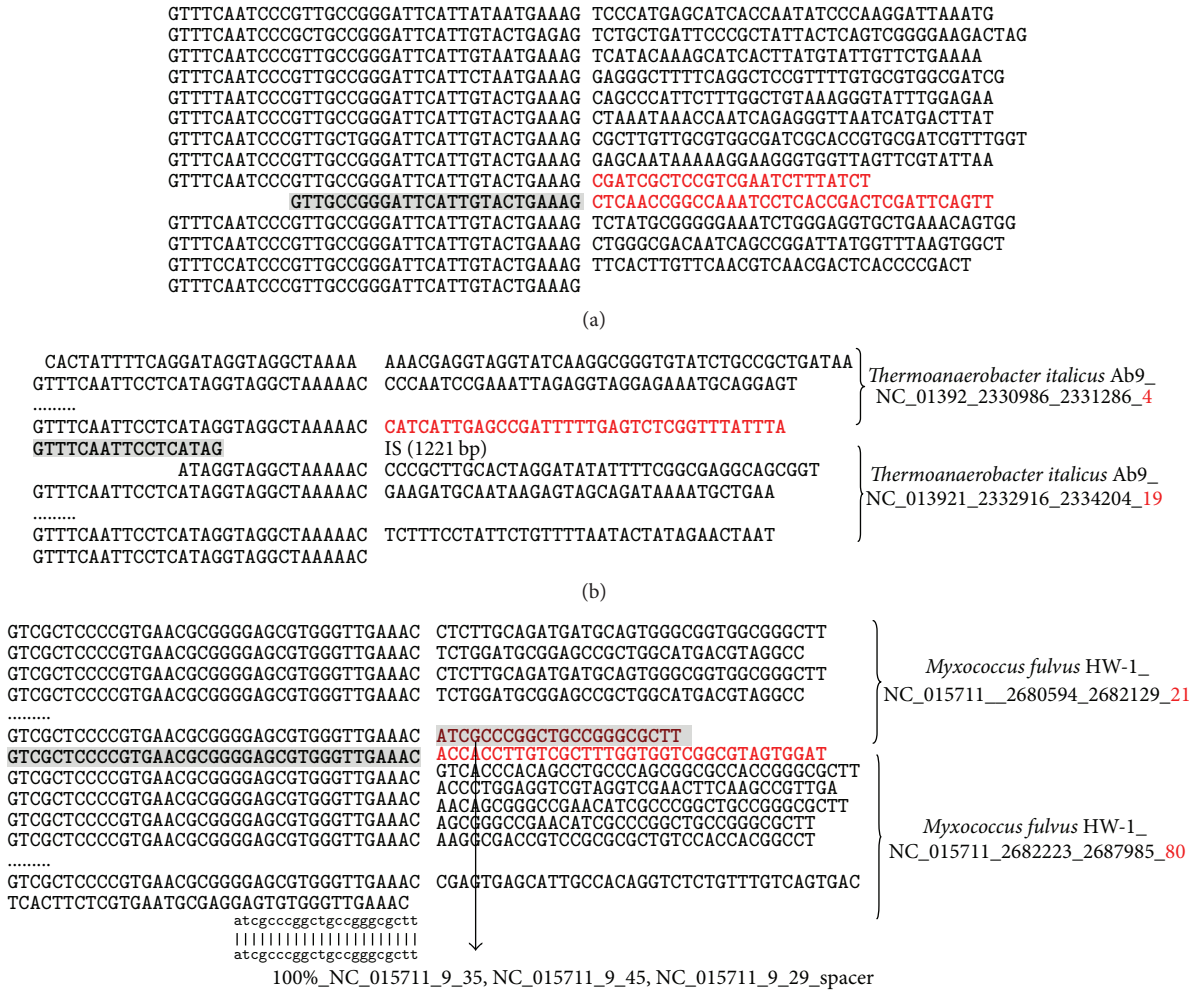


Figure 2: Detection of new DRs and spacers. (a) A CRISPR in *Oscillatoria acuminata* PCC 6304 has a long spacer with a truncated DR inside. The long spacer is in red and the truncated DR is in shade. DRs and spacers are represented in the left and right columns, respectively. (b) A CRISPR with DRs inserted by transposons. The newly annotated spacer regions are in red, and the new DRs are in shade. The last number in the CRISPR name is the DR copy number. (c) A CRISPR with undetected DRs inside. The two neighboring CRISPRs have almost identical DRs and one undetected DR in between. The undetected DR may be a full copy. The added region is highlighted in bold. The regions of new DRs matching the existing DRs are in shade. The regions of new spacers matching the spacers in other genomes are in shade. The last number in the CRISPR name is the DR copy number. (d) A CRISPR is broken into two at the beginning of a circular chromosome of *Thermococcus litoralis* DSM 5473. One more spacer is proposed to combine the two CRISPRs into one longer CRISPR. The added region (spacer) is highlighted in red.

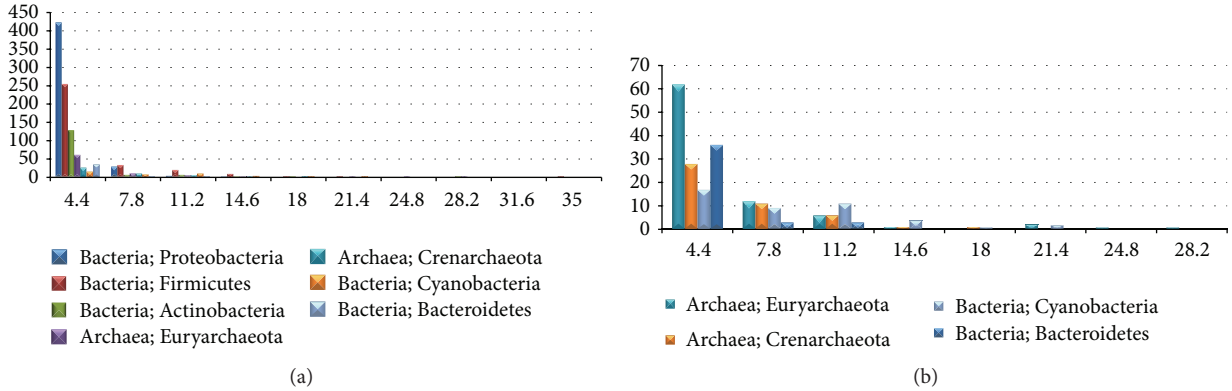


Figure 3: Histograms of the genome numbers (in vertical axis) versus the CRISPR numbers per genome (in horizontal axis). The summaries are conducted for the taxonomic domains of (a) at least 30 genomes and (b) 30–100 genomes, each of which has at least one CRISPR.

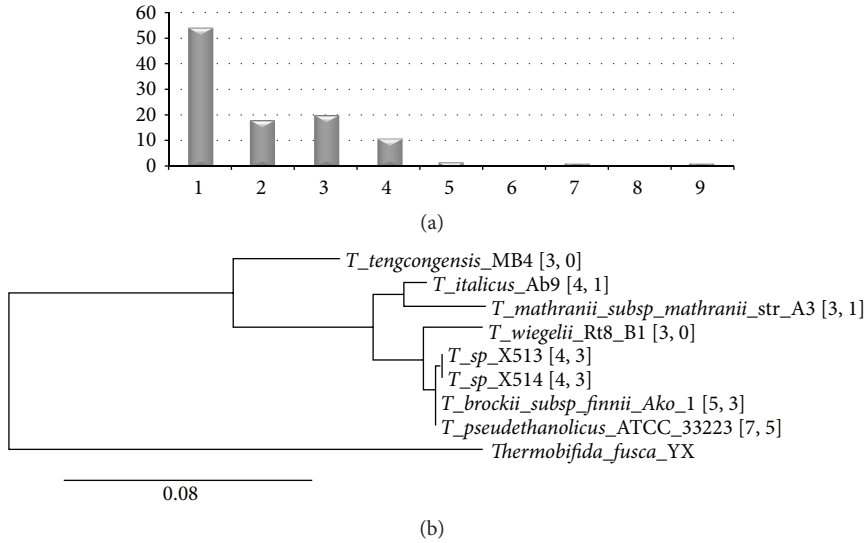


Figure 4: CRISPR distribution dynamics in the prokaryotic genomes. (a) Histogram of the standard deviation (StdEv) values of the 107 genera, each of which has at least three genomes in the CRISPR annotations. (b) The 16S rRNA phylogenetic tree of the 8 completely sequenced *Thermoanaerobacter* genomes, rooted at a closely related species *Thermobifida fusca YX*. The two numbers in the brackets are the numbers of annotated CRISPRs and CRISPRs with transposon insertions, respectively.

not detected by the database DbCRISPR, as shown in Figure S6.

Unlike the eukaryotic counterparts, most of the microbial chromosomes are in the circular shape [17], but the database DbCRISPR regards a CRISPR spanning the beginning point of a circular chromosome as two. We manually checked the 4,065 CRISPRs annotated in the database DbCRISPR and detected 8 such cases. Figure 2(d) shows two CRISPRs NC_022084_2214800_2215162 and NC_022084_29_2779 in the archaea *Thermococcus litoralis* DSM 5473, with 5 and 40 spacers, respectively. The identical DRs and the shared Cas genes suggest that these two closely located CRISPRs may be joined into one by the 38 bp sequence between them. This updated information is important, since the missing spacer may be a key anti-invasion factor. 7 other cases were detected in the database DbCRISPR, as demonstrated in Figure S7.

3.2. An Updated CRISPR Definition. Based on the curated annotations of all the CRISPRs in the prokaryotic genomes, this study proposes an updated definition of a CRISPR, as demonstrated in Figure 1 and summarized in Table 1. The minimum number of DRs in a CRISPR may be as low as 2. And a DR may have a length between 14 and 55 bps. A spacer is 9–95 bps in length. The length ratio between a spacer and a DR is proposed to be between 0.3 and 2.5.

The three previous CRISPR annotation programs do not have a consensus agreement on the range of a DR length. In the default settings, CRT, CRISPRFinder, and PILER-CR assume that a DR is at least 19, 23, and 16 bps, respectively. But the Cyanobacteria *Microcystis aeruginosa* NIES 843 and the Firmicutes *Thermacetogenium phaeum* DSM 12270 have a CRISPR with the minimum DR lengths of 14 and 15, respectively. The program CRT requires a DR to be at least

Table 1: Summary of CRISPR definitions in different studies.

# program	minDR #	minDR	maxDR	minSpacer	maxSpacer
CRT	3	19 bp	38 bp	19 bp	48 bp
CRISPRFinder	3	23 bp	55 bp	0.6 DR	2.5 DR
PILER-CR	3	16 bp	64 bp	8 bp	64 bp
caCRISPR	*3*	*14 bp*	*55 bp*	*9 bp (0.3 DR)*	*95 bp (2.5 DR)*

Column "minDR #" gives the minimum number of DRs required to define a CRISPR. Columns "minDR" and "maxDR" are the minimum and maximum DR lengths. The other two columns "minSpacer" and "maxSpacer" give the minimum and maximum spacer lengths for a CRISPR. The proposed definition for a CRISPR in this study is denoted as "caCRISPR," and the other three computer programs compared in this study are CRT (CRISPR Recognition Tool) [7], CRISPRFinder [8], and PILER-CR [6].

19 bps in length, which will miss CRISPRs with a short 17 bp DR in the 7 Firmicutes and an Actinobacteria genomes. The maximum DR length observed in the curated CRISPR annotations is 55 bps. So the program PILER-CR's default value for this feature 64 bps is not strictly supported by the observations. CRT requires the maximum CRISPR DR to be at most 38 bps, which will not recognize CRISPRs in the 30 bacterial genomes. CRISPRFinder has the same setting with caCRISPR for the maximum DR length of 55 bps.

This study proposes the range of a spacer length in two measurements, that is, 9–95 bps and 0.3–2.5 DRs. The program CRT assumes a spacer to be at least 19 bps in length, which will miss CRISPRs in the four Archaea Crenarchaeota genomes and 21 bacterial genomes (14/21~66.67% are Proteobacteria). CRISPRs in the 191 and 49 prokaryotic genomes will not be recognized by the programs CRT and PILER-CR due to their assumptions of the maximum spacer lengths of 48 and 64 bps, respectively. CRISPRFinder has the same requirement as caCRISPR for the maximum spacer length as 2.5 DRs, but its assumption of a minimum spacer length 0.6 DR will miss a CRISPR with the minimum spacer/DR ratio 0.594 and CRISPRs in the 15 bacterial genomes. So the data suggests that the spacer length in two measurements will provide higher specificity and cover all the known CRISPRs.

For the convenience of further analysis, the curated CRISPR annotations are released in the formats of both FASTA and GFF3 in the Supplementary Materials. Other file formats may be provided upon request.

3.3. Taxonomical Distributions of CRISPRs in Prokaryotic Genomes. Among the 4,052 annotated CRISPRs in the 1,302 genomes, the majority comes from the 7 domains (1,149/1,302~88.25%). The seven domains of genomes harbor 3,458/4,052~85.34% of the known CRISPRs, and the 460 (460/1,302~35.33%) Proteobacteria alone have 25.42% (~1,030/4,052) of the annotated CRISPRs. Firmicutes is the second largest domain of the annotated CRISPRs, and 1,030 CRISPRs come from the 323 Firmicutes bacterial genomes. Another 148 CRISPRs are detected in the 411 actinobacterial genomes. And Figure 3(a) shows that Firmicutes tends to have more CRISPRs in one genome, since Firmicutes has more genomes with at least 4.4 CRISPRs (the upper limit of the first bin) than any other domains.

After excluding the top three domains of genomes with CRISPRs, the other four largest domains of CRISPRs are the two archaea domains Euryarchaeota and Crenarchaeota and the other two bacteria domains Cyanobacteria and Bacteriodetes, as shown in Figure 3(b). All the three domains show the trend that the number of genomes decreases with the increased CRISPR number per genome, except the domain Cyanobacteria. The number of Cyanobacteria genomes (11) in the third bin is larger than that (9) in the second bin, as shown in Figure 3(b), suggesting that cyanobacterial genomes tend to harbor a large number of CRISPRs. Actually Cyanobacteria is the domain with the maximum average CRISPR number per genome (6.71 for the 44 genomes), if the bacterial domain Fibrobacteres is omitted. There is only one species *Fibrobacter succinogenes* subsp. *succinogenes* S85 in this domain, and 29 CRISPRs are annotated in this genome.

3.4. Dynamic Patterns of CRISPRs in the Prokaryotic Genomes. CRISPRs show significant variations in its distributions between closely related prokaryotic genomes. We calculate the standard deviation (StdEv) of CRISPR number per genome in the genus with at least three genomes in the annotations. 54 of the 107 genera have a StdEv smaller than 1.0, as shown in Figure 4(a), and only 8 genera demonstrate StdEv = 0, suggesting that these closely related genomes have the same numbers of CRISPRs. 38 genera also show variable CRISPR numbers, with StdEv $\in$ $(1.0, 3.0]$. The archaea genus *Methanocaldococcus* evens reaches StdEv = 8.26 for the CRISPR numbers in its six genomes, ranging from 1 to 22.

We further investigate the CRISPR number variations among different species of the same genus *Thermoanaerobacter*, for its high appearing rate in the CRISPR annotation corrections. Figure 4(b) illustrates how actively CRISPRs in the 8 completely sequenced species of *Thermoanaerobacter* are manipulated by the transposon insertions. Six of the eight species carry CRISPRs inserted by Insertion Sequences (IS) or Miniature Inverted-Repeat Transposable Elements (MITE), as shown in Figure S4. Four CRISPRs in the two genomes *T.* sp. X513 and X514 originate from the same long CRISPR inserted by three copies of the transposon IS110, according to the evidences of almost identical DR sequences and spacers with similar lengths. The same insertion flanking sequences and close phylogenetic distance suggest that the three insertions happen in the common ancestor of these two species. After the divergence of the two genomes, the fourth CRISPR in *T.* sp. X514 continues to expand with 14 more spacers. IS110 also plays an active role in splitting a long CRISPR into shorter ones in the two other *Thermoanaerobacter* genomes, that is, *T. brockii* subsp. *finnii* Ako-1 and

T. pseudethanolicus ATCC 33223. A long CRISPR in *T. brockii* subsp. *finnii* Ako-1 is broken into four by two insertions of IS110 and one IS1634 insertion. The long CRISPR in *T. pseudethanolicus* ATCC 33223 shares the same DR sequence and spacers with similar lengths but undergoes more diversified manipulations, as shown in Figure S4. Besides the three single-copy IS110 insertions, an IS110 dimer and a Miniature Inverted-Repeat Transposable Element (MITE) are also detected in the split of this long CRISPR into shorter ones. These data demonstrate that CRISPRs in the *Thermoanaerobacter* genomes are under active evolutionary manipulation and expansion.

Additional Points

Supplementary data are available at Bioinformatics online.

Competing Interests

The authors declare that there are no competing interests.

Authors' Contributions

Guoqin Mai and Ruiquan Ge contribute equally to this work.

Acknowledgments

This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB13040400), the startup grant from the Jilin University, Shenzhen Peacock Plan (KQCX20130628112914301 and KQCX20130628112914291), the China 863 program (SS2015AA020109-4), Shenzhen Research Grants (JCYJ20130401170306884), and Key Laboratory of Human-Machine-Intelligence Synergic Systems, Chinese Academy of Sciences. Computing resources were partly provided by the Dawning supercomputing clusters at SIAT CAS.

References

[1] R. Sorek, V. Kunin, and P. Hugenholtz, "CRISPR—a widespread system that provides acquired resistance against phages in bacteria and archaea," *Nature Reviews Microbiology*, vol. 6, no. 3, pp. 181–186, 2008.

[2] H. Deveau, J. E. Garneau, and S. Moineau, "CRISPR/Cas system and its role in phage-bacteria interactions," *Annual Review of Microbiology*, vol. 64, pp. 475–493, 2010.

[3] G. Wang, F. Zhou, V. Olman, F. Li, and Y. Xu, "Prediction of pathogenicity islands in enterohemorrhagic *Escherichia coli* O157:H7 using genomic barcodes," *FEBS Letters*, vol. 584, no. 1, pp. 194–198, 2010.

[4] K. Belhaj, A. Chaparro-Garcia, S. Kamoun, N. J. Patron, and V. Nekrasov, "Editing plant genomes with CRISPR/Cas9," *Current Opinion in Biotechnology*, vol. 32, pp. 76–84, 2015.

[5] J. A. Doudna and E. Charpentier, "The new frontier of genome engineering with CRISPR-Cas9," *Science*, vol. 346, no. 6213, Article ID 1258096, 2014.

[6] R. C. Edgar, "PILER-CR: fast and accurate identification of CRISPR repeats," *BMC Bioinformatics*, vol. 8, article 18, 2007.

[7] C. Bland, T. L. Ramsey, F. Sabree et al., "CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats," *BMC Bioinformatics*, vol. 8, article 209, 2007.

[8] I. Grissa, G. Vergnaud, and C. Pourcel, "CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats," *Nucleic Acids Research*, vol. 35, no. 2, pp. W52–W57, 2007.

[9] I. Grissa, G. Vergnaud, and C. Pourcel, "The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats," *BMC Bioinformatics*, vol. 8, article 172, 2007.

[10] D. L. Wheeler, T. Barrett, D. A. Benson et al., "Database resources of the National Center for Biotechnology Information," *Nucleic Acids Research*, vol. 36, no. 1, pp. D13–D21, 2008.

[11] C. M. Lawrence and M. F. White, "Recognition of archaeal CRISPR RNA: No P in the alindromic repeat?" *Structure*, vol. 19, no. 2, pp. 142–144, 2011.

[12] X. Wang, H. Tang, J. E. Bowers, and A. H. Paterson, "Comparative inference of illegitimate recombination between rice and sorghum duplicated genes produced by polyploidization," *Genome Research*, vol. 19, no. 6, pp. 1026–1032, 2009.

[13] T. Liu and J. Huang, "Quality control of homologous recombination," *Cellular and Molecular Life Sciences*, vol. 71, no. 19, pp. 3779–3797, 2014.

[14] P. Siguier, J. Filée, and M. Chandler, "Insertion sequences in prokaryotic genomes," *Current Opinion in Microbiology*, vol. 9, no. 5, pp. 526–531, 2006.

[15] F. Zhou, V. Olman, and Y. Xu, "Insertion sequences show diverse recent activities in *Cyanobacteria* and *Archaea*," *BMC Genomics*, vol. 9, article 36, 2008.

[16] Y. Chen, F. Zhou, G. Li, and Y. Xu, "A recently active miniature inverted-repeat transposable element, Chunjie, inserted into an operon without disturbing the operon structure in Geobacter uraniireducens Rf4," *Genetics*, vol. 179, no. 4, pp. 2291–2297, 2008.

[17] I. G. Duggin, R. G. Wake, S. D. Bell, and T. M. Hill, "The replication fork trap and termination of chromosome replication," *Molecular Microbiology*, vol. 70, no. 6, pp. 1323–1333, 2008.

Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 2061248, 8 pages
http://dx.doi.org/10.1155/2016/2061248

Research Article

Methylation Status of SP1 Sites within miR-23a-27a-24-2 Promoter Region Influences Laryngeal Cancer Cell Proliferation and Apoptosis

Ye Wang,[1] Zhao-Xiong Zhang,[1] Sheng Chen,[1] Guang-Bin Qiu,[2] Zhen-Ming Xu,[3] and Wei-Neng Fu[1]

[1]*Department of Medical Genetics, China Medical University, Shenyang 110122, China*
[2]*Department of Laboratory Medicine, No. 202 Hospital of PLA, Shenyang 110003, China*
[3]*Department of Otolaryngology, No. 463 Hospital of PLA, Shenyang 110007, China*

Correspondence should be addressed to Guang-Bin Qiu; qiuguangbin@163.com, Zhen-Ming Xu; zs840817@163.com, and Wei-Neng Fu; wnfu@mail.cmu.edu.cn

Received 14 January 2016; Revised 26 February 2016; Accepted 8 March 2016

Academic Editor: Hongwei Wang

DNA methylation plays critical roles in regulation of microRNA expression and function. miR-23a-27a-24-2 cluster has various functions and aberrant expression of the cluster is a common event in many cancers. However, whether DNA methylation influences the cluster expression and function is not reported. Here we found a CG-rich region spanning two SP1 sites in the cluster promoter region. The SP1 sites in the cluster were demethylated and methylated in Hep2 cells and HEK293 cells, respectively. Meanwhile, the cluster was significantly upregulated and downregulated in Hep2 cells and HEK293 cells, respectively. The SP1 sites were remethylated and the cluster was significantly downregulated in Hep2 cells into which methyl donor, S-adenosyl-L-methionine, was introduced. Moreover, S-adenosyl-L-methionine significantly increased Hep2 cell viability and repressed Hep2 cell early apoptosis. We also found that construct with two SP1 sites had highest luciferase activity and SP1 specifically bound the gene cluster promoter in vitro. We conclude that demethylated SP1 sites in miR-23a-27a-24-2 cluster upregulate the cluster expression, leading to proliferation promotion and early apoptosis inhibition in laryngeal cancer cells.

1. Introduction

miR-23a, miR-27a, and miR-24-2 consist of miR-23a-27a-24-2 gene cluster which is highly conserved in different species. miR-23a-27a-24-2 cluster and its individual members play important roles in various biological and pathological processes such as cell development [1], proliferation [2], apoptosis [3], differentiation [4], immune response [5], and invasion and metastasis [6], respectively.

Aberrant miR-23a-27a-24-2 cluster expression is reported to be a common event in lots of cancers such as acute lymphoblastic leukemia [7], acute myeloid leukemia [8], chronic lymphocytic leukemia [9], prostate and breast cancer [10], gastric cancer [11], cholangiocarcinoma [12], hepatocellular cancer (HCC) [13], acute promyelocytic leukemia [14], and colorectal cancer [15]. Because three members in the cluster are derived from a single primary transcript, they have similar expression pattern in general. For example, the cluster has been found to be upregulated in acute lymphoblastic leukemia [7], acute myeloid leukemia [8], chronic lymphocytic leukemia [9], prostate and breast cancer [10], gastric cancer [11], cholangiocarcinoma [12], and hepatocellular cancer (HCC) [13], respectively. In several cancers, such as acute promyelocytic leukemia (APL), the cluster is downregulated [14]. In colorectal cancer, however, the first two miRNAs of the cluster are overexpressed and the third is underexpressed [15]. The complex expression patterns suggest that the cluster is tissue-specific and is involved in complicated regulatory mechanism in gene expression.

However, why miR-23a-27a-24-2 cluster is aberrantly expressed is seldom reported. From genetics level, only

Table 1: Primer sequences of miR-23a/27a/24-2 cluster used in the study.

Primer name	Sequence
miR-23a	F: 5′-ATCAC ATTCGCAGGGATTTCC-3′ RTQ-UNIr: 5′-CGAATTCTAGACGTCGAGCGAGCGGACATGCGTGCGTAGTTAACGTTGGTACCGACGTCGGAT-CCACTAGTCC (T)-3′
miR-27a	F: 5′-TTCACAGTGCGTAAGTTCCCG-3′ RTQ-UNIr: 5′-CGAATTCTAGACGTCGAGCGAGCGGACATGCGTGCGTAGTTAACGTTGGTACCGACGTCGGAT-CCACTAGTCC (T)-3′
miR-24-2	F: 5′-TGCGTCAGTTCACGAGGAACAG-3′ RTQ-UNIr: 5′-CGAATTCTAGACGTCGAGCGAGCGGACATGCGTGCGTAGTTAACGTTGGTACCGACGTCGGAT-CCACTAGTCC (T)-3′
U6	F: 5′-CTCCGTTCGCGACGACA-3′ R: 5′-AACCGTTCACGAATTTCGGT-3′

amplification is confirmed to upregulate the cluster expression in gastric cancer cells [16]. In epigenetics, SNPs and methylation are reported to be associated with regulation of the cluster expression. For example, the polymorphisms miR-23a rs3745453, miR-27a rs895819, and rs11671784 could modulate the cluster member's expression [17–19]. In hepatocellular carcinoma, He et al. found that hypomethylation contributes to aberrant miR-23a and miR-27a expression by genome-wide methylated DNA immunoprecipitation chip and miRNA expression microarray assays [20]. miR-23a gene is hypermethylated and upregulated after demethylation in osteosarcoma cells [21]. These suggest that methylation status affects the cluster expression regulation. Unfortunately, how methylation regulates miR-23a and miR-27a expression is not reported.

In our previous study, we found that miR-23a and miR-27a are upregulated in laryngeal cancer [22, 23]. We also analyzed relationship between miR-23a-27a-24-2 cluster polymorphism rs10422126 and laryngeal cancer occurrence. However, the result showed no significant difference between them (data not shown).

In the study, we predicted CG-rich region of miR-23a-27a-24-2 cluster promoter and detected the methylation status in the region spanning two SP1 sites. We also investigated whether methylation status of SP1 sites affects the cluster expression and proliferation and apoptosis in Hep2 cells.

2. Materials and Methods

2.1. Cells and Cell Culture. Human laryngeal carcinoma cells Hep2 and human embryonic kidney cells HEK-293 were obtained from Cell Biology Institute of Shanghai, Chinese Academy of Science. Hep2 and HEK 293 cells were maintained in RPMI-1640 and Dulbecco's high glucose modified Eagle's medium (DMEM), respectively, with 10% fetal bovine serum, 100 nits/mL penicillin, and 100 μg/mL streptomycin in a humidified atmosphere at 37°C with 5% CO_2. S-Adenosyl-L-methionine (SAM) was purchased from Sigma Corporation (MO, USA).

2.2. Quantitative Reverse Transcription-Polymerase Chain Reaction (qRT-PCR). To detect the expression of miR-23a/27a/24-2 cluster in Hep2 and HEK293 cell lines, total RNA was isolated using Trizol reagent (Invitrogen, Carlsbad, CA, USA) following the protocol of the manufacturer. Reverse transcription was performed using the One Step Prime Script miRNA cDNA Synthesis Kit (Takara, Dalian, China) following the manufacturer's instructions. qRT-PCR was performed using SYBR® Premix Ex Taq™II (Takara, Dalian, China) according to the manufacturer's instructions using 7500 Real-time RT-PCR system (Applied Biosystems, Foster City, CA, USA). PCR results were normalized to endogenous U6 and quantified in relation to the controls using the delta-delta CT method. All primers for miR-23a/27a/24-2 cluster used in the study are shown in Table 1.

2.3. Bisulfite Modification and Bisulfite-Specific PCR (BSP). Hep2 cells were treated by SAM. Genomic DNAs isolated from Hep2, HEK-293, and SAM-treated Hep2 cells were used to detect methylation status of CG-rich region in miR-23a/27a/24-2 cluster promoter. Genomic DNA was then bisulfite-modified using the EZ DNA Methylation-Gold™ kit (Zymo Research, Orange, CA, USA) according to the manufacturer's recommendation. Based on the promoter CG-rich region sequence of the cluster, bisulfite PCR primers were designed according to the online primers program "MethPrimer" (http://www.urogene.org/methprimer/). Primers used for BSP are as follows: forward 5′-TTTGTATTTTGGAGTTTGGATTTTG-3′ and reverse 5′-CCTCATTAAACCCTAAACAAACCA-3′. BSP products were then cloned into a T-vector (Takara, Japan) and transformed into JM109 *E. coli* competent cells (Takara, Japan) according to the manufacturer's instructions.

2.4. Proliferation Assay. Hep2 cells were treated by SAM at 0.2 mM, 0.4 mM, 0.6 mM, 0.8 mM, and 1.0 mM concentrations, respectively. SAM-untreated Hep2 cells were used as controls. 3-4 × 10^4 cells were seeded into each well of a 96-well culture plate to a final volume of 100 μL. After

culture for 24 h, 48 h, 72 h, and 96 h, 10 μL of CCK-8 was added to each well and incubated for 1–4 h at 37°C in a 5% CO_2 incubator. Absorbance at 450 nm was measured using a microplate reader. Growth inhibition rate was then calculated. A proliferation curve was plotted based on SAM concentration and growth inhibition rate. The subsequent concentration of SAM treatment was based on IC_{50} value.

2.5. Apoptosis Assay. Apoptotic cells were measured by using an Annexin-V:FITC Apoptosis Detection Kit I (BD Biosciences, San Jose, CA, USA) according to the manufacturer's protocol. Hep2 cells were incubated with 0.2 mM, 0.4 mM, 0.6 mM, 0.8 mM, and 1.0 mM SAM for 72 h, respectively. Cells were then harvested, washed twice with 1x PBS, and resuspended in 100 μL of binding buffer. Cells were incubated with Annexin-V and PI at room temperature for 15 min in the dark. Apoptosis was detected by flow cytometer (FACSCalibur, Becton-Dickinson, USA). Signals Annexin-V^-/PI^-, Annexin-V^+/PI^-, and Annexin-V^+/PI^+ indicate living, early, and late apoptotic cells, respectively.

2.6. Transient Transfection and Luciferase Assays. p450, p498, and p603 constructs containing zero, one, and two SP1 sites in the cluster, respectively, were obtained from GENECHEM (Shanghai, China). Cells seeded in 96-well plate in triplicate were transfected with different constructs by using Lipofectamine 2000™ in accordance with the manufacturer's procedure. pRL-TK (Promega Corporation, Madison, WI, USA) was used as a normalization control. Cells were collected at 48 h after transfection and luciferase activity was measured using a dual-luciferase reporter assay kit (Promega Corporation) by Dual Luciferase Assay System (Promega, USA). Relative luciferase activity was calculated as firefly/*Renilla* luciferase ratio.

2.7. Electrophoretic Mobility Shift Assay (EMSA). Nuclear extracts of Hep2 cells were prepared using a nuclear extract kit (Pierce, USA) following the manufacturer's instructions. Oligonucleotides used in EMSA were synthesized by Sangene (Beijing, China), and their sequences were as follows: SP1 wild type: 5′-CTCTGGGGGCGGGGGGGTCGG-3′ and mutant: 5′-CTCTGGAGAATAAGAGGTCGG-3′. The oligonucleotides were labeled using the biotin 3′ end DNA Labeling Kit (Pierce, USA). EMSA was performed by LightShift Chemiluminescent EMSA kit (Pierce, USA) according to the protocol provided. In brief, nuclear protein extracts were incubated with 3′-end-biotin-labeled probes in binding buffer for 20 min on ice, separated on a 6% nondenaturing polyacrylamide gel, and then transferred onto a nylon membrane and fixed by ultraviolet cross-linking. Protein-DNA complexes were visualized by streptavidin-horseradish peroxidase followed by chemiluminescent detection (Pierce, USA). For competition assays, nuclear protein extracts were incubated with a 100-fold excess of the unlabeled wild type and mutated oligonucleotide duplex competitors, respectively. For supershift reaction, anti-SP1 antibody (Abcam, USA) was incubated with nuclear extracts for 1 h at 4°C prior to the addition of the biotin-labeled DNA probes.

2.8. Statistical Analysis. Unless otherwise stated, each experiment was performed for a minimum of three times. Data were subjected to statistical analysis by SPSS 16.0 software and shown as mean ± standard deviation (SD). A paired samples t-test was used to analyze differences in miR-23a/27a/24-2 cluster expression. Results obtained from cell-based experiments were analyzed by independent samples t-test and one-way ANOVA. $P < 0.05$ is considered statistically significant.

3. Results and Discussion

3.1. Results

3.1.1. A CG-Rich Region in miR-23a-27a-24-2 Cluster Promoter Is Hypomethylated in Hep2 Cells. As shown in Figure 1(a), a CG-rich region with 6 CpGs overlapping two SP1 sites was found in the cluster promoter −530~−410. Bisulfite DNA sequencing results displayed that 4 cytosines in 6 CpGs (66.7%) spanning the two SP1 sites were methylated in HEK-293 cells. In Hep2 cells, none of them were methylated (Figure 1(b)), suggesting that the CG-rich region including two SP 1 sites in Hep2 cells is demethylated compared to that in HEK-293 cells.

3.1.2. Hypomethylation of the CG-Rich Region Contributes to Upregulation of miR-23a-27a-24-2 Cluster in Hep2 Cells. qRT-PCR results indicated that three members of miR-23a-27a-24-2 cluster were significantly overexpressed in Hep2 cells compared to HEK-293 cells (Figure 1(c)). Furthermore, bisulfite DNA sequencing results showed that 3 cytosines in 6 CpGs (50%) including two SP1 sites were remethylated in SAM-treated Hep2 cells (Figure 1(d)), implying that SAM can alter the methylation status in the CpG-rich region. qRT-PCR result displayed that miR-23a-27a-24-2 cluster expression level was significantly lower in SAM-treated Hep2 cells than that in SAM-untreated ones (Figure 1(e)). These results suggest that hypomethylation of the CG-rich region especially two SP1 sites upregulates the cluster expression.

3.1.3. Hypomethylation of the CG-Rich Region Participates in Regulation of Hep2 Cell Proliferation and Apoptosis. To identify whether methylation status of the CpG-rich region affects Hep2 cell functions, we detected Hep2 cell proliferation and apoptosis by MTT and flow cytometry methods, respectively. MTT results showed a decreased viability tendency with the increase of concentration and treatment duration of SAM compared to the controls (Figure 2(a)). Moreover, Hep2 cell proliferation reached a peak at 0.8 mmol/L of SAM treatment on either day and showed significant differences at 0.8 mmol/L on the second and third day (Figure 2(b)). On the contrary, SAM-treated Hep2 cells displayed an increased trend in early apoptosis with an increase in SAM concentration on the third day and began to reveal a significant difference at 0.6 mmol/L compared to the controls (Figure 2(c)). We speculate that the CG-rich region hypomethylation partly regulates the Hep2 cell proliferation and early apoptosis.

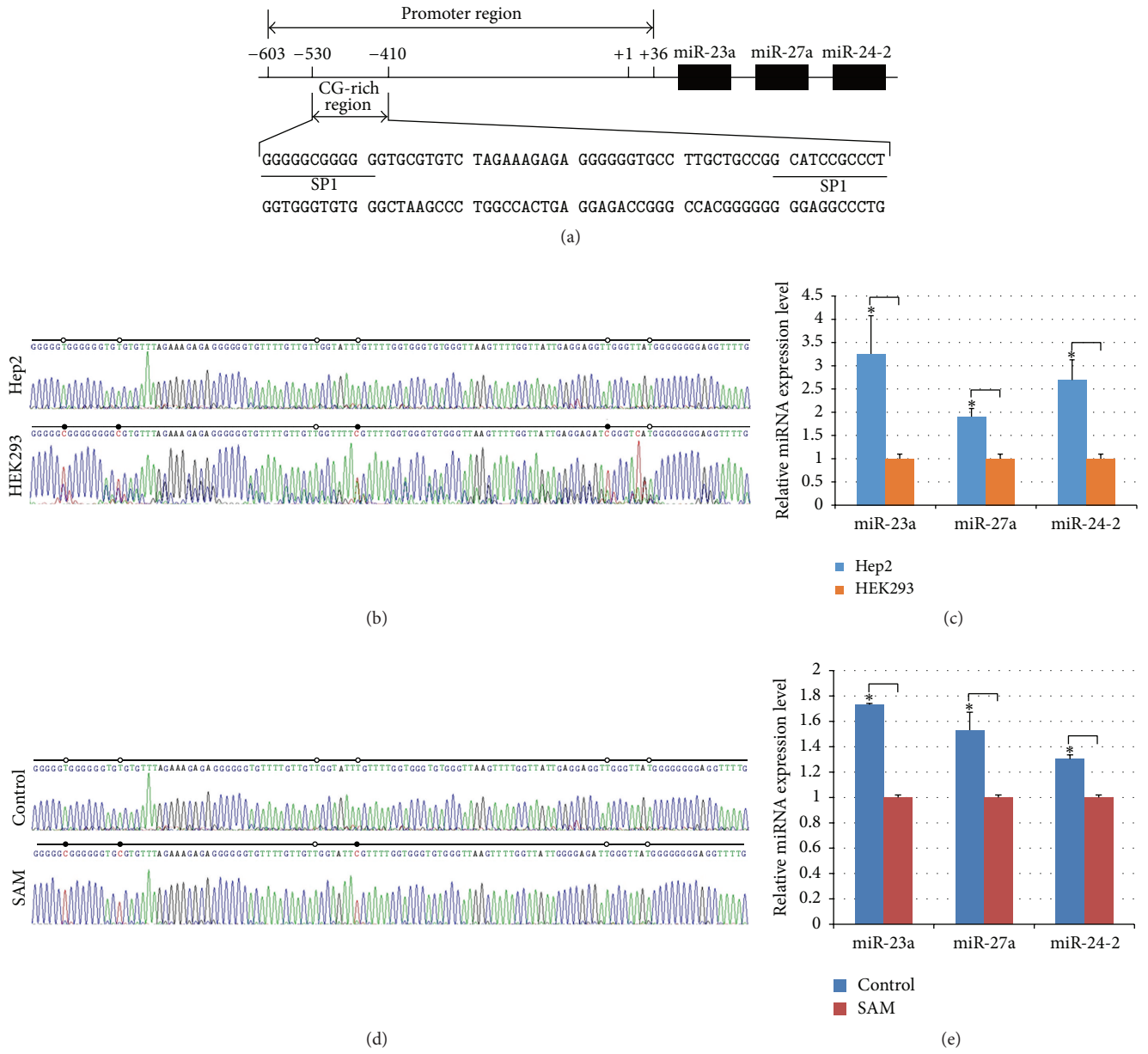


Figure 1: Effects of DNA methylation status of miR-23a-27a-24-2 cluster promoter CG-rich region on the cluster expression. (a) Prediction of miR-23a-27a-24-2 cluster promoter CG-rich region. CG-rich region spanning two SP1 sites is located at the cluster promoter, −530~−410. (b) DNA methylation status of the CG-rich region in Hep2 and HEK-293 cells. Cytosines of CG dinucleotides in the two SP1 sites were hypermethylated in HEK-293 cells compared to Hep2 cells. (c) Expression of miR-23a-27a-24-2 cluster in Hep2 and HEK-293 cells. Members of the cluster were upregulated in Hep2 cells compared to HEK-293 cells. (d) DNA methylation status of the CG-rich region in SAM-treated and SAM-untreated Hep2 cells. Cytosines of CG dinucleotides in the two SP1 sites were remethylated in SAM-treated Hep2 cells compared to SAM-untreated cells. (e) Expression of miR-23a-27a-24-2 cluster in SAM-treated and SAM-untreated Hep2 cells. Members of the cluster were downregulated in SAM-treated Hep2 cells compared to SAM-untreated cells. Hep2 cells were incubated with 1 mmol/L SAM. SAM-untreated cells were used as controls. ∗ indicates $P < 0.05$.

3.1.4. SP1 Sites within the CG-Rich Region Are Important in Regulation of miR-23a-27a-24-2 Cluster. Luciferase reporter assay result indicated that construct overlapping two SP1 sites had strongest luciferase activity and even that harbouring one SP1 site showed significantly higher luciferase ability compared to the controls (Figure 3(a)), implying that the SP1 sites probably regulate miR-23a-27a-24-2 cluster expression. EMSA result displayed that SP1-labled wild type probe had strong binding ability to nuclear proteins compared to the probe-free group. By addition of anti-SP1 antibody, a shifting band was detected. SP1-unlabeled probe reduced the binding and SP1-mutant probe had no effect on

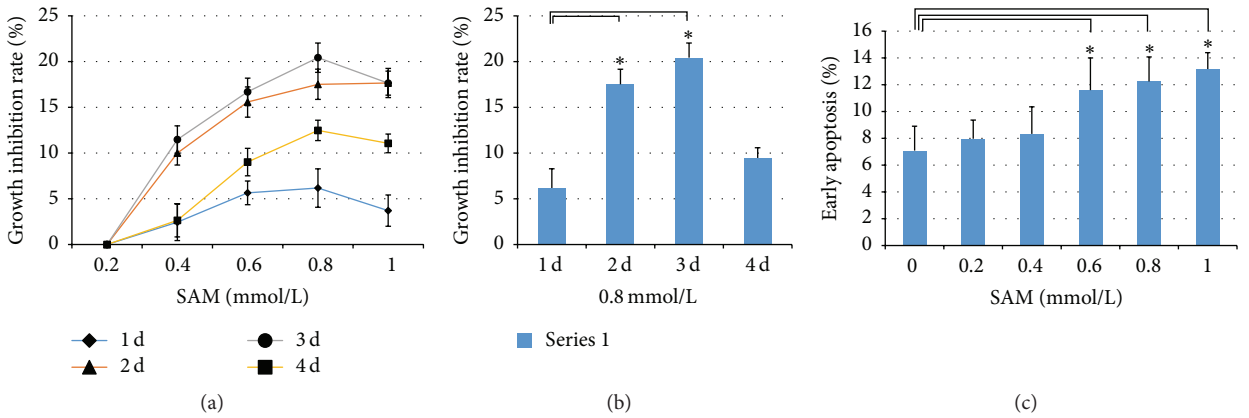
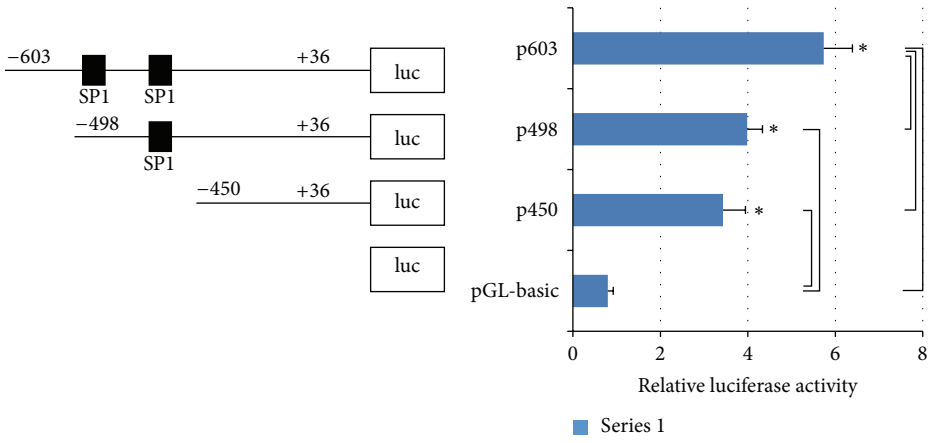


Figure 2: Effects of DNA methylation status of miR-23a-27a-24-2 cluster promoter CG-rich region on proliferation and apoptosis. (a) Inhibition of different concentrations of SAM on Hep2 cell growth. (b) Inhibition analysis of 0.8 mmol/L SAM on Hep2 cell growth. (c) Early apoptosis analysis of different concentrations of SAM on Hep2 cells on the third day. $*$ indicates $P < 0.05$.



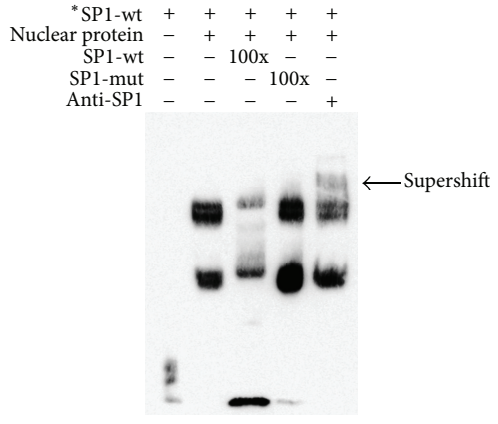


Figure 3: Activities of miR-23a-27a-24-2 cluster promote and binding of SP1 to the cluster CG-rich region. (a) Relative luciferases of different constructs in miR-23a-27a-24-2 cluster are promoted. (b) Binding of SP1 to the cluster CG-rich region in vitro. $*$ indicates $P < 0.05$.

the binding (Figure 3(b)). The findings confirmed that both SP1 sites are import cis-acting elements in miR-23a-27a-24-2 cluster regulation.

3.2. Discussion. It is well-known that transcription factors regulate target gene expression via binding cis-acting elements [24]. As a basal transcription factor, SP1 plays a critical role in regulation of so-called housekeeping genes especially in the absence of TATA box [25].

miR-23a-27a-24-2 cluster promoter is located at the region −603 bp~+38 bp that lacks the known common promoter element TATA box [26, 27]. In the study, we first found a CG-rich region with six CpG dinucleotides covering two SP1 sites. Compared to HEK-293 cells, we discovered that the region is hypomethylated where both SP1 sites are demethylated in Hep2 cells. Meanwhile, all members of the cluster are significantly upregulated in Hep2 cells compared to HEK-293 cells, suggesting that hypomethylation especially demethylation of both SP1 sites in the region might contribute to overexpression of the cluster. SAM, S-adenosyl-L-methionine, is a very useful methyl donor in epigenetic mechanism study [28]. In the present study, we also found most cytosines in the CG-rich region are methylated in Hep2 cells after being treated by SAM. All three miRNAs of miR-23a-27a-24-2 cluster are significantly downregulated in SAM-treated Hep2 cells compared to the controls, further indicating that abnormal methylation in the CG-rich region regulates the cluster expression.

Recently, miR-23a and 27a are reported in an aberrant methylation status in several studies. Similar to our results, miR-23a overexpression is speculated to be associated with its hypomethylation in leukemic cells [29]. In hepatocellular carcinoma, hypomethylation is considered to be responsible for upregulation of miR-23a and miR-27a [13]. On the contrary, the miR-23a gene promoter region was found to be hypermethylated, leading to downregulation of miR-23a in osteosarcoma cells [21]. However, these results do not tell us the detailed information on alteration of methylation status and molecular mechanism of which methylation-related sites could affect cancer cell function.

It is also well-known that miRNAs participate in regulation of cancer cell proliferation, apoptosis, differentiation, migration, and metastasis via different targets [30–33], respectively. At present, lots of targets of miR-23a-27a-24-2 cluster have been identified such as GJA1, BCL2, CDC27, 14-3-3θ, NAIF1, and SOX7 [34–39]. In our previous study, we found that miR-23a and miR-27a overexpression in Hep2 cells promotes cancer cell proliferation and represses apoptosis by targeting APAF-1 and PLK2 [22, 23], respectively.

In the present study, SAM-treated Hep2 cells showed significantly lower level of proliferation and higher level of apoptosis than controls, indicating that SAM increases proliferation and decreases apoptosis in laryngeal cancer cells. SAM-induced proliferation prevention and apoptosis activation have been found in other studies. For example, SAM inhibits osteosarcoma and colorectal cancer cell proliferation [40, 41]. SAM causes apoptosis in normal liver L-02 cells and undifferentiated pheochromocytoma PC12 cells [42, 43]. Our EMSA result indicates that SP1 is an important *trans*-acting factor of miR-23a-27a-24-2 cluster and mutant SP1 reduces the banding ability of SP1 to the cluster promoter. Similar to the SP1 variation, we suggest that methylation status of SP1 sites may interfere with the binding of SP1 to miR-23a-27a-24-2 cluster. It has been shown that CpG methylation in the promoter region of BLU could prevent itself from binding to SP1 [44].

4. Conclusion

We conclude that demethylated SP1 sites in CG-rich region of miR-23a-27a-24-2 cluster promoter result in the cluster overexpression, leading to proliferation promotion and apoptosis inhibition probably via targeting the related targets such as APAF-1 and PLK2 in laryngeal cancer cells. These provide us with an important basis in our future work on miR-23a-27a-24-2 cluster promoter methylation in human cancer tissues and its clinical significance in tumorigenesis.

Disclosure

Ye Wang and Zhao-Xiong Zhang are co-first authors

Competing Interests

The authors declare that they have no competing interests.

Authors' Contributions

Ye Wang, Zhao-Xiong Zhang, and Sheng Chen performed the experiment. Wei-Neng Fu and Zhen-Ming Xu designed the research. Guang-Bin Qiu carried out data analysis. Wei-Neng Fu wrote the paper.

Acknowledgments

This work was supported by two grants from the National Natural Science Foundations of China (81172577 and 81372876).

References

[1] K. Y. Kong, K. S. Owens, J. H. Rogers et al., "MIR-23A microRNA cluster inhibits B-cell development," *Experimental Hematology*, vol. 38, no. 8, pp. 629–640.e1, 2010.

[2] H. Peng, X. Wang, P. Zhang, T. Sun, X. Ren, and Z. Xia, "miR-27a promotes cell proliferation and metastasis in renal cell carcinoma," *International Journal of Clinical and Experimental Pathology*, vol. 8, no. 2, pp. 2259–2266, 2015.

[3] A. Musto, A. Navarra, A. Vocca et al., "miR-23a, miR-24 and miR-27a protect differentiating ESCs from BMP4-induced apoptosis," *Cell Death and Differentiation*, vol. 22, no. 6, pp. 1047–1057, 2015.

[4] M. Q. Hassan, J. A. Gordon, M. M. Beloti et al., "A network connecting Runx2, SATB2, and the miR-23a~27a~24-2 cluster regulates the osteoblast differentiation program," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, pp. 19879–19884, 2010.

[5] R. Lin, J. H. Sampson, Q. Li, and B. Zhu, "miR-23a blockade enhances adoptive T cell transfer therapy by preserving immune-competence in the tumor microenvironment," *Onco-Immunology*, vol. 4, no. 3, 2015.

[6] X. Li, X. Liu, W. Xu et al., "c-MYC-regulated miR-23a/24-2/27a cluster promotes mammary carcinoma cell invasion and hepatic metastasis by targeting Sprouty2," *The Journal of Biological Chemistry*, vol. 288, no. 25, pp. 18121–18133, 2013.

[7] S. Mi, J. Lu, M. Sun et al., "MicroRNA expression signatures accurately discriminate acute lymphoblastic leukemia from acute myeloid leukemia," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 50, pp. 19971–19976, 2007.

[8] J. Organista-Nava, Y. Gómez-Gómez, B. Illades-Aguiar et al., "High miR-24 expression is associated with risk of relapse and poor survival in acute leukemia," *Oncology Reports*, vol. 33, no. 4, pp. 1639–1649, 2015.

[9] V. Fulci, S. Chiaretti, M. Goldoni et al., "Quantitative technologies establish a novel microRNA profile of chronic lymphocytic leukemia," *Blood*, vol. 109, no. 11, pp. 4944–4951, 2007.

[10] M. D. Mattie, C. C. Benz, J. Bowers et al., "Optimized high-throughput microRNA expression profiling provides novel biomarker assessment of clinical prostate and breast cancer biopsies," *Molecular Cancer*, vol. 5, article 24, 2006.

[11] G. Ma, W. Dai, A. Sang, X. Yang, and C. Gao, "Upregulation of microRNA-23a/b promotes tumor progression and confers poor prognosis in patients with gastric cancer," *International Journal of Clinical and Experimental Pathology*, vol. 7, no. 12, pp. 8833–8840, 2014.

[12] F. Meng, R. Henson, M. Lang et al., "Involvement of human micro-RNA in growth and response to chemotherapy in human cholangiocarcinoma cell lines," *Gastroenterology*, vol. 130, no. 7, pp. 2113–2129, 2006.

[13] S. Huang, X. He, J. Ding et al., "Upregulation of miR-23a~27a~24 decreases transforming growth factor-beta-induced tumor-suppressive activities in human hepatocellular carcinoma cells," *International Journal of Cancer*, vol. 123, no. 4, pp. 972–978, 2008.

[14] A. Saumet, G. Vetter, M. Bouttier et al., "Transcriptional repression of microRNA genes by PML-RARA increases expression of key cancer proteins in acute promyelocytic leukemia," *Blood*, vol. 113, no. 2, pp. 412–421, 2009.

[15] S. Volinia, G. A. Calin, C.-G. Liu et al., "A microRNA expression signature of human solid tumors defines cancer gene targets," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 7, pp. 2257–2261, 2006.

[16] J. An, Y. Pan, Z. Yan et al., "MiR-23a in amplified 19p13.13 loci targets metallothionein 2A and promotes growth in gastric cancer cells," *Journal of Cellular Biochemistry*, vol. 114, no. 9, pp. 2160–2169, 2013.

[17] E. Ridolfi, C. Fenoglio, C. Cantoni et al., "Expression and genetic analysis of microRNAs involved in multiple sclerosis," *International Journal of Molecular Sciences*, vol. 14, no. 3, pp. 4375–4384, 2013.

[18] J.-Y. Ma, H.-J. Yan, Z.-H. Yang, and W. Gu, "Rs895819 within miR-27a might be involved in development of non small cell lung cancer in the Chinese Han population," *Asian Pacific Journal of Cancer Prevention*, vol. 16, no. 5, pp. 1939–1944, 2015.

[19] Y. Deng, H. Bai, and H. Hu, "Rs11671784 G/A variation in miR-27a decreases chemo-sensitivity of bladder cancer by decreasing miR-27a and increasing the target RUNX-1 expression," *Biochemical and Biophysical Research Communications*, vol. 458, no. 2, pp. 321–327, 2015.

[20] X.-X. He, S.-Z. Kuang, J.-Z. Liao et al., "The regulation of microRNA expression by DNA methylation in hepatocellular carcinoma," *Molecular Biosystems*, vol. 11, no. 2, pp. 532–539, 2015.

[21] Y. He, C. Meng, Z. Shao, H. Wang, and S. Yang, "MiR-23a functions as a tumor suppressor in osteosarcoma," *Cellular Physiology and Biochemistry*, vol. 34, no. 5, pp. 1485–1496, 2014.

[22] X.-W. Zhang, N. Liu, S. Chen et al., "Upregulation of microRNA-23a regulates proliferation and apoptosis by targeting APAF-1 in laryngeal carcinoma," *Oncology Letters*, vol. 10, no. 1, pp. 410–416, 2015.

[23] Y. Tian, S. Fu, G.-B. Qiu et al., "MicroRNA-27a promotes proliferation and suppresses apoptosis by targeting PLK2 in laryngeal carcinoma," *BMC Cancer*, vol. 14, no. 1, article 678, 2014.

[24] S. L. Noton and R. Fearns, "Initiation and regulation of paramyxovirus transcription and replication," *Virology*, vol. 479-480, pp. 545–554, 2015.

[25] K. Beishline and J. Azizkhan-Clifford, "Sp1 and the 'hallmarks of cancer'," *FEBS Journal*, vol. 282, no. 2, pp. 224–258, 2015.

[26] R. Chhabra, R. Dubey, and N. Saini, "Cooperative and individualistic functions of the microRNAs in the miR-23a~27a~24-2 cluster and its implication in human diseases," *Molecular Cancer*, vol. 9, article 232, 2010.

[27] Y. Lee, M. Kim, J. Han et al., "MicroRNA genes are transcribed by RNA polymerase II," *The EMBO Journal*, vol. 23, no. 20, pp. 4051–4060, 2004.

[28] I. Afanas'ev, "Mechanisms of superoxide signaling in epigenetic processes: relation to aging and cancer," *Aging and Disease*, vol. 6, no. 3, pp. 216–227, 2015.

[29] Z. Xishan, L. Xianjun, L. Ziying, C. Guangxin, and L. Gang, "The malignancy suppression role of miR-23a by targeting the BCR/ABL oncogene in chromic myeloid leukemia," *Cancer Gene Therapy*, vol. 21, no. 9, pp. 397–404, 2014.

[30] C. Y. Ok, L. Li, and K. H. Young, "EBV-driven B-cell lymphoproliferative disorders: from biology, classification and differential diagnosis to clinical management," *Experimental and Molecular Medicine*, vol. 47, no. 1, article e132, 2015.

[31] H. Liang, Z. Fu, X. Jiang et al., "miR-16 promotes the apoptosis of human cancer cells by targeting FEAT," *BMC Cancer*, vol. 15, no. 1, article 448, 2015.

[32] J. Yuan, G. Xiao, G. Peng et al., "MiRNA-125a-5p inhibits glioblastoma cell proliferation and promotes cell differentiation by targeting TAZ," *Biochemical and Biophysical Research Communications*, vol. 457, no. 2, pp. 171–176, 2015.

[33] M. Xu, M. Gu, K. Zhang, J. Zhou, Z. Wang, and J. Da, "miR-203 inhibition of renal cancer cell proliferation, migration and invasion by targeting of FGF2," *Diagnostic Pathology*, vol. 10, article 24, 2015.

[34] N. Wang, L.-Y. Sun, S.-C. Zhang et al., "MicroRNA-23a participates in estrogen deficiency induced gap junction remodeling of rats by targeting GJA1," *International Journal of Biological Sciences*, vol. 11, no. 4, pp. 390–403, 2015.

[35] B. Sabirzhanov, Z. Zhao, B. A. Stoica et al., "Downregulation of miR-23a and miR-27a following experimental traumatic brain injury induces neuronal cell death through activation of proapoptotic Bcl-2 proteins," *The Journal of Neuroscience*, vol. 34, no. 30, pp. 10055–10071, 2014.

[36] Y.-Q. Ren, F. Fu, and J. Han, "MiR-27a modulates radiosensitivity of triple- negative breast cancer (TNBC) cells by targeting CDC27," *Medical Science Monitor*, vol. 21, pp. 1297–1303, 2015.

[37] K. A. Scheibner, B. Teaboldt, M. C. Hauer et al., "MiR-27a functions as a tumor suppressor in acute leukemia by regulating 14-3-3θ," *PLoS ONE*, vol. 7, no. 12, Article ID e50895, 2012.

[38] G. Zhao, L. Liu, T. Zhao et al., "Upregulation of miR-24 promotes cell proliferation by targeting NAIF1 in non-small cell lung cancer," *Tumor Biology*, vol. 36, no. 5, pp. 3693–3701, 2015.

[39] Y. Ma, X.-G. She, Y.-Z. Ming, and Q.-Q. Wan, "miR-24 promotes the proliferation and invasion of HCC cells by targeting SOX7," *Tumor Biology*, vol. 35, no. 11, pp. 10731–10736, 2014.

[40] S. Parashar, D. Cheishvili, A. Arakelian et al., "S-adenosylmethionine blocks osteosarcoma cells proliferation and invasion in vitro and tumor metastasis in vivo: therapeutic and diagnostic clinical applications," *Cancer Medicine*, vol. 4, no. 5, pp. 732–744, 2015.

[41] K. Módis, C. Coletta, A. Asimakopoulou et al., "Effect of S-adenosyl-L-methionine (SAM), an allosteric activator of cystathionine-β-synthase (CBS) on colorectal cancer cell proliferation and bioenergetics in vitro," *Nitric Oxide*, vol. 41, pp. 146–156, 2014.

[42] W. Q. Zhao, Z. Williams, K. R. Shepherd et al., "S-adenosylmethionine-induced apoptosis in PC12 cells," *Journal of Neuroscience Research*, vol. 69, no. 4, pp. 519–529, 2002.

[43] L. Ji, Y. Chen, and Z. Wang, "Protection of S-adenosyl methionine against the toxicity of clivorine on hepatocytes," *Environmental Toxicology and Pharmacology*, vol. 26, no. 3, pp. 331–335, 2008.

[44] K. Xiao, Z. Yu, D.-T. Shi et al., "Inactivation of BLU is associated with methylation of Sp1-binding site of BLU promoter in gastric cancer," *International Journal of Oncology*, vol. 47, no. 2, pp. 621–631, 2015.

Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 3524908, 10 pages
http://dx.doi.org/10.1155/2016/3524908

Research Article

SNP Mining in Functional Genes from Nonmodel Species by Next-Generation Sequencing: A Case of Flowering, Pre-Harvest Sprouting, and Dehydration Resistant Genes in Wheat

Zhong-Xu Chen, Mei Deng, and Ji-Rui Wang

Triticeae Research Institute, Sichuan Agricultural University, Chengdu 611130, China

Correspondence should be addressed to Ji-Rui Wang; wangjirui@gmail.com

Received 30 December 2015; Accepted 18 February 2016

Academic Editor: Hongwei Wang

As plenty of nonmodel plants are without genomic sequences, the combination of molecular technologies and the next generation sequencing (NGS) platform has led to a new approach to study the genetic variations of these plants. Software GATK, SOAPsnp, samtools, and others are often used to deal with the NGS data. In this study, BLAST was applied to call SNPs from 16 mixed functional gene's sequence data of polyploidy wheat. In total 1.2 million reads were obtained with the average of 7500 reads per genes. To get accurate information, 390,992 pair reads were successfully assembled before aligning to those functional genes. Standalone BLAST tools were used to map assembled sequence to functional genes, respectively. Polynomial fitting was applied to find the suitable minor allele frequency (MAF) threshold at 6% for assembled reads of each functional gene. SNPs accuracy form assembled reads, pretrimmed reads, and original reads were compared, which declared that SNPs mined from the assembled reads were more reliable than others. It was also demonstrated that mixed samples' NGS sequences and then analysis by BLAST were an effective, low-cost, and accurate way to mine SNPs for nonmodel species. Assembled reads and polynomial fitting threshold were recommended for more accurate SNPs target.

1. Background

Next-generation sequencing technologies (e.g., Solexa/Illumina, SOLiD/ABI, 454/Roche, and HeliScope/Helicos) have opened a new field to genotype analysis of nonmodel organisms. Technologies generating long-sequence reads (200–400 bp) are increasingly used in evolutionary studies of nonmodel organisms, but the short sequence reads (30–150 bp) can be produced at lower cost and with more sequence data per run [1]. Short-read technologies are typically best suitable for resequencing projects where a reference genome on which the reads can be mapped is already available [2, 3]. Because of the short-read lengths, the application of NGS technologies has generally been restricted to nonmodel organisms for which the genome sequences is not available. However, recently many algorithmic and experimental advances have made it possible to succeed at de novo sequence projects [4–6]. But some of these reads (both long and short) contain adapters and other exogenous contents by experimental designs. On other cases, adapters were sequenced inadvertently when they are out of operational errors and other unknown reasons. If these adapters were not trimmed out, they would interfere with the downstream data analysis, such as mapping the reads to the reference genome and de novo assembly [7, 8].

For most of the next-generation sequencing technologies (both single-read and paired-end libraries), the quality of the sequencing gets lower while approaching the end of the reads. If excessive sequencing errors occurred in the end of the reads, this would affect the accuracy of mapping and other downstream analysis, even if the reads contain high-quality bases in the beginning. To prevent otherwise high-quality reads from being rejected during quality filtering or from influencing mapping or assembly processes, it can be beneficial to trim bases from poor-quality ends of reads [9].

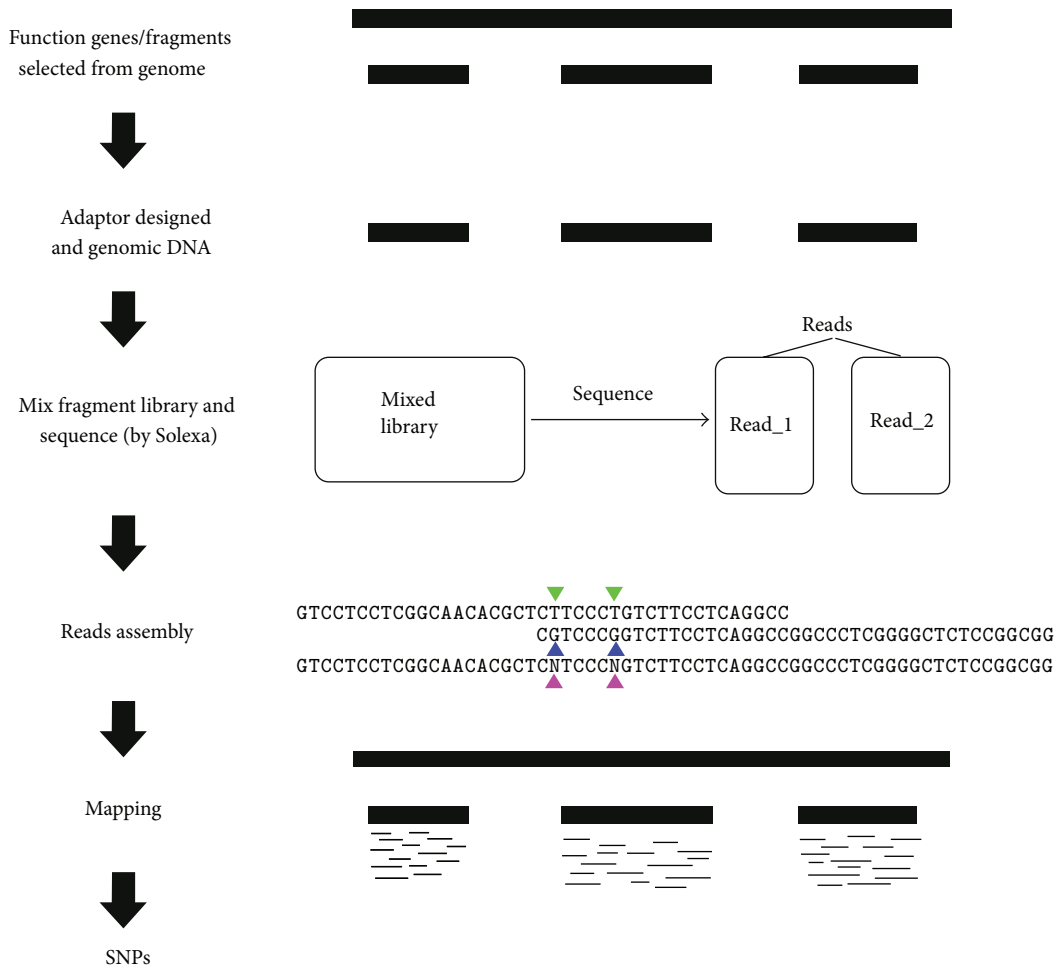


Figure 1: The main steps to mine SNPs on function genes.

Illumina's sequencing by synthesis (SBS) technology (Illumina report) is the most successful and widely adopted next-generation sequencing platform worldwide, which is also the only platform that offers a short-insert paired-end capability for high-resolution genome sequencing as well as long-insert paired-end reads using the same robust chemistry for efficient sequence assembly, de novo sequencing, large-scale structural variation detection, and so on [10].

Triticeae has large and complex genomes with a great abundance of repeated sequences, which does not have a very good whole genome reference available now. Studies on these plants whose polyploidy has further increased genome size and complexity have not been able to fully take advantage of next-generation sequencing for SNP discovery (since SNPs are of more importance on functional genes coding region, 16 genes were molecular-cloned and resequenced form wheat as a case). After these genes were cloned and mixed, these genes were resequenced by NGS Solexa platform and SNPs were called following our pipelines in Figure 1. The polynomial fitting equation was applied to find the best threshold value to filter the low quality SNPs.

2. Materials and Methods

2.1. DNA Isolation and PCR Amplification. Genomic DNA was extracted from leaves of single plants of about 2 weeks old with a modified CTAB protocol. 16 functional genes were randomly selected from NCBI database with the sequences as reference in the following study (Table 1).

Anchored primers were designed on the basis of conserved sequences outside of the polymorphic regions. PCR amplification was performed with GeneAmp PTC-240 cycler (Bio-Rad) in 50 μL volume which consisted of 100 ng of genomic DNA, 100 μM of each dNTP, 1 μM of each primer, 1 U Taq polymerase with high fidelity, 1.5 mM Mg^{2+}, and 1x PCR buffer. The cycling parameters were 95°C for 5 min to predenature, followed by 35 cycles of 95°C for 50 sec, 50–60°C for 30 sec and 72°C 45 sec, and a final extension at 72°C for 5 min. Desired PCR products were obtained by agarose gel. The fragments of genes were mixed with similar concentration.

2.2. Sequence Data Quantity and Quality. Ten mixed DNA samples were sequenced in one run with Illumina Solexa

Table 1: Information of sixteen functional genes.

Name	NCBI number	Length	Product
ABA8OH	[GenBank: AB455560]	654	ABA 8-hydroxylase
ABI5	[GenBank: AB238934]	1540	bZip-type transcription factor TaABI5
ACC1	[GenBank: EU660901]	1131	Plastid acetyl-CoA carboxylase
Apx	[GenBank: AY513261]	1354	Thylakoid ascorbate peroxidase
DRF	[GenBank: FJ560492]	963	Dehydration responsive factor 1 variant
EMH5	[GenBank: X73228.1]	443	Early-methionine-labeled protein
ERD4	[GenBank: AK330512]	810	Transmembrane protein 63B-like
FUC3	[GenBank: BQ806797]	564	Predicted protein
GSK	[GenBank: DQ678922]	527	GSK-like kinase 1A
HKT8	[GenBank: DQ646339]	866	High affinity K_+ transporters
LEA1	[GenBank: AY148490]	816	Late embryogenesis abundant protein
LEC1	[GenBank: BT009029]	910	Nuclear transcription factor Y subunit B1
PhyC	[GenBank: AJ295224]	934	Phytochrome C
Q	[GenBank: AY702960]	809	Floral homeotic protein
WDAI	[GenBank: AY729672]	446	Dimeric alpha-amylase inhibitor
ZCCT1	[GenBank: AY485644]	669	Zinc finger-CCT domain

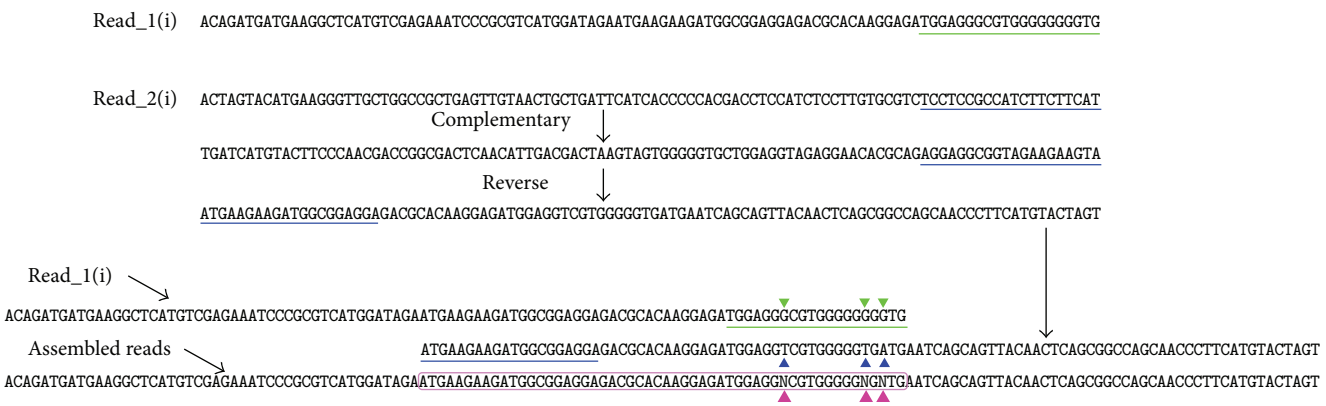


Figure 2: Reads assembly. A(i).fastq and B(i).fastq were one-paired-end reads. The color lines were low quality parts (20 bp). Purple wireframe was the assembled reads part. Solid triangle was the locus which was not consistent in two reads. Paired-end reads were reverse compliment reads. To assemble the two reads, reverse compliment sequence should be calculated by one of them and the other one should be kept. The entire mismatch locus would be set as "N."

platform. We get the sequencing result as pairing reads, which was stored in two fastq files, "read_1.fq" and "read_2.fq," respectively. The sequences at the same position from read_1.fq and read_2.fq are pairing. In each file there were about 0.6 million reads and all reads were the same in length. Each pair should belong to the same reference gene and the paired sequences reversed complementary to each other. File read_1 and file read_2 are corresponding to each other in lines. read_1 is positive sequencing result while read_2 is reverse complementary sequencing result and they could be assembled into one tag if both reads were of high quality (Figure 2). Usually raw reads that only have 3′ adaptor fragments should be removed before data analysis. The following analysis was carried out after the dirty raw reads were removed (Illumina report).

2.3. Assembly and Alignment. Theoretically, the overlap part of two assembled reads should have totally consistent code. But because the sequencing techniques still have read errors, there will be some low quality locus at the end of the sequence. Generally, when we intend to map reads to reference, we will take a reads quality inspection and cut some length to control the read quality. In this study, to avoid the influence of the final SNP sites statistic caused by such case, we set such locus of each assemble sequence as "N" (Figure 2). In the following basic group frequency statistic of reference sequence, "N" is

not participated in the statistic. Thus it eliminates the problem of bad quality of reads in the end; meanwhile it reduces the influence of the SNP quality sites caused by the whole segment sequencing.

As there was no genome reference in nonmodel plant, people usually do mapping works without a genome reference and then calculate the SNPs [11, 12]. Here the DNA sequences of known functional gene were used as reference. To make reads align to reference, we make all the assembled reads into databases with *standalone BLAST tool* (NCBI). Meanwhile to compare the quality difference between assembled reads and nonassembled reads from the same sequence file, among the rest of reads the nonassembled ones were also made into a new database. Then we used the function genes as the query sequence to blast in the database by basic local alignment algorithm [13]. In some of our function genes there are several low-complexity fragments and at the same time the BLAST tool will not calculate the low-complexity part as default. Therefore, we should set the "-F" as "F" to close the low-complexity filter when we use the *blast all* command. To compare the quality of the assembled reads and nonassembled reads, another database was set up by nonassembled reads and the 16 function genes were blast in each database. Blast of 16 genes (with 800 bp average length) in one database containing 0.4 million reads could be completed in 10 minutes by regular PC.

2.4. SNPs Calling. Researchers selected SNPs when the MAF is more than 1% for human sequences, while they selected MAF > 5% for plant sequences. All of those are an estimate threshold. As we all know, different experiments may have their own errors and the sequence quality is also different when different technology platforms were used. In this study, we present a new way to find a reasonable MAF for each independent experiment. First we selected some stable genes which were already known as comparable samples and sequence with other samples together. Then the ratios of SNPs change by the MAF were calculated. To observe those trends of SNPs rations variation feature better, polynomial equation was applied to fit the curves (theoretically, N-order polynomial can approximate to any nonlinear function). We derived the first-order differential equation of fitting polynomial equation and that is the accelerating equation of initial equation. The stable value of the accelerated curve was the best threshold.

To check the result of SNPs' ratio by this process, the pretrimmed reads and original reads (clean and adapts discarded) were also used to map and screen SNPs. Three kinds of reads data were compared by SNPs' ratio and position. The assembled reads data should have less SNPs than other reads at the same MAF threshold.

3. Result and Discussion

3.1. Assembled Reads. 16 function gene samples were sequenced in one run and 2 fastq files (each file contains 589573 reads) were output. The usage of the methods referred above to assembled reads and 390992 pairs of reads were successfully assembled. The assembled reads rate was about

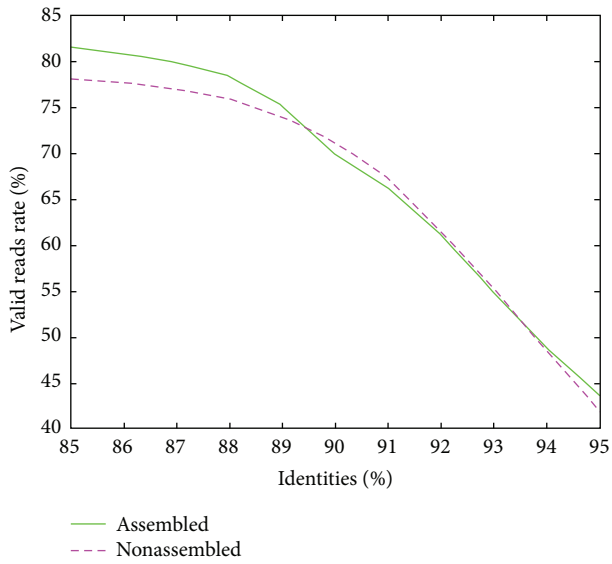


Figure 3: Rate curve of reads can be aligned to reference by identity varied. The valid contigs rate equals the number of the contigs which successfully aligned to references dividing the total reads number in the database.

66.32%. The average length of assembled reads was 155.10, which illustrated that when two reads assembled nearly 50 bp locus will be overlapped. Over 98.56% assembled reads were assembled by reverse complementary reads; meanwhile the 1.5% assembled reads from others may have very low quality. To get accurate result, raw data were reprocessed (Figure 1), and only assembled reads with both forward and reverse complementary reads were selected for accurate sequence. As we checked the sequence data, only 15~20 bp of original reads in the end were of low quality. Thus the low quality segment of the two reads will be aligned to the other reads (Figure 2). If there is any different code at the alignment locus, that locus will be set as "N" and when we align reads to references sequence, "N" will not be calculated. Thus, the problem of low quality segment in the reads will be solved. In blast result of the nonassembled reads database, most contigs are longer than 80 bp; meanwhile when blasting in assembled reads database, there were many short contigs (more or less than 20 bp) aligned to references. We use standalone BLAST tool to blast function genes in local database. To compare the sequence quality of the assembled and nonassembled reads, we made two local databases.

One database consists of assembled reads and the other consists of nonassembled reads. When blasting in the assembled reads database, 321919 contigs have successfully aligned to the function genes when the identity threshold was set as 85% identities and the number of contigs changed to 249076 by the threshold 90% identities. As a result of blasting in nonassembled database, 314977 contigs from 397162 recorders were aligned to the same query sequence (Table 2). Comparing both assembled and nonassembled valid reads by different blast thresholds, assembled sequence performed high mapping rate (Figure 3). We found that the rates of the successful aligned contigs in each database, both assembled

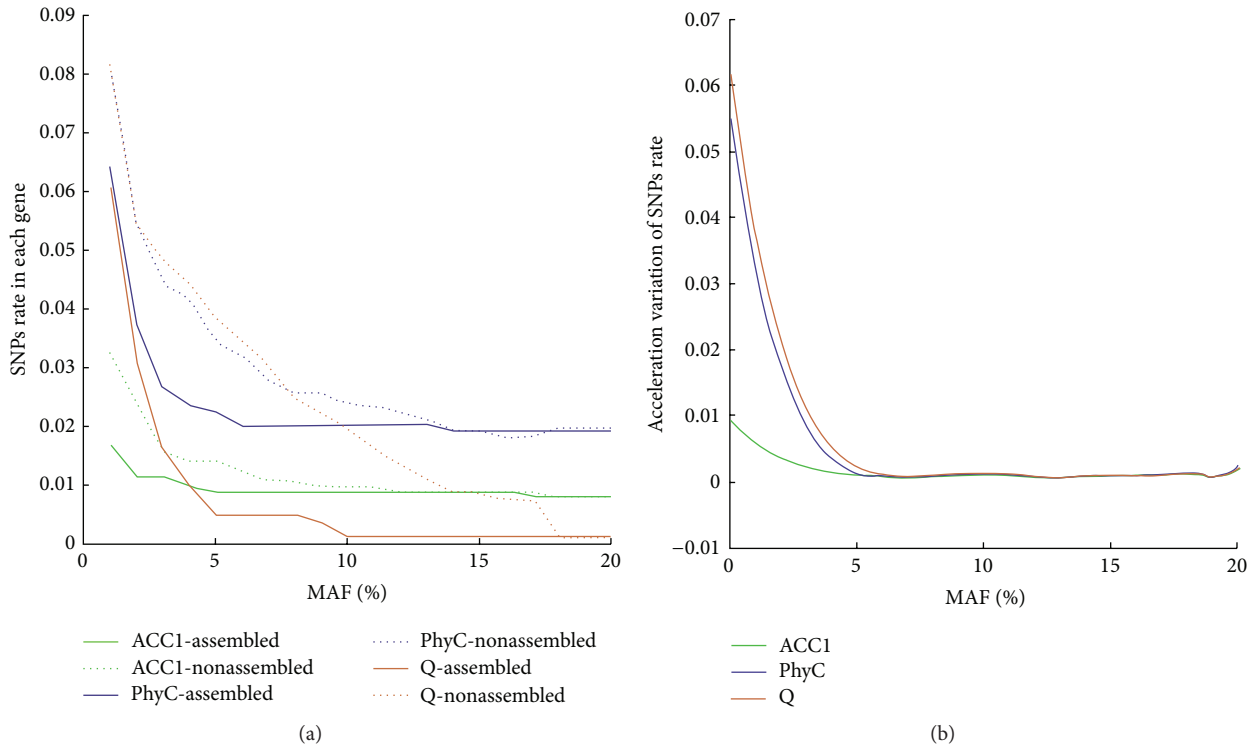


Figure 4: Curve of SNPs rate with the threshold value of MAF variation. (a) SNPs rate curves. The x-axis shows the MAF variation and the y-axis was the SNPs' proportion in each gene. Solid lines are a result of assembled reads and dotted lines are of nonassembled reads. (b) The curve of accelerating equation from assembled database. The x-axis is also the MAF variation, but the y-axis was the acceleration of SNPs variation by MAF. The curve was calculated by the fitting polynomial from (a).

Table 2: Elementary information about the reads.

		Reads number	Average length
Assembled	Original reads	390992 (pair)	155.10
	Aligned to reference	219433 (pair)	156.90
Nonassembled	Original reads	198581 (pair)	100
	Aligned to reference	206362 (single)	81.99

The threshold of the aligned identities was 85%.

and nonassembled reads, were of little difference. When the identity threshold rises up, the two rate curves become nearly coincident. This illustrated that the assembled reads and nonassembled reads may have similar correct alignment contigs rate on quantity.

3.2. SNPs Calling. It is widely believed that SNP could be identified in human gene if the frequency of second base was above 1%. But there is no uniform criterion in plant gene. People who study on crops usually select the threshold value of SNP screening by experience or according to actual condition. If the threshold was too low, some false SNPs would be selected; if the threshold was high, a few of real SNPs could not be completely selected. When threshold was set at 1% in our study, there were too many SNPs in some genes, such as *WCOR14*, *LEA1*, and *LEC1*; the rate of SNP in each reference gene almost reached 40%. That is not reasonable.

To find better threshold value for SNP screening in this study, we set the threshold from 1% to 20% by 1% increase and plotted the curves of *ACC1*, *PhyC*, and *Q* genes' SNP rate variation trend (Figure 4(a)). Because *ACC1*, *PhyC*, and *Q* genes were known as stable genes, we use their SNPs rate variation curve as reference to analysis curve character and find out the best second code rate threshold. We could clearly find that the SNPs rate by nonassembled reads was higher than SNPs rate by assembled reads. Meanwhile the SNPs rate curve of nonassembled keeps descending most of the time, but the assembled reads SNPs curve becomes relatively stable when the second code rises up to 4%. It revealed that nonassembled reads will bring in more SNPs. Most contigs aligned to the function genes in nonassembled reads database were over 80 bp with some low quality locus at the end of the sequence. But the assembled reads have about 50 bp overlap locus at average. When two reads assembled into a sequence, those overlap loci with different codes from two reads were set as "N" (Figure 2) and will not be statistic as potential SNPs. That is why assembled reads are of higher quality than nonassembled reads. So we suggest using assembled sequence to get accurate SNPs or MSV information.

In our study, to find the best MAF to screen SNPs, we only discuss the results from assembled reads database. All the solid lines descended fast in the beginning and become stable

Table 3: The sequence variation information of functional gene by 6% MAF.

Name	Length	Assembled reads		Nonassembled reads		Pretrimmed reads		Original reads	
		Number	Ratio	Number	Ratio	Number	Ratio	Number	Ratio
ABA8OH	654	5	0.76%	19	2.91%	13	1.99%	20	3.06%
ABI5	1540	159	10.32%	171	11.10%	145	9.42%	152	9.87%
ACC1	1131	10	0.88%	14	1.24%	10	0.88%	16	1.41%
APX	1354	97	7.16%	86	6.35%	83	6.13%	96	7.09%
DRF	963	42	4.36%	39	4.05%	41	4.26%	46	4.78%
EMH5	443	21	4.74%	29	6.55%	20	4.51%	27	6.09%
ERD4	810	46	5.68%	44	5.43%	42	5.19%	45	5.56%
FUC3	564	18	3.19%	13	2.30%	20	3.55%	22	3.90%
GSK	527	9	1.71%	9	1.71%	15	2.85%	20	3.80%
HKT8	866	51	5.89%	36	4.16%	44	5.08%	56	6.47%
LEA1	816	14	1.72%	44	5.39%	44	5.39%	61	7.48%
LEC1	910	92	10.11%	90	9.89%	83	9.12%	110	12.09%
PhyC	934	19	2.03%	30	3.21%	31	3.32%	36	4.07%
Q	809	4	0.49%	28	3.46%	25	3.09%	35	4.20%
WDAI	446	44	9.87%	47	10.54%	38	8.52%	34	7.62%
ZCCT1	669	28	4.19%	34	5.08%	21	3.14%	28	4.19%

after 4% or 5% second code rate (Figure 4). The severe decline may be caused by the sequencing precision. To eliminate the problem by sequencing quality reasonably, selecting an appropriate threshold is more significant. Polynomial fitting method was used to fit the curve to get more information about the curve variation rate. After examination, the 6-order polynomial turned out to be the best one to fit the curves. Then we computed first-order differential of the fitted equation and got the curve variation equations. From derivation equation curve (Figure 4), it showed us the acceleration of SNPs rate descent. When the acceleration became near 0, there were few variations in the initial curve. It means that the rate of SNPs will remain unchanged when the threshold rises up. According to Figure 4, we chose 6% as the second threshold in our study. In future research, the new MAF threshold should be calculated based on the new sequence result.

As designed, the assembled reads have high quality and when they are aligned to reference genes, they will perform more quality than others reads. Here we compared the castoff length while reads aligned to sequence with nonassembled reads, assembled reads, pretrimmed reads, and original reads. The pretrimmed reads were original reads cut by the end of 20 bp before being used to align to reference. Original reads came from the sequence result without any process. It declared that most reads were zero-cut in the process of alignment (Figure 5). But the assembled reads have more proportion of zero-cut; over 65% reads were zero-cut. Obviously the nonassembled reads have the longest length cut than the other three reads, which illustrated that the reads that cannot be assembled from original reads were of lower quality than the reads that can be assembled. Consequently, if we just use the part of assembled reads for SNPs, we could get more accurate result.

There are not as much reads as pretrimmed and original reads in assembled database. The overlaps of each gene from assembled reads were lower than other two databases (Figure 6). But in assembled reads database the lowest overlap in *Q* gene still exceeds 100. Although the number of assembled reads is not as much as others, it still has a reliable overlap. We can see that the average overlap of each gene is not homogeneous; *PhyC* gene had 341.83 overlaps, *ACC1* gene 793.03, and *Q* gene 1764.03. That is because the PCR samples concentration we mixed was not under the same uniformity. To get more average overlap, the sample concentration should be as equal as possible.

The advantage of assembled reads in SNPs analysis is that they perform more accurately. In Table 3, there were

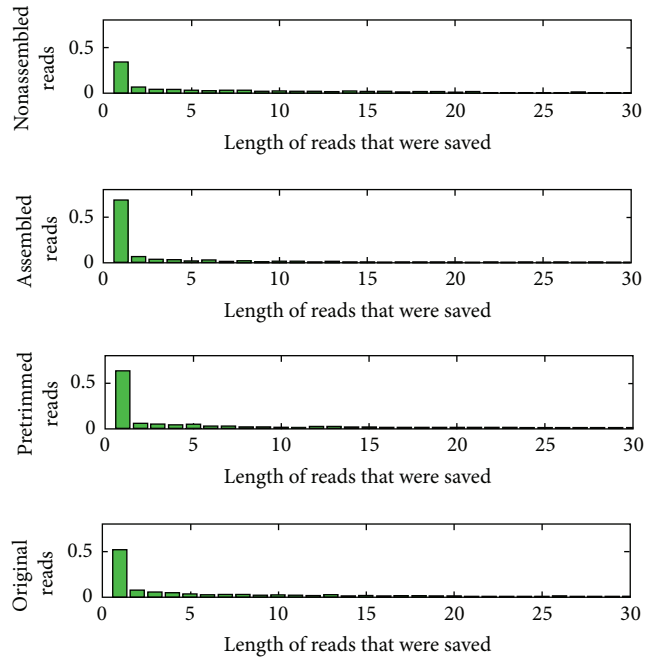


Figure 5: Proportions of reads were trimmed by different length. The *x*-axis was the lengths of reads which were trimmed by local blast algorithm. The *y*-axis was the proportion of each trimmed length. The less the length was trimmed the less the low quality parts the reads have.

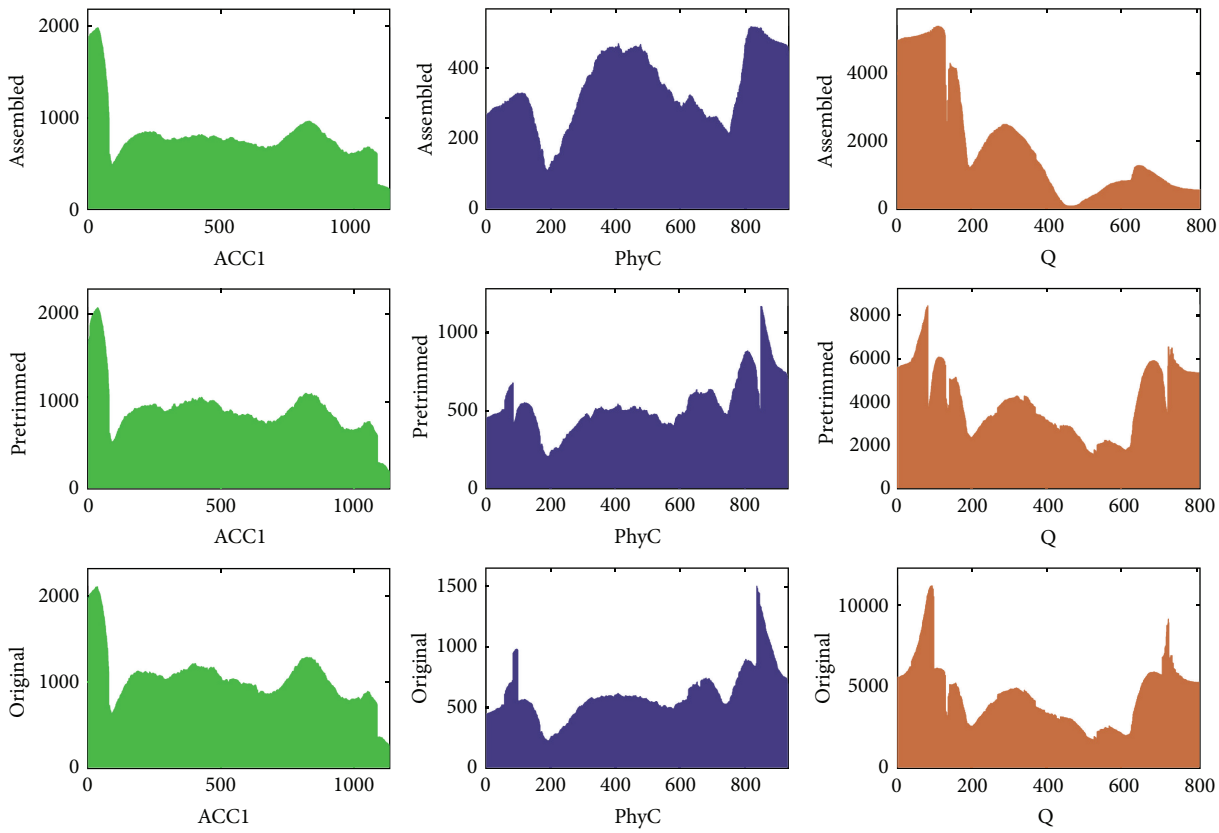


FIGURE 6: Bar chart of genes locus overlaps by contigs mapping. In each subgraph, the x-axis was the whole function gene locus; the y-axis was the total number of contigs on each locus.

SNPs from the main stable genes we discussed before. By the same MAF threshold (6%), *ACC1* gene had 10 SNPs from assembled and pretrimmed reads database and had 16 SNPs when aligned by original reads, but in *PhyC* and *Q* gene, less SNPs were screened by assembly. The quality of reads will determine the reliability of SNPs. As original reads have low sequence quality at the end of 15 bp, the pretrimmed reads will surely have high sequence quality and alignment quality. The high-quality reads could avoid bringing too much false SNPs and be aligned to reference more accurate. The SNPs of each gene screened by pretrimmed reads and assembled reads were all overlapped with SNPs from original reads (Figure 7(a)). It is as estimated that assembled and pretrimmed reads will screen less SNPs than original reads. Form the SNPs relationship diagram we can find that most SNPs in assembled reads were overlapped with pretrimmed reads. Only one SNP of *ACC1* gene was not matched. Then we checked that the unmatched SNPs were at 80th (assembled) and 387th (pretrimmed) loci. At the 80th locus, main code was C and minor one is T. The proportion of T from assembled reads was more than that from both original and pretrimmed (Figure 7(b)). Judging from the result of sequencing, different reads had different sequence quality at the same locus, which caused gravity of code skewing to main code. But we set the mismatched locus as "N" without considering the gravity of code when we assembled reads. In that way, the skewing of main code gravity whose low sequence reads brought in was relieved and allowed us to use high-quality reads to get accurate SNPs. At the 387th locus, the proportion of minor code decreased progressively from original to assembled reads. Based on our design ideas, the decrease of minor code proportion may be caused by high-quality reads which we used to align to reference.

We marked all the SNPs from the assembled and nonassembled reads on the genes (Figure 8). There was large amount of distributed SNPs which only discovered in nonassembled reads (orange color) even in stable genes *ACC1*, *PhyC*, and *Q*. Many of them may be false SNPs because of the low quality reads. SNPs markers only from assembled reads (green color) were less than those from nonassembled. It was proved that the reads with higher quality could be assembled easier than that without enough quality. We suggest discarding the reads that could not be assembled when using this method to mine SNPs for getting more reliable information.

The blue and green markers were the final SNPs position tags we found in this study. There were incredible quantities of SNPs in some genes (Figure 8). As wheat was one of organics which have the most complex genome, it has a large genome size and a high proportion of repetitive elements (85–90%) [14, 15]. Many duplicate SNPs may be nothing more than paralogous sequence variants (PSVs). Alternatively,

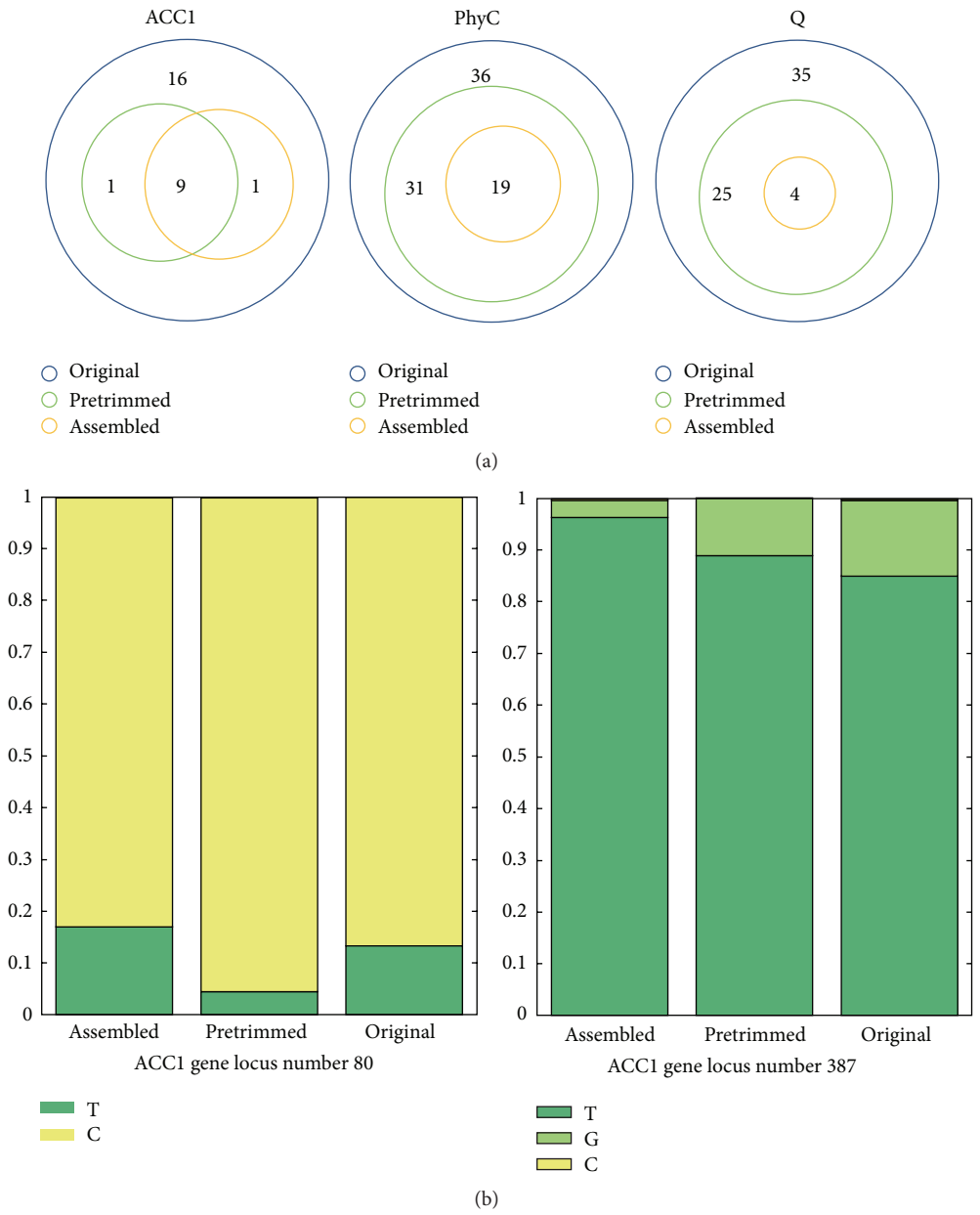


Figure 7: Relationship diagram of SNPs from different reads mapping. (a) The relationship of the SNPs calculated by different data in each gene. (b) The base frequency on mismatch SNPs locus. Different color means different code. The y-axis was the proportion.

gene conversion in duplicate may generate allelic diversity. So the SNPs in our result could be explained as the PSVs or polymorphism multisite variation (MSV) [16, 17].

4. Conclusion

As the high throughput next-generation sequence technology is progressing almost every year, more long read sequence will be brought to us, such as PacBio that will make more easy way for calling SNPs in nonreference species [18]. Particularly for plants with large and complex genome, more long and accurate technology will be helpful in calling SNP [19, 20] (what a pity that PacBio is still a very high-cost way compared to Illumina system). This study aims at finding an efficient and flexible pipeline to mine SNPs with low cost for function genes of nonmodel plant. In outline, our strategy is to mix as much DNA samples as we required and sequence by one run and then use assembled reads to make database for mapping by local blast algorithm computational tools and meanwhile utilize function gene sequence as reference and finally analyze the resulting genotyping data and screen SNPs. The result demonstrated that several function genes of nonmodel plants can be molecular-cloned, mixed to sequence, and analyzed after being assembled and aligned. The assembled reads performed more accurately than the trimmed reads when they are aligned to references (functional genes). Utilizing

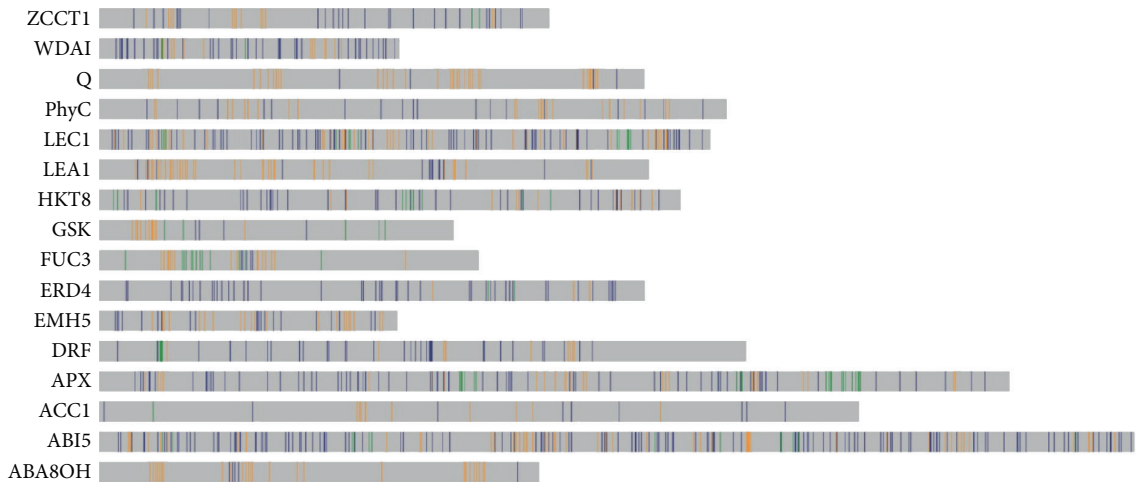


Figure 8: The position of SNPs on the gene. Comparison of SNPs position of the assembled reads and nonassembled reads. The vertical bars were the potential SNPs locus. The green bars form assembled reads, the orange bars form nonassembled reads, and the blue bars belonged to both assembled and nonassembled reads.

polynomial fitting and differential equation to find the best MAF threshold is more reasonable.

Competing 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 Basic Research Program of China (2014CB147200) and the National Natural Science Foundation of China (31171555 and 31571654). Zhong-Xu Chen was supported by the Miaozi Project from Science & Technology Department of Sichuan Province (2015RZ0026).

References

[1] E. R. Mardis, "The impact of next-generation sequencing technology on genetics," *Trends in Genetics*, vol. 24, no. 3, pp. 133–141, 2008.

[2] N. Whiteford, N. Haslam, G. Weber et al., "An analysis of the feasibility of short read sequencing," *Nucleic Acids Research*, vol. 33, no. 19, article e171, 2005.

[3] S. C. Schuster, "Next-generation sequencing transforms today's biology," *Nature Methods*, vol. 5, no. 1, pp. 16–18, 2008.

[4] D. R. Zerbino and E. Birney, "Velvet: algorithms for de novo short read assembly using de Bruijn graphs," *Genome Research*, vol. 18, no. 5, pp. 821–829, 2008.

[5] J. Butler, I. MacCallum, M. Kleber et al., "ALLPATHS: de novo assembly of whole-genome shotgun microreads," *Genome Research*, vol. 18, no. 5, pp. 810–820, 2008.

[6] J. T. Simpson, K. Wong, S. D. Jackman, J. E. Schein, S. J. M. Jones, and I. Birol, "ABySS: a parallel assembler for short read sequence data," *Genome Research*, vol. 19, no. 6, pp. 1117–1123, 2009.

[7] R. Schmieder and R. Edwards, "Quality control and preprocessing of metagenomic datasets," *Bioinformatics*, vol. 27, no. 6, Article ID btr026, pp. 863–864, 2011.

[8] R. K. Patel and M. Jain, "NGS QC toolkit: a toolkit for quality control of next generation sequencing data," *PLoS ONE*, vol. 7, no. 2, Article ID e30619, 2012.

[9] D. Blankenberg, A. Gordon, G. Von Kuster et al., "Manipulation of FASTQ data with galaxy," *Bioinformatics*, vol. 26, no. 14, pp. 1783–1785, 2010.

[10] Illumina Technology, http://www.illumina.com/techniques/sequencing.html.

[11] A. Ratan, Y. Zhang, V. M. Hayes, S. C. Schuster, and W. Miller, "Calling SNPs without a reference sequence," *BMC Bioinformatics*, vol. 11, article 130, 2010.

[12] F. M. You, N. Huo, K. R. Deal et al., "Annotation-based genome-wide SNP discovery in the large and complex *Aegilops tauschii* genome using next-generation sequencing without a reference genome sequence," *BMC Genomics*, vol. 12, article 59, 2011.

[13] 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.

[14] R. B. Flavell, M. D. Bennett, J. B. Smith, and D. B. Smith, "Genome size and the proportion of repeated nucleotide sequence DNA in plants," *Biochemical Genetics*, vol. 12, no. 4, pp. 257–269, 1974.

[15] M. Trick, N. M. Adamski, S. G. Mugford, C.-C. Jiang, M. Febrer, and C. Uauy, "Combining SNP discovery from next-generation sequencing data with bulked segregant analysis (BSA) to fine-map genes in polyploid wheat," *BMC Plant Biology*, vol. 12, article 14, 2012.

[16] X. Estivill, J. Cheung, M. A. Pujana, K. Nakabayashi, S. W. Scherer, and L.-C. Tsui, "Chromosomal regions containing high-density and ambiguously mapped putative single nucleotide polymorphisms (SNPs) correlate with segmental duplications in the human genome," *Human Molecular Genetics*, vol. 11, no. 17, pp. 1987–1995, 2002.

[17] D. Fredman, S. J. White, S. Potter, E. E. Eichler, J. T. Den Dunnen, and A. J. Brookes, "Complex SNP-related sequence

variation in segmental genome duplications," *Nature Genetics*, vol. 36, no. 8, pp. 861–866, 2004.

[18] R. J. Roberts, M. O. Carneiro, and M. C. Schatz, "The advantages of SMRT sequencing," *Genome Biology*, vol. 14, no. 7, p. 405, 2013.

[19] J. Huddleston, S. Ranade, M. Malig et al., "Reconstructing complex regions of genomes using long-read sequencing technology," *Genome Research*, vol. 24, no. 4, pp. 688–696, 2014.

[20] A. Rhoads and K. F. Au, "PacBio sequencing and its applications," *Genomics, Proteomics & Bioinformatics*, vol. 13, no. 5, pp. 278–289, 2015.