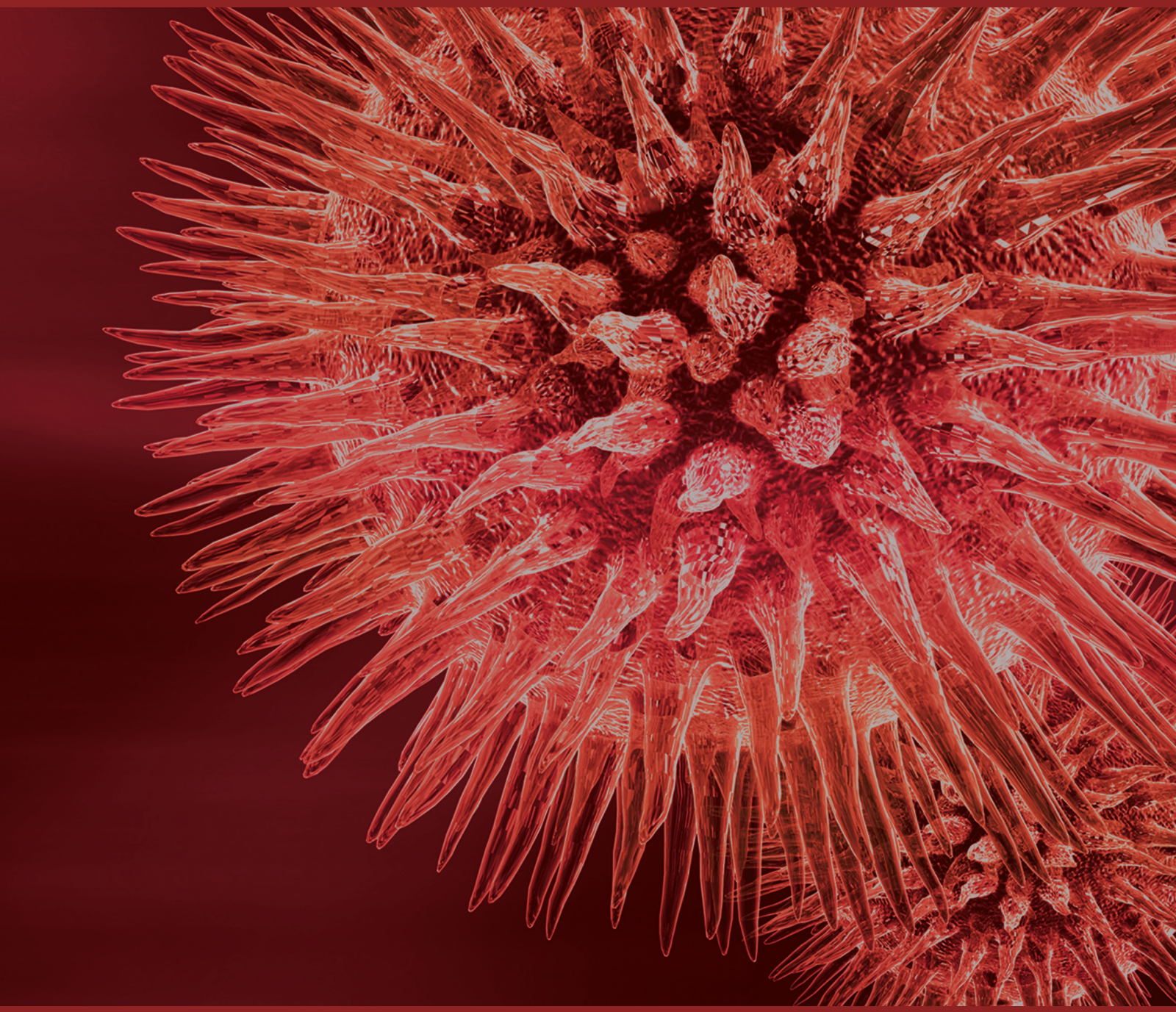
BioMed Research International

Functional Genomics, Genetics, and Bioinformatics

Guest Editors: Youping Deng, Hongwei Wang, Ryuji Hamamoto, David Schaffer, and Shiwei Duan



Functional Genomics, Genetics, and Bioinformatics

BioMed Research International

Functional Genomics, Genetics, and Bioinformatics

Guest Editors: ouping Deng, Hongwei Wang, Ryuji Hamamoto, David Schaffer, and Shiwei Duan

Contents

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 184824, 3 pages
http://dx.doi.org/10.1155/2015/184824

Editorial
Functional Genomics, Genetics, and Bioinformatics

Youping Deng,[1] Hongwei Wang,[2] Ryuji Hamamoto,[2] David Schaffer,[3] and Shiwei Duan[4]

[1]*Department of Internal Medicine, Rush University Cancer Center, Rush University Medical Center, Chicago, IL 60612, USA*
[2]*Department of Medicine, University of Chicago, Chicago, IL 60637, USA*
[3]*Department of Bioengineering, Binghamton University, Binghamton, NY 13902, USA*
[4]*School of Medicine, Ningbo University, Ningbo, Zhejiang 315211, China*

Correspondence should be addressed to Youping Deng; youping_deng@rush.edu

Received 10 December 2014; Accepted 10 December 2014

Biology has become the land of the "-omics," including genomics [1], transcriptomics [2, 3], epigenomics [4], proteomics [5], lipidomics [6, 7], and metabolomics [8]. Each of these "-omics" generates a huge amount of high-throughput data, and it is a challenge both to analyze these data and to further investigate the function of specific molecules. Though more genomes have been completed due to the rapid development of sequencing technology [9], we cannot understand the information contained within a genome until we mine out its implicated functions including downstream transcription, translation, epigenetics modulation, and metabolic pathways. In this special issue, we mainly focus on functional "-omics" and bioinformatics.

The Peer-reviewed papers are collected in the special issue. They are approximately divided into three areas: bioinformatics, functional genomics, and functional genetics. The majority of the papers are purely bioinformatics related papers. We define bioinformatics papers as those using computational tools or developing methods to analyze functional "-omics" data without using wet labs. Two papers fell into the category of functional gen-omics, which is focused on using whole genome level wet-lab technology to find important molecules and investigate their potential functions. Five papers are considered as functional genetics papers. Functional genetics is a broad concept here and these papers are concentrated on studying the molecular functions and mechanisms of individual molecules using wet-lab experimental approaches.

Bioinformatics. In the bioinformatics papers, four papers deal with transcriptomics data. F. Wang et al. developed a novel approach for coexpression analysis of E2F1-3 and MYC target genes in chronic myelogenous leukemia (CML); they found a significant difference in the coexpression patterns of those candidate target genes between the normal and the CML groups. It is challenging to analyze the quantity of image data on gene expression. A. Shlemov et al. developed a method called 2D singular spectrum analysis (2D-SSA) for application to 2D and 3D datasets of embryo images related to gene expression; it turned out to work pretty well. J. Li et al. characterized putative *cis*-regulatory elements (CREs) associated with male meiocyte-expressed genes using *in silico* tools. They found that the upstream regions (1 kb) of the top 50 genes preferentially expressed in *Arabidopsis* meiocytes possessed conserved motifs, which were potential binding sites of transcription factors. NAGNAG alternative splicing plays an important role in biological processes and represents a highly adaptable system for posttranslational regulation of gene function. Interestingly, X. Sun et al. identified about 31 NAGNAG alternative splicing sites that were identified in human large intergenic noncoding RNAs (lincRNAs).

Three papers are focused on the deification of new gene family members and gene evolution. Conotoxins are small disulfide-rich neurotoxic peptides, which can bind to ion channels with very high specificity and regulate their activities. H. Ding et al. developed a novel method called iCTX-Type, which is a sequence-based predictor that can be used to

identify the types of conotoxins in targeting ion channels. A user-friendly web tool is also available. Y.-Z. Zhou et al. analyzed the evolution pattern and function diversity of PPAR gene family members based on 63 homology sequences of PPAR genes from 31 species. They found that gene duplication events, selection pressures on HOLI domain, and the variants on promoter and 3′UTR are critical for PPARs evolution and acquiring diversity functions. There has recently been considerable focus on its two human pathogenic species *N. meningitidis* and *N. gonorrhoeae*, which belong to *Neisseria*, a genus of gram-negative bacteria. D. Yu et al. selected 18 *Neisseria* genomes, preformed a comparative genome analysis, and identified 635 genes with recombination signals and 10 genes that showed significant evidence of positive selection. Further functional analyses revealed that no functional bias was found in the recombined genes. The data help us to understand the adaptive evolution in *Neisseria*.

One paper tried to solve the key algorithm issue called the all-pairs suffix-prefix matching problem, which is crucial for de novo genome assembly. M. H. Rachid et al. developed a space-economical solution to the problem using the generalized Sadakane compressed suffix tree. One paper conducted a comparative genomics analysis. R. Cecagno et al. found that the versatile gene repertoire in the genome of rhizosphere bacterium *Azospirillum amazonense* could have been acquired from distantly related bacteria from horizontal transfer. They also demonstrated that the coding sequence related to production of phytohormones, such as flavin monooxygenase and aldehyde oxidase, is likely to represent the tryptophan-dependent TAM pathway for auxin production in this bacterium. They conclude that the genomic structure of the bacteria has evolved to meet the requirement for adaptation to the rhizosphere and interaction with host plants.

One article conducted a meta-analysis. H. Ye et al. have demonstrated that rs2228671 is a protective factor of CHD in Europeans. One paper is concentrated on the microorganism bioinformatics. Y. Ding et al. recognized the roles of the synonymous codon usage in the formation of nsp1α structure of porcine reproductive and respiratory syndrome virus PRRSV.

Functional Genomics. There are two papers that conducted gene association studies based on genome wide data. J. Li et al. found that the presence of ATTε4haplotype was associated with an increased risk of mental retardation (MR) in children but did not find any significant association between single loci of the four common ApoE polymorphisms (−491A/T, −427T/C, −219T/G, and ε2/3/4) and MR or borderline MR. J. Zhou et al. did not find an association between rs7529229 and chronic heart disease (CHD) in Han Chinese. However, their meta-analyses indicated that rs7529229 was associated with the CHD risk in Europeans.

Functional Genetics. There are 5 articles that investigate the individual gene function in different areas. Two papers are related to neural diseases. G.-M. Chang et al. found that activating NF-κB signaling pathway can protect intestinal epithelial cell No. 6 against fission neutron irradiation. X.-S. Liu et al. demonstrated that hepatocyte growth factor (HGF) could promote the regeneration of damaged Parkinson's disease (PD) cells at higher efficacy than the supernatant from hUC-MSCs alone. Thus, the combination of hUC-MSC with HGF could potentially be a new biological treatment for PD. One paper is focused on cancer. N. Ji et al. found that celastrol had antiprostate cancer effects partially through the downregulation of the expression level of hERG channel in DU145 cells, suggesting that celastrol may be a potential agent against prostate cancer with a mechanism of blocking the hERG channel. One paper is studying heart disease. Z. Lu et al. reported that the levels of NT-proBNP and CCR were closely related to the occurrence of HF and were independent risk factors for heart failure (HF). Meanwhile, there was a significant negative correlation between the levels of NT-proBNP and CCR. One interesting paper is trying to understand the function of Japanese encephalitis virus (JEV), and they have demonstrated that RNA recombination in JEV occurs unequally in different cell types. They conclude that the adjustment of viral RNA to an appropriately lower level in mosquito cells prevents overgrowth of the virus and is beneficial for cells to survive the infection.

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

Youping Deng
Hongwei Wang
Ryuji Hamamoto
David Schaffer
Shiwei Duan

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, article S16, 2013.

[2] 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.

[3] H. Jiang, Y. Deng, H.-S. Chen et al., "Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes," *BMC Bioinformatics*, vol. 5, article 81, 2004.

[4] 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.

[5] F. Xu, G. Li, C. Zhao et al., "Global protein interactome exploration through mining genome-scale data in Arabidopsis thaliana," *BMC Genomics*, vol. 11, article S2, supplement 2, 2010.

[6] Y. Wang, C. Zhao, J. Mao et al., "Integrated lipidomics and transcriptomic analysis of peripheral blood reveals significantly enriched pathways in type 2 diabetes mellitus," *BMC Medical Genomics*, vol. 6, no. 1, article S12, 2013.

[7] X. Zhou, J. Mao, J. Ai et al., "Identification of plasma lipid biomarkers for prostate cancer by lipidomics and bioinformatics," *PLoS ONE*, vol. 7, no. 11, Article ID e48889, 2012.

[8] T. W.-M. Fan, A. N. Lane, and R. M. Higashi, "The promise of metabolomics in cancer molecular therapeutics," *Current Opinion in Molecular Therapeutics*, vol. 6, no. 6, pp. 584–592, 2004.

[9] F. Wang, L. Lu, C. Yu et al., "Development of a novel DNA sequencing method not only for hepatitis B virus genotyping but also for drug resistant mutation detection," *BMC Medical Genomics*, vol. 6, no. 1, article S15, 2013.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 613910, 11 pages
http://dx.doi.org/10.1155/2015/613910

Research Article

Evolutionary Pattern and Regulation Analysis to Support Why Diversity Functions Existed within PPAR Gene Family Members

Tianyu Zhou,[1] Xiping Yan,[1] Guosong Wang,[1] Hehe Liu,[1,2] Xiang Gan,[1] Tao Zhang,[1] Jiwen Wang,[1] and Liang Li[1]

[1]*Key Lab of Sichuan Province, Institute of Animal Genetics and Breeding, Sichuan Agricultural University, Ya'an, Sichuan 625014, China*
[2]*College of Animal Science and Technology, Sichuan Agricultural University, Ya'an, Sichuan 625014, China*

Correspondence should be addressed to Hehe Liu; liuee1985@gmail.com

Received 30 June 2014; Accepted 4 November 2014

Academic Editor: Ryuji Hamamoto

Peroxisome proliferators-activated receptor (PPAR) gene family members exhibit distinct patterns of distribution in tissues and differ in functions. The purpose of this study is to investigate the evolutionary impacts on diversity functions of PPAR members and the regulatory differences on gene expression patterns. 63 homology sequences of PPAR genes from 31 species were collected and analyzed. The results showed that three isolated types of PPAR gene family may emerge from twice times of gene duplication events. The conserved domains of HOLI (ligand binding domain of hormone receptors) domain and ZnF_C4 (C4 zinc finger in nuclear in hormone receptors) are essential for keeping basic roles of PPAR gene family, and the variant domains of LCRs may be responsible for their divergence in functions. The positive selection sites in HOLI domain are benefit for PPARs to evolve towards diversity functions. The evolutionary variants in the promoter regions and 3′ UTR regions of PPARs result into differential transcription factors and miRNAs involved in regulating PPAR members, which may eventually affect their expressions and tissues distributions. These results indicate that gene duplication event, selection pressure on HOLI domain, and the variants on promoter and 3′ UTR are essential for PPARs evolution and diversity functions acquired.

1. Introduction

Peroxisome proliferators-activated receptors (PPARs) are transcription factors belonging to the ligand-activated nuclear receptor superfamily, which play key roles in regulating metabolism, inflammation, and immunity. In vertebrates, the gene family of PPAR consisted of PPARα, PPARβ (also called PPARb/d or PPARδ), and PPARγ [1]. Recently, a considerable number of papers have reviewed their importance in functions within various physiological and biochemistry processes [2–5]. Their special effects and functional manners of depending on a ligand-activated way even have attracted some scientists to consider them as a drug target for therapy of some metabolic disorders, such as the type 2 diabetes mellitus and atherosclerosis [6].

It has been well established that the PPARs can be divided into five distinct functional regions, which include DBD (DNA-binding domain), LBD (ligand-binding domain), AF1 (activation function 1), AF2 (activation function 2), and a variable hinge region. The DBD and LBD consist of a highly conserved DNA-binding domain and a moderately conserved ligand-binding domain, respectively. The AF1 and AF2 are two ligand-independent activation function domains. All these regions except the variable hinge region are highly conserved among PPAR members and are responsible for keeping their functions [3]. Although the PPARs share high similarities with each other in structures, they exhibit distinct patterns of distribution in tissues and differ in functions [7]. It has been summarized that PPARα mainly is involved in the oxidation process of hepatocytes, PPARβ mainly targets within the adipocyte proliferation, and PPARγ plays essential roles in origination and fate determination of preadipocyte. In adult rat, it has shown that PPARs had different expression patterns [8]. Definitely, PPARα is highly expressed in

hepatocytes, cardiomyocytes, enterocytes, and the proximal tubule cells of kidney, PPARβ is expressed ubiquitously and often at higher levels than PPARα and PPARγ, and PPARγ is expressed predominantly in adipose tissue and the immune tissues [4].

It is interesting to investigate why PPARs exhibit distinct patterns of distribution in tissues and differ in functions even if they share high similarity of regions. There may be at least two main aspects of molecular reasons accounting for their differences. Firstly, it could be explained by the molecular evolutionary process, for example, the gene duplication event and the selective patterns. PPAR gene family as one of the nuclear hormone receptor (NHR) superfamilies evolves together with other NHR members. It has been demonstrated that a large number of NHR members are likely to result from two waves of gene duplication events. The first wave occurs before the arthropod/vertebrate divergence and has generated the ancestors of the NHR subfamilies, for instance, PPARs, RARs, and RXRs. The second wave of duplication is vertebrate-specific and leads to a diversification inside the subfamilies, with the emergence of the presently known isotypes such as PPARα, PPARβ, and PPARγ [3, 7]. However, it is still unknown which one is the common ancestor gene in PPAR members, and what the impacts of PPARs divergence on their functions are. Secondly, the special transcriptions factors binging in the promoter regions and the miRNAs target at 3′ UTRs of PPARs may be responsible for the distinct patterns of distribution in tissues. Numerous reports have established the basis for gene expression patterns in distribution by predicting and comprising the transcription factors and miRNAs of interested genes [9].

Therefore, in this present study, we took advantage of the availability of gene sequence data to analyze the PPAR gene family based on a view of molecular evolutionary relationship by deducing the possibility of evolution in PPAR gene family, as well as by predicting and comparing their transcription factors and miRNAs to primarily understand the reasons for diversity functions and distinct patterns of expressions in tissues of PPAR members. These analyses may contribute to a comprehensive understanding for the functions of PPAR gene family.

2. Materials and Methods

2.1. PPAR Gene Homology Sequence Collection. The Genomic Blast function (http://blast.ncbi.nlm.nih.gov/Blast.cgi) was used for collecting homologous sequences of PPAR gene family members in species. The parameters were set as the default value. For the minority of the PPAR gene sequences unfound by blast, we separated supplement in the website of Nucleotide database (http://www.ncbi.nlm.nih.gov/nuccore/) by manually using keywords. Through blasting the homology sequences of PPARα, PPARβ, and PPARγ on NCBI, we finally obtained 63 homology sequences that belong to 31 species (Table S1 in Supplementary Material available online at http://dx.doi.org/10.1155/2015/613910). Most of these sequences were from mammals, and a few of them were obtained from fish and birds. These collected sequences were edited and aligned by the MegAlign in DNAStar (Madison, Wisconsin, USA).

2.2. Search for Protein Domains. The open reading frames (ORF) of PPAR sequences in different species were predicted using online software (http://www.ncbi.nlm.nih.gov/gorf/orfig.cgi). Next, these ORF sequences were confirmed by Pfam (http://pfam.sanger.ac.uk/). Only if there were homology amino acid sequences blasted, Pfam would show the ORF sequences being correctly predicted. Furthermore, the correct amino acid sequences were entered into SMART (http://smart.embl-heidelberg.de/) platform for a prediction of protein structure domain.

2.3. Construction of Phylogenetic Tree. The format of each PPAR homologous protein sequence was edited by BioEdit software [10]. Then, the protein sequences were used for constructing phylogenetic tree through a model of maximum likelihood method (ML) by Mega 5.1 [11]. The topological stability of the maximum likelihood tree was evaluated by 1000 bootstrap replications. The Atlantic salmon PPARγ protein sequence (NM_001123546.1) was selected as the outgroup of the protein phylogenetic tree.

2.4. Amino Acid Site Selection Pressure Analysis. The sequences of two conserved protein domains (ZnF_C4 and HOLI domains) were chosen and compared by BioEdit, and then they were classified and merged. According to the analysis of Bayesian tree phylogeny, we used the site model in PAML software package in Codeml program [12] to analyze these two domains.

The site model was constructed to test whether PPAR gene is subjected to positive selection ($\omega > 1$) or negative selection ($\omega < 1$) [13]. This model allows different sites to have different selection pressure, while there is no difference in different branches of the phylogenetic tree. The models named M1a (neutral) and M2a (selection) [13, 14] in the current study were used twice the log-likelihood difference ($2\Delta L$) following χ^2 distribution of likelihood ratio test (LRT), the difference degree of freedom for the two parameters of the model number.

2.5. Analysis of Transcription Factors. By using Gene (http://www.ncbi.nlm.nih.gov/gene/) of the NCBI, the location of the PPAR gene was determined on the chromosome corresponding species. And then, we confirmed the first exon of the PPAR gene transcription initiation site on a chromosome. Sequence about 1000 bp was selected to use as the predicted promoter regions from the upstream of the first exon. On the TRANSFAC, the Alibaba (http://www.gene-regulation.com/pub/programs/alibaba2/index.html) can estimate transcription factor binding sites (TFBS) in unknown DNA sequences.

2.6. Predictions of miRNAs in 3′ UTR Region of PPAR Members. The miRNAs in 3′ UTR region of PPAR members and their regulatory sites were predicted by TargetScan release (http://www.targetscan.org/). In the TargetScan, the 3′ UTR

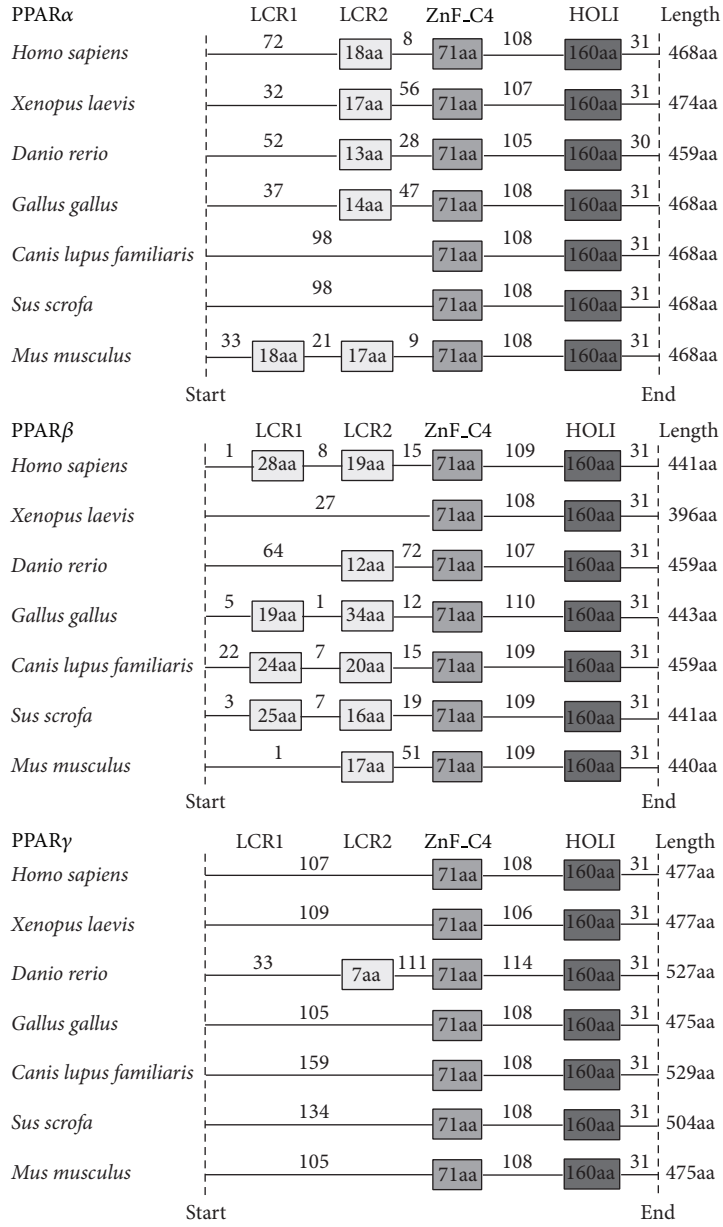


Figure 1: The protein domains of PPARs were predicted in 7 representative species. A box represents a conserved domain. The numerals labeled in the boxes and lines represent the number of amino acid residues. The PPARs coding domain sequences were collected in 7 representative species including human, xenopus, zebrafish, chicken, dog, pig, and mouse.

region of PPAR members of human was searched for miRNAs. The search results were sorted in the miR2Disease Base (http://www.mir2disease.org/) for predicting functions of the predicted conservative miRNAs.

3. Results and Analysis

3.1. The Unique Homology and Conserved Domains in PPAR Gene Family. As it was shown in Table S1, the coding regions of all PPAR nucleotides were in average length of about 1400 bp, which encode about 466 amino acids. The average length of nucleotides of PPARα coding domain is 1406 bp, whereas the average length of nucleotides of PPARβ is 1284 bp which is lower than the average value of the entire PPAR family. The nucleotide of PPARγ is 1479 bp which is obviously higher than the average value.

The protein domains were predicted corresponding to each sequence in the coding region through SMART. The PPAR coding domain sequences in 7 representative species including human, xenopus, zebrafish, chicken, dog, pig, and mouse were obtained for a further analysis (Figure 1). The data demonstrated that all PPARs family members contained

the ZnF_C4 and HOLI domains, which are conserved among species. In addition to the conserved domains, low complexity 1 and low complexity 2 regions (LCRs) were in great differences among PPAR members and species. In PPARα, it was found that LCR2 widely existed in most species, and LCR1 only existed in mice. It is also worth noticing that more than half of the studied species contained the LCRs domains in PPARβ, except for the absence of LCR2 in xenopus. In PPARγ, the LCR2 domain was only found in zebrafish, whereas the LCR1 domain was absent in all studied species.

3.2. The Phylogenetic Tree of PPAR Gene Family. In order to investigate the homologous relationships among PPAR gene family members, we constructed phylogenetic tree based on the amino acid level. The phylogenetic tree was constructed based on the 63 amino acid sequences from 31 species (Table S1), and the results were shown in Figure 2. The orthologs of PPAR members from fishes were placed at the base of the three branches of the tree. Furthermore, the PPAR genes were spitted into three lineages (support value = 100%). Through the branches and distances of the phylogenetic tree, PPARα and PPARβ were clustered together. The branch of PPARγ stood alone and was closer to the outgroup than the other two branches. PPARγ might be the earliest ancestor form of the PPAR gene family. According to the classification, it suggested that the first independent duplication event may occur in bony fishes before separation from the birds and mammals during the whole evolutionary process of PPAR gene family. And after a second duplication event, the isolated types of PPARα and PPARβ may emerge as the paralogs of PPARγ.

3.3. Selection Pressure of Amino Acid Residues in PPAR Gene. To determine the selection states of each amino acid site in conserved structure of PPARs during the evolution process, the tools of selective pressure were used for investigating the different selection patterns based on the conserved motifs of ZnF_C4 domain and HOLI domain, which were widely included and conserved in PPAR gene family. In branch-site models (Table 1), we found the estimated ω value ≥ 1 with the M2a model for HOLI domain and ZnF_C4 domain. It suggests that PPAR genes were under positive selection. By the LRT test, M1a and M2a were compared with their corresponding null models (M0), respectively. The results suggested that M2a ($P < 0.05$) was more in coincidence with the data than M1a ($P > 0.05$). What is more, the LRT tests of all PPAR members were different. The HOLI domain could be accepted by M2a, indicating a positive selection pressure of HOLI domain during the molecular evolution process, whereas the ZnF_C4 domain was rejected.

In a 95% posterior probability, the results (Figures 3(a) and 3(b)) showed that the positive selection sites in PPARα HOLI domain were 118G, 137S, and 143I, in PPARβ HOLI domain 20S, 21S, 58S, and 117P, and in PPARγ HOLI domain 16S and 75G, whereas in the ZnF_C4 domain, there were no positive selection sites observed in all PPAR members, except for only one suspected amino acid residue with ω value between 0.5 and 1 observed in ZnF_C4 domain of PPARα and PPARβ, respectively. In PPARγ ZnF_C4 domain, there were no positive selection sites observed either.

3.4. Prediction of Transcription Factors. The transcription factors and their binding sites in promoter regions of PPAR gene family were predicted in human and chicken, respectively, and the results were listed in Table S2. In chicken, 45, 44, and 39 transcription factors were predicted and targeted at the promoter regions of PPARα, PPARβ, and PPARγ, respectively. In human, only a total of 31, 36, and 40 transcription factors have been predicted at promoter regions of PPARα, PPARβ, and PPARγ, respectively, which were different from it in chicken.

Through comparing transcription factors, we found that numerous common transcription factors existed among PPAR members. Then they were compared pairwise among the three PPAR members, and the results were listed in Table 2. The PPARs shared 9 common transcription factors which were targeted at the promoter regions, including Sp1, CPE_bind, CP1, Oct-1, GATA-1, AP-2α, NF-1, GR, and C/EBPα in human, while in chicken, 11 common transcription factors were predicted and targeted at the promoter regions of chicken PPARs, which included CREB, SRF, ICSBP, Ftz, AP-1, Oct-1, GATA-1, AP-2α, NF-1, GR, and C/EBPα. However, the binding sites for each common transcription factor were varied among PPAR members.

Finally, we quantified the coexisting transcription factors among PPAR members (Table 3). In human, the amount of the identical transcription factors between PPARα and PPARβ was 18, while the amount between PPARβ and PPARγ is 16. The number of identical transcription factors of PPARα and PPARγ was 12. In chicken, the group of PPARα/γ and PPARα/β shared 20 and 15 identical transcription factors, respectively.

3.5. Prediction of miRNAs Target at the 3′ UTR Region of PPAR Members. The miRNAs in 3′ UTR of PPAR members were predicted in human. The results (Table S3) showed that, in the 3′ UTR region of PPARα, a total of 23 conserved binding sites of miRNAs were predicted in vertebrates, and 4 conserved sites of miRNA families were predicted in mammals. In the 3′ UTR region of PPARβ (Figure 4(b)) and PPARγ (Figure 4(c)), 5 and 3 conserved sites of miRNA families were predicted in vertebrates, respectively. Notably, the miR-17 and miR-9 were predicted in both 3′ UTR regions of PPARα and PPARβ, and the miR-27abc and miR-128 were predicted in both 3′ UTR regions of PPARα and PPARγ (Figure 4(a)).

The functions of these miRNAs were enriched in PUBMED online. Among the 27 miRNA families, the vast majority were closely related with cancer. For example, the miR-142-3p [15], miR-19a [16], and miR-124 [17] were reported to be involved in hepatocellular carcinoma; the miR-9 [18] targeting to the 3′ UTR region of PPARα was associated with Hodgkin's lymphoma. In the 3′ UTR region of PPARβ, the miR-138 [19] were reported to be linked to anaplastic thyroid carcinoma; the miR-17 [20] was related to B-cell lymphoma; the miR-29c [21] was interrelated with chronic lymphocytic

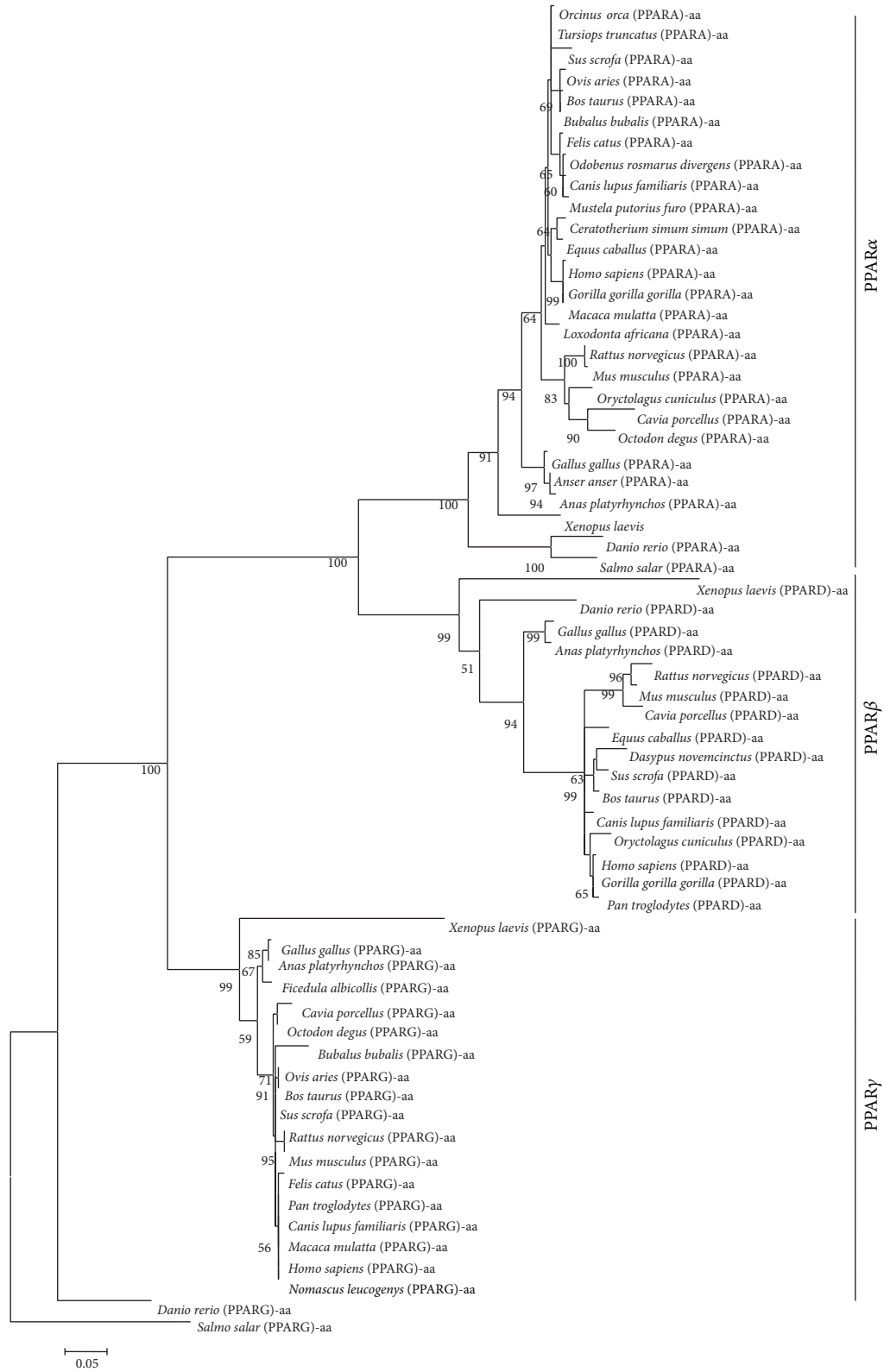


Figure 2: The phylogenetic tree based amino acid sequences. The phylogenetic tree was constructed by amino acid sequences. The sequences information was provided in Table S1. The phylogenetic tree was constructed by the maximum likelihood method with Mega 5.1. The numbers on nodes indicate the support values. It showed the bootstrap values were more than 50%.

Table 1: Selection pressure analysis of amino acid sites in PPARs.

Model	ln$L$	Parameters estimates	2ΔL
Model 0			
α-HOLI	−4720.995579		
β-HOLI	−2960.312353		
γ-HOLI	−2719.120808		
α-ZnF_C4	−1050.374116		
β-ZnF_C4	−1033.465323		
γ-ZnF_C4	−1276.57302		
Model 1a			
α-HOLI	−4622.51301	$P0 = 0.93383, P1 = 0.06617$ $\omega0 = 0.01746, \omega1 = 1.00000$	
β-HOLI	−2894.97063	$P0 = 0.94599, P1 = 0.05401$ $\omega0 = 0.02167, \omega1 = 1.00000$	
γ-HOLI	−2689.49404	$P0 = 0.97407, P1 = 0.02593$ $\omega0 = 0.00539, \omega1 = 1.00000$	
α-ZnF_C4	−1062.29037	$P0 = 0.98590, P1 = 0.01410$ $\omega0 = 0.00862, \omega1 = 1.00000$	
β-ZnF_C4	−1027.39487	$P0 = 0.97625, P1 = 0.02375$ $\omega0 = 0.00722, \omega1 = 1.00000$	
γ-ZnF_C4	−1276.57373	$P0 = 0.99999, P1 = 0.00001$ $\omega0 = 0.00168, \omega1 = 1.00000$	
Model 2a			
α-HOLI	−4622.51301	$P0 = 0.93383, P1 = 0.03245, P2 = 0.03372$ $\omega0 = 0.01746, \omega1 = 1.00000, \omega2 = 1.00000$	196.96513
β-HOLI	−2894.97063	$P0 = 0.94599, P1 = 0.04003, P2 = 0.01398$ $\omega0 = 0.02167, \omega1 = 1.00000, \omega2 = 1.00000$	130.683442
γ-HOLI	−2689.49404	$P0 = 0.97407, P1 = 0.00429, P2 = 0.02163$ $\omega0 = 0.00539, \omega1 = 1.00000, \omega2 = 1.00000$	59.253532
α-ZnF_C4	−1044.04256	$P0 = 0.98590, P1 = 0.01410, P2 = 0.00000$ $\omega0 = 0.00737, \omega1 = 1.00000, \omega2 = 6.14876$	12.66312
β-ZnF_C4	−1027.39487	$P0 = 0.97625, P1 = 0.00871, P2 = 0.01504$ $\omega0 = 0.00722, \omega1 = 1.00000, \omega2 = 1.00000$	12.140902
γ-ZnF_C4	−1276.57302	$P0 = 1.00000, P1 = 0.00000, P2 = 0.00000$ $\omega0 = 0.00168, \omega1 = 1.00000, \omega2 = 1.00000$	0.000004

Note: selection pressure on amino acid sites of the inspection is based on the calculation of dN/dS (ω), where dN is nonsynonymous coding sequences of each base mutation rate (nonsynonymous substitution rate) and dS is a synonymous mutation rate (synonymous substitution rate). When the $\omega > 1$, the gene is by positive selection; $\omega = 1$, no selection pressure; $\omega < 1$, by purifying selection.

leukemia. In the 3′ UTR region of PPARγ, the miR-128 [22] was associated with glioma.

4. Discussions

One new gene is mainly generated by the gene or genome duplication event [23]. PPARs as one of the NHR superfamilies evolve together with other NHR members, and after it has undergone twice time of gene duplication events, the vertebrate-specific PPAR is eventually diverged into three different isotypes [3, 7]. The phylogenetic tree of PPARs in the present study demonstrated that PPAR gene family may have yielded a gene duplication event, which first occurs in bony fishes before separation from the birds and mammals during the whole evolution process. PPARγ is closer to the outgroup than the other two branches, supporting that PPARγ might be the original ancestor gene in PPAR gene family. After being firstly duplicated in fish, PPAR begins to divide into two subtypes, including the PPARγ and the common ancestor of PPARα and PPARβ. These findings are consistent with the previous studies by Michalik et al., which depicted an evolutionary process of PPARs. Moreover, PPARα and PPARβ were clustered closer than others, supporting that they may originate from a homology ancestor gene, and their divergence may result from another gene duplication event in vertebrates; however, there is no sufficient evidence to support this hypothesis currently.

Following the gene duplication event in PPARs, the newly emerging receptors would have acquired the ligand binding capacities in an independent fashion [24]. Once such capacity

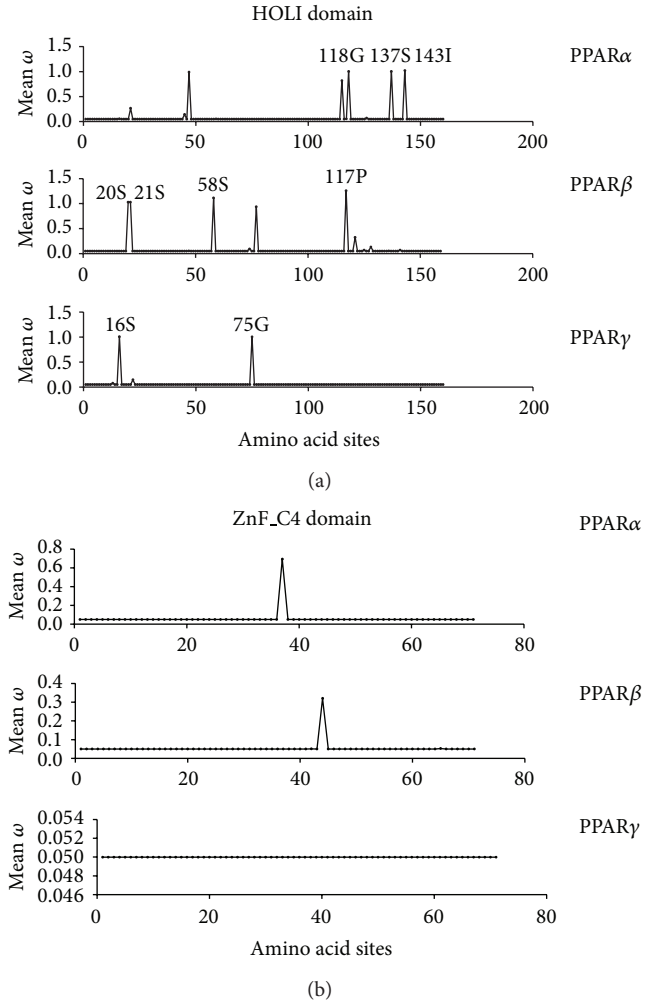


Figure 3: Approximate posterior mean of amino acid sites. There was a list of amino acids in each sequence of the corresponding ω value. The amino acid residue marked on the image represents the $\omega > 1$ with probability of more than 95%.

was acquired, each receptor of PPARs may begin to further evolve and refine its specificity for a given ligand. Each PPAR isotype may then evolve by mutations, which lead to a more specific range of ligands across species. These hypotheses could be supported by the sequence variants among PPARs across species in the present study. Our results showed that all PPAR members contained the conserved HOLI and ZnF_C4 domains, which are important for keeping the functions of PPAR gene family. HOLI domain located in N-terminal of the PPAR protein is also known as ligand binding domain of hormone receptors [25]. It belongs to the LBD region that acts in response to ligand binding, which caused a conformational change in the receptor to induce a response, thereby acting as a molecular switch to turn on transcriptional activity [26]. In addition, ZnF_C4 domain is also called C4 zinc finger in nuclear hormone receptors. This domain was the DBD region, which recognizes specific sequences, connected via a linker region to a C-terminal LBD. Both HOLI and ZnF_C4 domains are highly conserved among PPAR members and are responsible for keeping their basic functions for PPAR family members.

In addition to the two conserved domains, PPAR family contained low complexity regions (LCRs). LCRs located near the left of ZnF_C4 domain are in great differences among PPAR members across species. Studies suggested that the positions of LCRs within a sequence might be important to both determine their binding properties and maintain biological functions [27]. There are no LCRs existing in PPARγ, suggesting that PPARγ might only keep the basic function of PPAR family. The number of LCRs in PPARα and PPARβ is similar and obviously more than PPARγ, indicating differential functions of PPARα and PPARβ from PPARγ. The results showed that the variants in LCRs might be involved in the diversity functions of PPAR members and supported a common origin of PPARα and PPARβ.

Due to the reason that ZnF_C4 and HOLI domain are important for keeping roles of PPAR members, we used patterns of selection pressure to analyze the adaptive evolution of the conserved protein sequences. The results showed that the HOLI domain was selected under a natural pressure in the evolutionary process, whereas the ZnF_C4 domain was not. It showed that ZnF_C4 domain was more conservative than HOLI domain in PPAR family, supporting a more important role of PPAR zinc finger in keeping PPARs' functions [28]. The HOLI domain in PPARβ with the most amounts of positive selection sites among PPAR members suggested that the variations in these positive selection sites were more beneficial for PPARβ phylogenetic towards diversity functions. Studies have confirmed that these chemical properties of amino acid residues were important to sustain normal protein folding and keep functions [29]. For instance, sulfhydryl groups of the peptide chain of two cysteines (cysteine, referred to as S) form two disulfide linkages with oxidation reaction. Whether it breaks or reshapes into a new one, it also could adjust protein to perform certain function [30]. Therefore, it can be inferred that the nucleotide variants in HOLI domain could be responsible for diversity functions of PPAR members. In a 95% posterior probability, the positive selection sites were 118G, 137S, and 143I in PPARα HOLI domain, were 20S, 21S, 58S, and 117P in PPARβ HOLI domain, and were 16S and 75G in PPARγ HOLI domain. It is interesting to point out that the positive selection sites in HOLI domain of PPARα and PPARβ share more similarity in locations and amino acid residues, supporting a homology function of PPARα and PPARβ.

The regulatory mechanism of gene expressions plays an important role in tissue distribution and distinct biological functions of genes. In eukaryotes, most genes are initiated and transcribed by lots of specific transcription factors targeting at their promoter regions [31]. Through predicting the transcription factors and their binding sites in promoter region of PPARs, we found that the transcription factors were varied among PPAR members in human and chicken, which may account for the specific tissue expression and distinct functions of PPARs. Some of these predicted transcription factors and their regulatory effects on PPARs are consistent

Table 2: The common transcription factors predicted in human and in chicken.

Transcription factor	Binding sites and position					
	Chicken (α)	Human (α)	Chicken (β)	Human (β)	Chicken (γ)	Human (γ)
Oct-1	TTAT (−205)	TGCAT (−50)	TTAwTTk (−463)	GCTkT (−737)	AATAT (−18)	AATT (−75)
C/EBPα	TTGA (−62)	GTTGC (−302)	ACAT (−29)	ATCCCA (−23)	ACTC (−71)	TTGC (−192)
AP-2α	GGGG (−84)	GGCyG (−239)	GGCT (−108)	CCCrG (−65)	AGCCTG (−684)	GCCTG (−136)
NF-1	TTTTGG (−457)	TGGCCA (−127)	GCCAA (−140)	TGsC (−15)	TGCCA (−560)	GCCAA (−383)
GR	TGTTCT (−137)	ACAA (−185)	AGAACA (−26)	ACAsA (−123)	ACAG (−128)	AGAAC (−679)
GATA-1	TTAT (−205)	GsATT (−51)	GCAGA (−312)	CwGAT (−175)	AGATA (−58)	CTTATC (−438)
CREB	GTCA (−942)		CGTCA (−941)		ACrTCA (−432)	
SRF	GCCwT (−385)		TTCCGG (−896)		AnATGG (−174)	
ICSBP	GGAAA (−399)		CCCT (−39)		GTTT (−42)	
Ftz	TAAT (−840)		TTAATT (−463)		TAAwTG (−343)	
AP-1	TGAsT (−776)		TCAGC (−556)		TGACTC (−69)	
Sp1		GGAGGG (−12)		GrGG (−38)		TGGG (−139)
CPE_bind		CrTCA (−74)		TGACGT (−968)		CCCC (−876)
CP1		ATTGG (−125)		ATTGG (−913)		AkTGGT (−401)

Table 3: The number of identical transcription factors among PPARs in human and in chicken.

	Human			Chicken		
	PPARα	PPARβ	PPARγ	PPARα	PPARβ	PPARγ
PPARα	—	18	12	—	15	20
PPARβ	18	—	16	15	—	18
PPARγ	12	16	—	20	18	—

with the previous reports; for example, the transcription factors AP1 and NF-kB were proved to enhance the expression of PPARβ activity [32]. Some of these transcription factors are also tissue specific, for example, the SP1 expressed in adenocarcinomas of the stomach [33], CP1 highly expressed in liver, kidney, and intestine but weakly expressed in adrenals and in lactating mammary glands [34, 35], and NF-1 detected in brain, peripheral nerve, lung, colon, and muscle [36], and so forth. It can be speculated that the variants in the promoter regions of PPARα and PPARβ result into differential transcription factors binding on them that eventually influence their expressions and tissues distributions. Additionally, there are 18 common transcription factors between PPARβ and PPARα, whereas the PPARγ shared the least amount of common transcription factors with the other two members, which may contribute to the similarity in expression characteristics between PPARβ and PPARα.

The miRNA can combine with the target mRNA by base pair, which leads to degradation or inhibition of the quantity levels of the target mRNA, thereby regulating gene expressions [37]. The regulation of miRNA on gene expressions is another path shaping gene expression patterns and biological processes [38]. In the present study, the miRNAs and their targets sites in 3′ UTR region of PPARs were predicted, and it was observed that the quantity of miRNAs was obviously differential in PPAR members. The number of miRNAs predicted in PPARα was significantly more than the other two members. Moreover, it was worth noticing that most of the miRNAs were predicted in PPARα, only a minority of them predicted in at least two PPAR isotypes; for example, only miRNA-128 was found in PPARα and PPARγ and miRNA-9 was found in PPARα and PPARβ. These differences may be correlated with the distinct functions of PPAR isotypes, and PPARα may be regulated by miRNAs in a much more complex way than the other two PPARs.

5. Conclusions

In the present study, the evolutionary pattern and regulation characteristics of PPARs were analyzed. The three isotypes of PPAR gene family may emerge from twice times of gene duplication events. PPARγ might be the original ancestor gene in PPAR gene family. The conserved domains of HOLI domain and ZnF_C4 are essential for keeping basic roles of PPAR gene family, and the variant domain of LCRs may be responsible for their divergences in functions. The positive selection sites in HOLI domain are beneficial for PPARs to evolve towards diversity functions. The variants in the promoter regions and 3′ UTR of PPARs resulted into differential transcription factors and miRNAs involved in regulating PPAR members that may eventually influence their expressions and tissue distributions.

Conflict of Interests

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

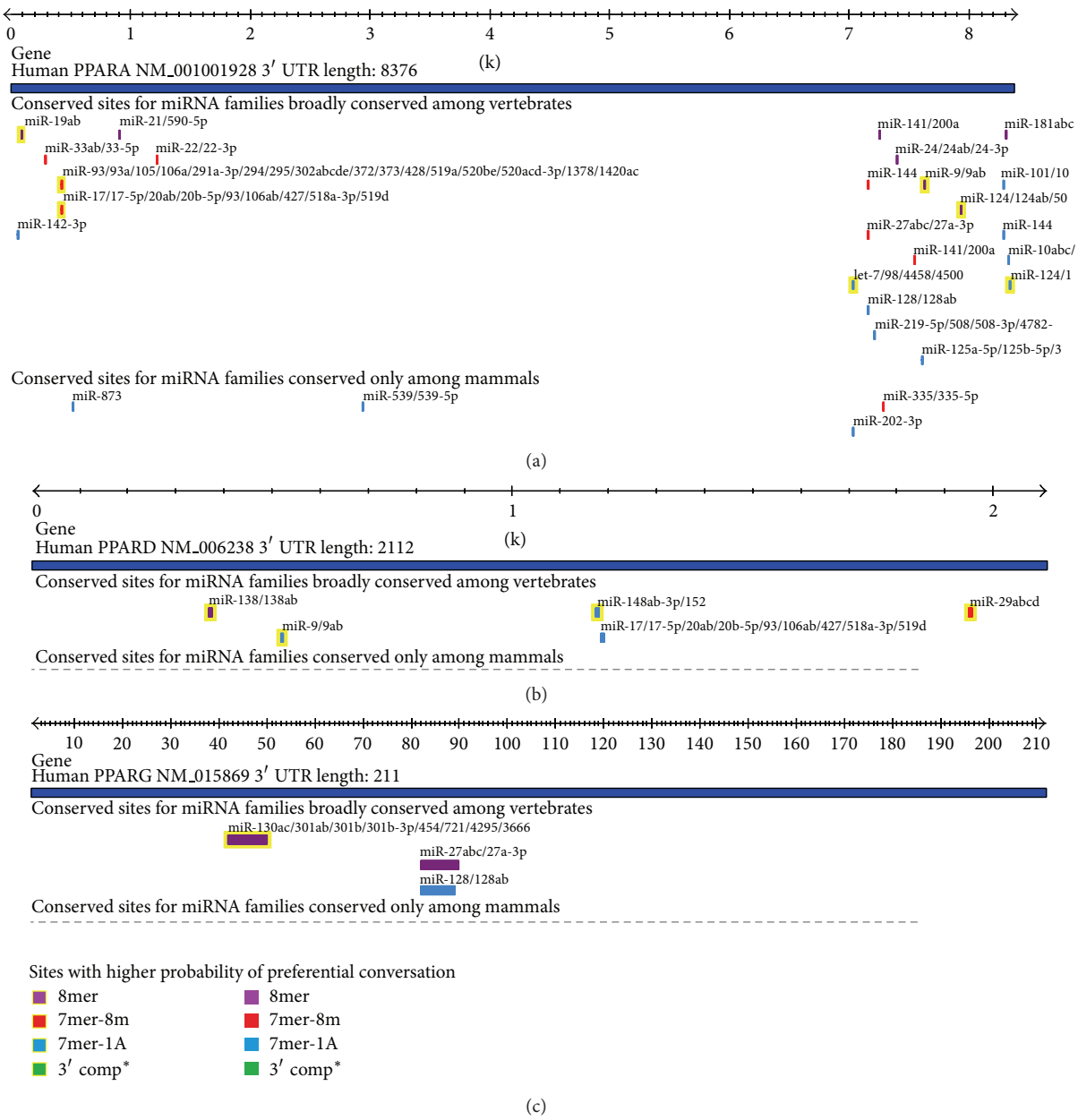


Figure 4: The miRNAs predicted and their targets sites in 3′ UTR region of PPAR genes in human. (a) PPARα; (b) PPARβ; (c) PPARγ. The miRNAs targets sites correspond to the 3′ UTR region of PPAR genes. The lower corner is the probability of preferential conservation for sites.

Authors' Contribution

Tianyu Zhou and Xiping Yan contribute equally as the co-first authors of the paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (no. 31301964), Chinese Agriculture Research Service (no. CARS-43-6), the Major Project of Sichuan Education Department (13ZA0252), and the Breeding of Multiple Crossbreeding Systems in Waterfowl (2011NZ0099-8).

References

[1] P. Tontonoz, E. Hu, and B. M. Spiegelman, "Stimulation of adipogenesis in fibroblasts by PPARγ2, a lipid-activated transcription factor," *Cell*, vol. 79, no. 7, pp. 1147–1156, 1994.

[2] R. B. Clark, "The role of PPARs in inflammation and immunity," *Journal of Leukocyte Biology*, vol. 71, no. 3, pp. 388–400, 2002.

[3] R. A. Daynes and D. C. Jones, "Emerging roles of PPARs in inflammation and immunity," *Nature Reviews Immunology*, vol. 2, no. 10, pp. 748–759, 2002.

[4] Y.-L. Shiue, L.-R. Chen, C.-J. Tsai, C.-Y. Yeh, and C.-T. Huang, "Emerging roles of peroxisome proliferator-activated receptors in the pituitary gland in female reproduction," *Biomarkers and Genomic Medicine*, vol. 5, no. 1-2, pp. 1–11, 2013.

[5] D. Bishop-Bailey and J. Bystrom, "Emerging roles of peroxisome proliferator-activated receptor-β/δ in inflammation," *Pharmacology and Therapeutics*, vol. 124, no. 2, pp. 141–150, 2009.

[6] J. P. Berger, T. E. Akiyama, and P. T. Meinke, "PPARs: therapeutic targets for metabolic disease," *Trends in Pharmacological Sciences*, vol. 26, no. 5, pp. 244–251, 2005.

[7] L. Michalik, B. Desvergne, C. Dreyer, M. Gavillet, R. N. Laurini, and W. Wahli, "PPAR expression and function during vertebrate development," *International Journal of Developmental Biology*, vol. 46, no. 1, pp. 105–114, 2002.

[8] O. Braissant, F. Foufelle, C. Scotto, M. Dauça, and W. Wahli, "Differential expression of peroxisome proliferator-activated receptors (PPARs): tissue distribution of PPAR-α, -β, and -γ in the adult rat," *Endocrinology*, vol. 137, no. 1, pp. 354–366, 1996.

[9] X. Wu, X. Zou, Q. Chang et al., "The evolutionary pattern and the regulation of stearoyl-CoA desaturase genes," *BioMed Research International*, vol. 2013, Article ID 856521, 12 pages, 2013.

[10] T. A. Hall, "BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT," *Nucleic Acids Symposium Series*, vol. 41, pp. 95–98, 1999.

[11] K. Tamura, J. Dudley, M. Nei, and S. Kumar, "MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0," *Molecular Biology and Evolution*, vol. 24, no. 8, pp. 1596–1599, 2007.

[12] Z. Yang, "PAML 4: phylogenetic analysis by maximum likelihood," *Molecular Biology and Evolution*, vol. 24, no. 8, pp. 1586–1591, 2007.

[13] R. Nielsen and Z. Yang, "Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene," *Genetics*, vol. 148, no. 3, pp. 929–936, 1998.

[14] Z. Yang, W. J. Swanson, and V. D. Vacquier, "Maximum-likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites," *Molecular Biology and Evolution*, vol. 17, no. 10, pp. 1446–1455, 2000.

[15] L. Gramantieri, M. Ferracin, F. Fornari et al., "Cyclin G1 is a target of miR-122a, a MicroRNA frequently down-regulated in human hepatocellular carcinoma," *Cancer Research*, vol. 67, no. 13, pp. 6092–6099, 2007.

[16] A. Budhu, H. L. Jia, M. Forgues et al., "Identification of metastasis-related microRNAs in hepatocellular carcinoma," *Hepatology*, vol. 47, no. 3, pp. 897–907, 2008.

[17] M. Furuta, K. I. Kozaki, S. Tanaka, S. Arii, I. Imoto, and J. Inazawa, "miR-124 and miR-203 are epigenetically silenced tumor-suppressive microRNAs in hepatocellular carcinoma," *Carcinogenesis*, vol. 31, no. 5, pp. 766–776, 2009.

[18] K. Nie, M. Gomez, P. Landgraf et al., "MicroRNA-mediated down-regulation of PRDM1/Blimp-1 in Hodgkin/Reed-Sternberg cells: a potential pathogenetic lesion in Hodgkin lymphomas," *American Journal of Pathology*, vol. 173, no. 1, pp. 242–252, 2008.

[19] S. Mitomo, C. Maesawa, S. Ogasawara et al., "Downregulation of miR-138 is associated with overexpression of human telomerase reverse transcriptase protein in human anaplastic thyroid carcinoma cell lines," *Cancer Science*, vol. 99, no. 2, pp. 280–286, 2008.

[20] M. Inomata, H. Tagawa, Y. M. Guo, Y. Kameoka, N. Takahashi, and K. Sawada, "MicroRNA-17-92 down-regulates expression of distinct targets in different B-cell lymphoma subtypes," *Blood*, vol. 113, no. 2, pp. 396–402, 2009.

[21] G. A. Calin, M. Ferracin, A. Cimmino et al., "A MicroRNA signature associated with prognosis and progression in chronic lymphocytic leukemia," *The New England Journal of Medicine*, vol. 353, no. 17, pp. 1793–1801, 2005.

[22] Y. Zhang, T. Chao, R. Li et al., "MicroRNA-128 inhibits glioma cells proliferation by targeting transcription factor E2F3a," *Journal of Molecular Medicine*, vol. 87, no. 1, pp. 43–51, 2009.

[23] V. E. Prince and F. B. Pickett, "Splitting pairs: the diverging fates of duplicated genes," *Nature Reviews Genetics*, vol. 3, no. 11, pp. 827–837, 2002.

[24] H. Escriva, S. Bertrand, and V. Laudet, "The evolution of the nuclear receptor superfamily," *Essays in Biochemistry*, vol. 40, pp. 11–26, 2004.

[25] J. Berger and D. E. Moller, "The mechanisms of action of PPARs," *Annual Review of Medicine*, vol. 53, pp. 409–435, 2002.

[26] D. P. Edwards, "The role of coactivators and corepressors in the biology and mechanism of action of steroid hormone receptors," *Journal of Mammary Gland Biology and Neoplasia*, vol. 5, no. 3, pp. 307–324, 2000.

[27] A. Coletta, J. W. Pinney, D. Y. W. Solís, J. Marsh, S. R. Pettifer, and T. K. Attwood, "Low-complexity regions within protein sequences have position-dependent roles," *BMC Systems Biology*, vol. 4, article 43, 2010.

[28] R. T. Nolte, G. B. Wisely, S. Westin et al., "Ligand binding and co-activator assembly of the peroxisome proliferator- activated receptor-γ," *Nature*, vol. 395, no. 6698, pp. 137–143, 1998.

[29] P. J. Hogg, "Disulfide bonds as switches for protein function," *Trends in Biochemical Sciences*, vol. 28, no. 4, pp. 210–214, 2003.

[30] W. J. Wedemeyer, E. Welker, M. Narayan, and H. A. Scheraga, "Disulfide bonds and protein folding," *Biochemistry*, vol. 39, no. 15, pp. 4207–4216, 2000.

[31] I. Rahman and W. MacNee, "Role of transcription factors in inflammatory lung diseases," *Thorax*, vol. 53, no. 7, pp. 601–612, 1998.

[32] P. Delerive, K. de Bosscher, S. Besnard et al., "Peroxisome proliferator-activated receptor α negatively regulates the vascular inflammatory gene response by negative cross-talk with transcription factors NF-κB and AP-1," *Journal of Biological Chemistry*, vol. 274, no. 45, pp. 32048–32054, 1999.

[33] V. Infantino, P. Convertini, F. Iacobazzi, I. Pisano, P. Scarcia, and V. Iacobazzi, "Identification of a novel Sp1 splice variant as a strong transcriptional activator," *Biochemical and Biophysical Research Communications*, vol. 412, no. 1, pp. 86–91, 2011.

[34] N. Schweifer and D. P. Barlow, "The Lx1 gene maps to mouse Chromosome 17 and codes for a protein that is homologous to glucose and polyspecific transmembrane transporters," *Mammalian Genome*, vol. 7, no. 10, pp. 735–740, 1996.

[35] Y. Alnouti, J. S. Petrick, and C. D. Klaassen, "Tissue distribution and ontogeny of organic cation transporters in mice," *Drug Metabolism and Disposition*, vol. 34, no. 3, pp. 477–482, 2006.

[36] L. B. Andersen, R. Ballester, D. A. Marchuk et al., "A conserved alternative splice in the von recklinghausen neurofibromatosis (NF1) gene produces two neurofibromin isoforms, both of which have GTPase-activating protein activity," *Molecular and Cellular Biology*, vol. 13, no. 1, pp. 487–495, 1993.

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

[38] K. Appasani, *MicroRNAs: From Basic Science to Disease Biology*, Cambridge University Press, Cambridge, Mass, USA, 2008.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 898592, 7 pages
http://dx.doi.org/10.1155/2015/898592

Research Article

The Plant Growth-Promoting Bacteria *Azospirillum amazonense*: Genomic Versatility and Phytohormone Pathway

Ricardo Cecagno,[1] Tiago Ebert Fritsch,[1] and Irene Silveira Schrank[1,2]

[1]*Centro de Biotecnologia, Laboratório de Microrganismos Diazotróficos, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, RS, Brazil*
[2]*Departamento de Biologia Molecular e Biotecnologia, Instituto de Biociências, Universidade Federal do Rio Grande do Sul (UFRGS), CP 15005, 91501-970 Porto Alegre, RS, Brazil*

Correspondence should be addressed to Irene Silveira Schrank; irene@cbiot.ufrgs.br

Received 2 July 2014; Revised 24 October 2014; Accepted 24 October 2014

Academic Editor: You-Ping Deng

The rhizosphere bacterium *Azospirillum amazonense* associates with plant roots to promote plant growth. Variation in replicon numbers and rearrangements is common among *Azospirillum* strains, and characterization of these naturally occurring differences can improve our understanding of genome evolution. We performed an *in silico* comparative genomic analysis to understand the genomic plasticity of *A. amazonense*. The number of *A. amazonense*-specific coding sequences was similar when compared with the six closely related bacteria regarding belonging or not to the *Azospirillum* genus. Our results suggest that the versatile gene repertoire found in *A. amazonense* genome could have been acquired from distantly related bacteria from horizontal transfer. Furthermore, the identification of coding sequence related to phytohormone production, such as flavin-monooxygenase and aldehyde oxidase, is likely to represent the tryptophan-dependent TAM pathway for auxin production in this bacterium. Moreover, the presence of the coding sequence for nitrilase indicates the presence of the alternative route that uses IAN as an intermediate for auxin synthesis, but it remains to be established whether the IAN pathway is the Trp-independent route. Future investigations are necessary to support the hypothesis that its genomic structure has evolved to meet the requirement for adaptation to the rhizosphere and interaction with host plants.

1. Introduction

The genus *Azospirillum* comprises free-living, nitrogen-fixing bacteria that are known as plant growth-promoting rhizobacteria (PGPR), which can colonize, by adhesion, the root surface or the intercellular spaces of the host plant roots. The potential role of the PGPR in association with economically important cereals and other grasses is to promote plant growth by several mechanisms including nitrogen fixation and phytohormone production [1]. Several species of *Azospirillum* are able to secrete phytohormones such as auxins, gibberellins, cytokinins, and nitric oxide as signals of plant growth promotion [2, 3].

Azospirillum genomes, as previously suggested for various strains, are larger and are comprised of multiple replicons indicating a potential for genome plasticity [4]. Genomic rearrangements can occur spontaneously where replicons can be lost upon the formation of new megaplasmids [5, 6]. Moreover, genome sequencing of some *Azospirillum* species revealed that significant part of the genome has been horizontally acquired [6]. Up until now, 16 *Azospirillum* species have been characterized; however complete genomic sequences of only *Azospirillum brasilense*, *Azospirillum lipoferum*, *Azospirillum* sp. B510, and a draft of *Azospirillum amazonense* genome have been published [7].

Azospirillum amazonense was found to be associated with the roots and rhizosphere of several grasses including sugarcane, maize, sorghum, and rice revealing a broad ecological distribution in Brazil. Studies revealed that *A. amazonense* is phylogenetically closer to *Rhodospirillum centenum* and *Azospirillum irakense* than to *A. brasilense*. Unlike other *Azospirillum* strains, *A. amazonense* can grow in the presence

of sucrose as sole carbon source and is also better adapted to soil acidity, which offers the bacterium additional advantages for colonization of plant root tissue in acid environments [8, 9]. Moreover, *A. amazonense* genomic analyses revealed the presence of genes not commonly distributed in other *Azospirillum* species such as those responsible for the utilization of salicin as carbon source (similar to *A. irakense*) and a gene cluster (RubisCO) implicated in carbon fixation (*A. lipoferum* is able to grow autotrophically by means of RubisCO, but the presence of the genes has not yet been demonstrated) [7]. However, our understanding of phytohormone production in *A. amazonense* is still incomplete.

The genomic plasticity of *A. amazonense* is probably related to the versatile gene repertoire present in the genome of this bacterium suggesting that horizontal gene transfer may have an impact on the adaptation and evolution of this species. Gene organization and phylogenetic analysis demonstrated that genes coding for proteins responsible for the nitrogen fixation process, carbon fixation (RubisCOs), and molecular hydrogen oxidation (hydrogenases) is more closely related to Rhizobiales members than to related species [7].

To further examine the importance of *A. amazonense* genetic variability, an *in silico* comparative genomic analysis using subtractive hybridization was performed using total coding sequences (CDS) from *A. amazonense* to compare with genomes of closely related bacteria. The analysis of conserved and specific *A. amazonense* coding sequences indicated features that distinguished *A. amazonense* from other *Azospirillum* species. Furthermore, the specific interesting features related to phytohormone production may provide several cues to establish *A. amazonense* pathways for auxin biosynthesis.

2. Material and Methods

2.1. Bacteria Selection and Genome Access. We have previously generated a good quality draft genome sequence of the *A. amazonense* Y2 (ATCC 35120) strain [7]. In this paper, the draft genome sequences were annotated and analyzed for the presence of specific regions, and during the BLAST search best-hits were detected with different bacteria, such as *Rhodospirillum*, *Azospirillum*, *Bradyrhizobium*, and *Caulobacter*.

Therefore, the *A. amazonense* comparative genomic analyses were performed using bacterial genomes including six species for which publicly closed genomes were available (Table 1). All genomes were downloaded from NCBI on January 10, 2013. The accession numbers used in this study are *Azospirillum amazonense* Y2 PRJNA73583, PRJNA65263; *Azospirillum* sp. B510 projects PRJNA46085, PRJDA32551; *Azospirillum brasilense* Sp245 PRJEA162161, PRJEA70627; *Azospirillum lipoferum* 4B PRJNA82343, PRJEA50367; *Rhodospirillum centenum* SW project PRJNA58805; *Bradyrhizobium japonicum* USDA 110 projects PRJNA57599 and PRJNA17; and *Caulobacter segnis* ATCC 21756 project PRJNA41709.

Table 1: General features for the bacteria genomes used in the comparative analysis.

Bacteria	Genome size	Total number of CDS	Assembly reference number
Azospirillum amazonense	7,044,835	3,319*	ASM22599v1
Azospirillum brasilense	7,530,241	7,557	ASM23736v1
Azospirillum lipoferum	6,846,400	6,093	ASM28365v1
Azospirillum sp. B510	7,599,738	6,309	ASM1072v1
Rhodospirillum centenum	4,355,543	4,003	ASM1618v1
Caulobacter segnis	4,655,622	4,139	ASM9228v1
Bradyrhizobium japonicum	9,105,828	8,317	ASM1136v1

*The total number of *A. amazonense* CDS was published by Sant'Anna et al. 2011 [7].

2.2. Annotation and Subtractive Hybridization. Reannotation of *A. amazonense* protein-coding genes was performed with a following procedure, which consists of two phases: initially the *A. amazonense* contigs were compared with the *Azospirillum* sp. B510 genome followed by functional annotation of each coding sequence (CDS) based on comparison with known sequences of the other six selected genomes using the Xbase Annotation Service [10]. All coding sequences predictions were manually checked for conservation in case of multiple hits, and only the alignments with best-hit results were selected from each genome. Information related to Cluster of Orthologous Group and KEGG pathway was added to the annotation using the server for metagenomic analysis (WebMGA) [11]. Annotation was based on comparison to protein clusters and on the BLAST results.

The subtractive hybridization using the mGenomeSubtractor program [12] was applied to run BLAST searches of the *A. amazonense* genome against multiple bacterial genomes for *in silico* comparative genomic analyses in order to characterize the unique sequences of *A. amazonense*. Proteins possibly related to phytohormones were analyzed in the Arabidopsis Hormone Database (AHD) [13], and proteins with homology (H) values more than 0.1 were arbitrarily defined as conserved coding sequences.

3. Results and Discussion

3.1. Comparative Analyses and Specific Protein Coding Sequences. The draft genome sequence of *A. amazonense* consists of 7,044,835 bp with 3,319 predicted coding sequences (CDS) where 2,299 have similarity with genes with known functions and 1,020 codes for hypothetical proteins or proteins of unknown function [7]. Although the estimated coverage of the genome was 35x, the number of predicted coding sequences was lower when compared with the other

species of *Azospirillum* where the total number of coding sequences ranges from 6,093 to 7,557 (Table 1).

In order to clarify the genomic coding sequences content of *A. amazonense*, two alternative comparative approaches using the Xbase Annotation Service were performed. Initially, to assess the coverage of the predicted gene repertoires a BLAST search was performed with only the *Azospirillum* sp. B510 genome. The total number of predicted protein-coding genes was 5,496 of which 2,165 were annotated as proteins of unknown function or hypothetical proteins. These numbers are similar to what is found in *A. lipoferum* and *Azospirillum* sp. B510 (Table 1). These results including the 5,496 sequences can be accessed using the http://www.xbase.ac.uk/annotation/results/rWn50Rn6LVRucf55SWfsvHfoqdpmd655/.

The second approach used an *A. amazonense*-vs-all (six selected genomes) BLAST to examine the overall similarity of the *A. amazonense* genome with closely related bacteria, and the results are shown in Table 2. The whole-genome comparisons revealed that the number of coding sequences found to be conserved and characterized as best-hits varied in each bacterium, from 3,126 (present in *Azospirillum* sp. B510) and 1,508 (present in *Rhodospirillum centenum*) to 2,846 (present in *R. centenum*) and 440 proteins (present in *A. lipoferum*), respectively. It is important to point out that the majority of the orthologs showing best-hit results were found with the *R. centenum* genome supporting previous suggestions of a close evolutionary relationship between *A. amazonense* and *R. centenum* [7, 14]. Interestingly, the number of coding sequences with best-hits orthologs found in the other *Azospirillum* species is almost equivalent to those found in the genome of bacteria from other genera, such as *Bradyrhizobium japonicum* and *Caulobacter segnis* (Table 2).

These unexpected results may support previous reports related with the genome repertoire of *A. amazonense* where horizontal gene transfer may be one of several events that result in an intragenera genomic plasticity. Phylogenetic analysis indicated the close relationship of the *A. amazonense* enzymes encoded by the gene cluster related to carbon fixation (RubisCo) and by genes related to nitrogen fixation (*nif*) processes with those from some species of the order Rhizobiales. Moreover, some features, such as the genetic organization of the carbon-fixation cluster and of the *nif* cluster of *A. amazonense,* are similar to the homolog cluster of the *Bradyrhizobium* species [7].

In conclusion, from the total 5,496 CDS found in the *A. amazonense* genome approximately half of the coding sequences have an ortholog in other closely related bacteria. However, using this methodology's numbers varying from 2,370 to 2,650 CDS showed lower degrees of similarity (*E* value > 10^{-10}) with coding sequences present in the genome of the compared bacteria.

Comparative genomic analysis using *in silico* subtractive hybridization allowed searching for specific proteins of the *A. amazonense* genome against multiple closely related bacterial genomes. Therefore, to determine the possible differences between the *A. amazonense* genome and each of the selected six closely related genomes, an *in silico* subtractive hybridization technique was applied. The histogram of *H*-values (Figure 1) was used to set the cutoff to discriminate between *A. amazonense*-specific and conserved coding sequences. Proteins with homology (*H*) values of less than 0.42 and more than 0.64 were arbitrarily defined as specific and conserved coding sequences, respectively [12]. This cutoff value was proposed by Shao et al. [12] and has been used in comparative genomic analyses to differentiate strains of pseudomonads [12, 15] or to compare genomes of species from the genus *Erwinia* [16].

Table 2: Predicted distribution of coding sequences (CDS) in *A. amazonense* draft genome and in the complete genome of other bacteria.

Comparisons	Conserved CDS		Specific CDS
	Best-hits	Total number	
A. a. versus *Azospirillum brasilense*	583	3,031	2,465
A. a. versus *Azospirillum lipoferum*	440	3,084	2,412
A. a. versus *Azospirillum* sp. B510	533	3,126	2,370
A. a. versus *Rhodospirillum centenum*	1,508	2,846	2,650
A. a. versus *Caulobacter segnis*	711	2,852	2,644
A. a. versus *Bradyrhizobium japonicum*	632	2,970	2,526

A. a.: *Azospirillum amazonense*.
Protein coding sequences with *E* value >10-10 were considered specific CDS (using the Xbase Annotation Service).

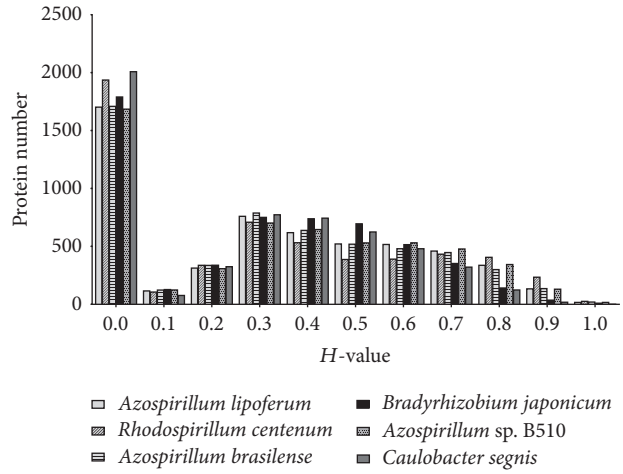


Figure 1: Histogram distribution of predicted proteins in *A. amazonense* compared with six closely related genomes using BLASTP-based homology value (*H* value). The *H*-value reflects the degree of similarity in terms of length of match and the degree of identity at amino acid level between the matching CDS in the subject genome and the query CDS examined with *E* value > 10^{-8}.

Table 3: Numbers of specific proteins for *A. amazonense* genome against six closely related genomes.

Comparisons	*A. amazonense*		
	Specific CDS	Conserved CDS	Other CDS
Azospirillum brasilense	3,689	948	859
Azospirillum lipoferum	3,606	746	1,144
Azospirillum sp. B510	3,571	793	1,132
Rhodospirillum centenum	3,697	770	1,029
Caulobacter segnis	4,043	382	1,071
Bradyrhizobium japonicum	3,880	317	1,299

The *in silico* subtractive hybridization analysis was performed with *A. amazonense* total coding sequences (CDS) against the proteins from the six genomes.
Proteins with homology (*H*) value less than 0.42 and more than 0.64 were arbitrarily defined as specific and conserved CDS, respectively, and other CDS were defined with *H* values between 0.42 and 0.64.

The subtractive hybridization approach revealed different profiles in gene number of specific and conserved proteins for the *A. amazonense* genome against the six others (Table 3). The number of proteins found to be conserved varied in each bacterium, from 948 CDS in *A. lipoferum* to 317 CDS in *B. japonicum*. Interestingly, the number of *A. amazonense*-specific proteins was similar when compared with the six bacteria varying from 4,043 (*C. segnis*) to 3,571 (*Azospirillum* sp. B510). Moreover, the specific proteins vary among the *Azospirillum* genomes analyzed from 3,571 to 3,689 only, indicating that these coding sequences are unique to the *A. amazonense* genome. Analyses of Figure 1 show that the majority of specific proteins in *A. amazonense* have *H*-values less than 0.1, suggesting that *Azospirillum* species are evolutionally diverse. This is consistent with previous studies that had proposed that some regions of the genome of *Azospirillum* species were acquired from distantly related bacteria from horizontal transfer [6].

To further characterize the global profile of *A. amazonense*-specific coding sequences, an *in silico* subtractive hybridization comparative analysis was performed with total *A. amazonense* putative coding sequences versus all six genomes, simultaneously. A total of 142 conserved CDS and 2,483 specific CDS were identified (see Supplementary Table 1 in Supplementary Material available online at http://dx.doi.org/10.1155/2014/898592). For function classification and pathway assignment, all specific and conserved *A. amazonense* coding sequences were classified to 20 different functional classes based on Clusters of Orthologous Groups (COG) (Table 4). The comparison of the *A. amazonense* genome and the other six available closely related genomes with regard to the functional category revealed that 1,196 specific CDS from *A. amazonense* were distributed among the different classes of orthologous clusters and that 1,287 specific CDS were unclassified being considered hypothetical products.

Table 4: Protein categories encoded by *A. amazonense* specific and conserved genes identified by *in silico* subtractive hybridization.

CDS assigned function*	Specific CDS	Conserved CDS
Transcription	143	11
Signal transduction mechanisms	137	5
Inorganic ion transport and metabolism	111	1
Carbohydrate transport and metabolism	101	4
Cell wall/membrane/envelope biogenesis	94	1
Amino acid transport and metabolism	86	19
Energy production and conversion	49	27
Secondary metabolites biosynthesis, transport, and catabolism	47	3
Cell motility	44	0
Coenzyme transport and metabolism	41	4
Lipid transport and metabolism	41	10
Intracellular trafficking, secretion, and vesicular transport	37	1
Defense mechanisms	34	0
Replication, recombination, and repair	33	3
Posttranslational modification, protein turnover, and chaperones	27	16
Translation, ribosomal structure, and biogenesis	14	27
Nucleotide transport and metabolism	12	12
Cell cycle control, cell division, and chromosome partitioning	4	1
RNA processing and modification	2	0
General function prediction only or function unknown	352	11

Proteins with homology (*H*) value less than 0.42 and more than 0.64 were arbitrarily defined as specific and conserved CDS, respectively.
*CDS assigned function was based on the COGs according to BLAST search.

Detailed analysis of the *A. amazonense*-specific CDS from Table 4 (and Supplementary Table 1) indicates special attention to the coding sequences classified in the Signal Transduction Mechanisms and Secondary Metabolites Biosynthesis, Transport and Catabolism functional class. Among the 137 CDS classified in the Signal Transduction Mechanisms functional category, there were protein coding sequences similar to cytokinins (ZP_08868173.1, ZP_08868457.1) and ethylene response (ZP_08867667.1) that could be related to phytohormone production (Supplementary Table 1). In particular, we have paid attention to a coding sequence related to lysine/ornithine N-monooxygenase (ZP_08869952.1) similar to flavin-containing monooxygenase from *Arabidopsis thaliana* (YUCCA9) involved in auxin synthesis in plants [17, 18] found in the Secondary Metabolites Biosynthesis, Transport and Catabolism functional class (Supplementary Table 1). Therefore, to better understand the auxin biosynthesis pathways in *A. amazonense*, studies attempting to define coding sequence related to phytohormone production were performed.

3.2. Phytohormone Production Related Sequences. The improvement of plant growth upon *Azospirillum* inoculation is attributed, as one of many factors, to the production of auxin by these bacteria [19]. Indole-3-acetic acid (IAA) is considered the most important auxin implicated in different aspects of plant growth. In bacteria, the two most common routes for indole-3-acetic acid biosynthesis are the IAM (indole-3-acetamide) and the IPyA (indole-3-pyruvate) pathways [20]. However, in *A. brasilense*, besides these two tryptophan-dependent pathways, an additional tryptophan-independent pathway was identified [21].

Similar to other *Azospirillum* species, *A. amazonense*, as typical plant-growth promoting rhizobacteria, stimulate root proliferation [22, 23]. However, genes responsible for biosynthesis and secretion of phytohormones are poorly described in this species. Previous works on *A. amazonense* genome sequence and annotation were unable to localize genes related to the IAM or IPyA pathways (*iaaM*, *iaaH*, and *ipdC*) and were able to identify only the presence of a coding sequence similar to nitrilases responsible for the conversion of indole 3-acetonitrile (IAN) to IAA in plants [7]. Therefore, the presence of coding sequences homologous to nitrilase and to flavin-containing monooxygenase (this paper) suggests that *A. amazonense* could use alternative pathways closely related to those found in plants.

Biosynthetic pathways for IAA have been fully investigated and tryptophan-dependent and Trp-independent routes have been studied [20, 24, 25]. Although genes coding for proteins related to the bacterial common routes IAM (indole-3-acetamide route) and IPyA (indole-3-pyruvic route) was not found in the *A. amazonense* genome, the identification of flavin-monooxygenase and nitrilase enzymes suggests the presence of the TAM (tryptamine route) and IAN (indole-3-acetamide route) pathways for IAA synthesis in this bacterium (Figure 2). It is well known that nitrilases in plants (maize and *Arabidopsis thaliana*) and also in *Bacillus amyloliquefaciens* were shown to hydrolyze indole-3-acetonitrile (IAN) to IAA [25, 26]. Moreover, evidence for the IAN and TAM pathways has been reported in *A. brasilense* [21].

Aiming to unveil the IAA pathways in *A. amazonense,* a search for other enzymes involving the TAM pathway was performed. The genome of *A. amazonense* contains an aldehyde oxidase-coding sequence (WP_004273557) homolog (query cover 91%; 33% identity; *E* value $e - 99$) to the *A. thaliana* AAO1 gene (AED92912) that is capable of oxidizing indole-3-acetaldehyde to indole-3-acetic acid with high efficiency [27]. Therefore, oxidation of indole-3-acetaldehyde by *A. amazonense* aldehyde oxidase is likely to represent the TAM route transforming indole-3-acetaldehyde (IAAld) to produce IAA phytohormone in this bacterium (Figure 2). To conclude, *A. amazonense* appears to possess only one regulated Trp-dependent route for IAA synthesis, the TAM pathway, while *A. brasilense* possesses two differently regulated routes, namely, the IPyA and the TAM pathways [21]. Furthermore, the alternative route that uses IAN as an intermediate, the IAN pathway, appears to be present in both species, but it remains to be established whether the IAN pathway is the Trp-independent route.

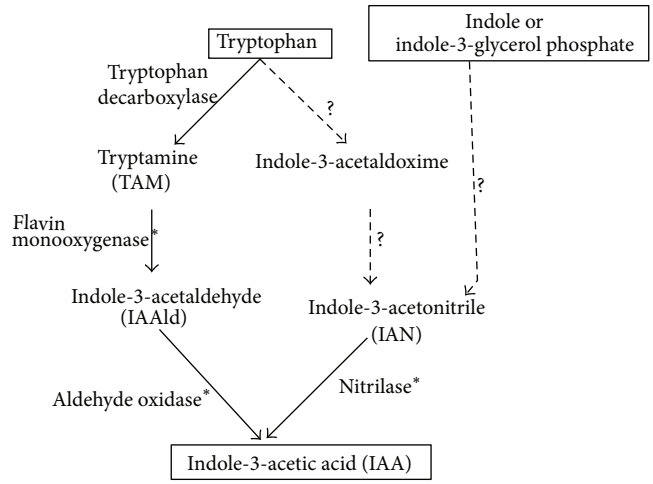


Figure 2: *A. amazonense* pathways of IAA biosynthesis. Tryptophan-dependent pathways or tryptophan-independent pathways (starting from indole or indole-3-glycerol phosphate) are indicated based on routes found in plants and bacteria. Enzymes indicated with an asterisk have been identified in *A. amazonense*, and routes indicated as dotted lines indicated that the precursor of IAN may or may not be tryptophan.

To further understand the phytohormone biosynthesis pathway in *A. amazonense,* an *in silico* comparative analysis was performed with total coding sequences from *A. amazonense* versus all proteins deposited in the Arabidopsis Hormone Database. Furthermore, the comparative analysis was also performed with coding sequences in the auxin response transcriptome data of *A. brasilense* [28]. A total of 54 *A. amazonense* CDS revealed similarity with proteins related to hormone production in plants and bacterial auxin signal transduction pathways (Supplementary Table 1). The presence of these coding sequences suggests that IAA could be a signal that alters gene expression in *A. amazonense* similar to that found in *A. brasilense* [28]. Moreover, the genome of *A. amazonense* has coding sequences that could be related to other hormone pathways similar to those described for other *Azospirillum* species [19, 21].

Another beneficial effect provided by the association of soil bacteria with plants could be due to the plant hormone ethylene, which can inhibit plant growth by regulating several developmental aspects [29]. Similar to other plant growth-promoting rhizobacteria, a coding sequence homologous to 1-aminocyclopropane-1-carboxylate (ACC) deaminase (WP_004272971.1) has been identified in the *A. amazonense* genome. Previous reports have suggested that *A. amazonense* can stimulate plant growth by producing or metabolizing plant hormones and the presence of ACC deaminase which can hydrolyze ACC, the immediate precursor of the plant hormone ethylene, could be involved by lowering the plant ethylene levels and increasing plant growth. Moreover, the identification of the octaprenyl diphosphate synthase enzyme (ZP_08868744) could be related to cytokinin biosynthesis by a known hormone that affects plant growth and yield. Although the auxin hormone is considered a major

class of hormones regulating plant growth, cytokinins or ethylene-related phytohormones could interact with auxin leading to root system development.

In conclusion, it appears that the rhizosphere bacterium *A. amazonense* is able to produce IAA through the tryptamine and indole-3-acetonitrile pathways and similar to *A. brasilense* could alter gene expression in response to the presence of auxin. Moreover, the role as plant-growth-promoting bacteria could be related to IAA production or to its ability to metabolize the ethylene precursor (ACC) and thereby increases the growth of the root system. Furthermore, the multiple genome comparison performed with *A. amazonense* and closely related bacteria supports previous evidence concerning *A. amazonense* genomic versatility and that several genes could have been acquired from distantly related bacteria [6]. Future investigation of *A. amazonense* is necessary to support the hypothesis that its genomic structures have evolved to meet the requirements for adaptation to the rhizosphere and interaction with host plants.

Conflict of Interests

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

Acknowledgments

This work was supported by grants from the Brazilian National Research Council (CNPq) and the Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS). R. Cecagno and T. E. Fritsch received scholarships from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

References

[1] Y. Bashan, G. Holguin, and L. E. de-Bashan, "*Azospirillum*-plant relationships: physiological, molecular, agricultural, and environmental advances (1997–2003)," *Canadian Journal of Microbiology*, vol. 50, no. 8, pp. 521–577, 2004.

[2] S. Fibach-Paldi, S. Burdman, and Y. Okon, "Key physiological properties contributing to rhizosphere adaptation and plant growth promotion abilities of *Azospirillum* brasilense," *FEMS Microbiology Letters*, vol. 326, no. 2, pp. 99–108, 2012.

[3] M. Kochar and S. Srivastava, "Surface colonization by Azospirillum brasilense SM in the indole-3-acetic acid dependent growth improvement of sorghum," *Journal of Basic Microbiology*, vol. 52, no. 2, pp. 123–131, 2012.

[4] C. C. G. Martin-Didonet, L. S. Chubatsu, E. M. Souza et al., "Genome structure of the genus *Azospirillum*," *Journal of Bacteriology*, vol. 182, no. 14, pp. 4113–4116, 2000.

[5] A. V. Shelud'ko, O. E. Varshalomidze, L. P. Petrova, and E. I. Katsy, "Effect of genomic rearrangement on heavy metal tolerance in the plant-growth-promoting rhizobacterium Azospirillum brasilense Sp245," *Folia Microbiologica*, vol. 57, no. 1, pp. 5–10, 2012.

[6] F. Wisniewski-Dyé, K. Borziak, G. Khalsa-Moyers et al., "*Azospirillum* genomes reveal transition of bacteria from aquatic to terrestrial environments," *PLoS Genetics*, vol. 7, no. 12, Article ID e1002430, 2011.

[7] F. H. Sant'Anna, L. G. P. Almeida, R. Cecagno et al., "Genomic insights into the versatility of the plant growth-promoting bacterium *Azospirillum amazonense*," *BMC Genomics*, vol. 12, article 409, 2011.

[8] F. M. Magalhães, J. I. Baldani, S. M. Souto, J. R. Kuykendall, and J. Dobereiner, "A new acid-tolerant *Azospirillum* species," *Anais da Academia Brasileira de Ciências*, vol. 55, pp. 417–430, 1983.

[9] J. I. Baldani and V. L. D. Baldani, "History on the biological nitrogen fixation research in graminaceous plants: special emphasis on the Brazilian experience," *Anais da Academia Brasileira de Ciencias*, vol. 77, no. 3, pp. 549–579, 2005.

[10] R. R. Chaudhuri, N. J. Loman, L. A. S. Snyder, C. M. Bailey, D. J. Stekel, and M. J. Pallen, "xBASE2: a comprehensive resource for comparative bacterial genomics," *Nucleic Acids Research*, vol. 36, no. 1, pp. D543–D546, 2008.

[11] S. Wu, Z. Zhu, L. Fu, B. Niu, and W. Li, "WebMGA: a customizable web server for fast metagenomic sequence analysis," *BMC Genomics*, vol. 12, article 444, 2011.

[12] Y. Shao, X. He, E. M. Harrison et al., "mGenomeSubtractor: a web-based tool for parallel in silico subtractive hybridization analysis of multiple bacterial genomes," *Nucleic Acids Research*, vol. 38, no. 2, Article ID gkq326, pp. W194–W200, 2010.

[13] Z.-Y. Peng, X. Zhou, L. Li et al., "*Arabidopsis* hormone database: a comprehensive genetic and phenotypic information database for plant hormone research in *Arabidopsis*," *Nucleic Acids Research*, vol. 37, no. 1, pp. D975–D982, 2009.

[14] M. Stoffels, T. Castellanos, and A. Hartmann, "Design and application of new 16S rRNA-targeted oligonucleotide probes for the Azospirillum-Skermanella-Rhodocista-cluster," *Systematic and Applied Microbiology*, vol. 24, no. 1, pp. 83–97, 2001.

[15] M. Qi, D. Wang, C. A. Bradley, and Y. Zhao, "Genome sequence analyses of *Pseudomonas savastanoi* pv. glycinea and subtractive hybridization-based comparative genomics with nine pseudomonads," *PLoS ONE*, vol. 6, no. 1, Article ID e16451, 2011.

[16] Y. Zhao and M. Qi, "Comparative genomics of Erwinia amylovora and related Erwinia species—what do we learn?" *Genes*, vol. 2, no. 3, pp. 627–639, 2011.

[17] Y. Zhao, S. K. Christensen, C. Fankhauser et al., "A role for flavin monooxygenase-like enzymes in auxin biosynthesis," *Science*, vol. 291, no. 5502, pp. 306–309, 2001.

[18] Y. Zhao, "Auxin biosynthesis: a simple two-step pathway converts tryptophan to indole-3-Acetic acid in plants," *Molecular Plant*, vol. 5, no. 2, pp. 334–338, 2012.

[19] O. Ona, J. van Impe, E. Prinsen, and J. Vanderleyden, "Growth and indole-3-acetic acid biosynthesis of *Azospirillum* brasilense Sp245 is environmentally controlled," *FEMS Microbiology Letters*, vol. 246, no. 1, pp. 125–132, 2005.

[20] M. Lambrecht, Y. Okon, A. V. Broek, and J. Vanderleyden, "Indole-3-acetic acid: a reciprocal signalling molecule in bacteria-plant interactions," *Trends in Microbiology*, vol. 8, no. 7, pp. 298–300, 2000.

[21] R. Carreño-Lopez, N. Campos-Reales, C. Elmerich, and B. E. Baca, "Physiological evidence for differently regulated tryptophan-dependent pathways for indole-3-acetic acid synthesis in *Azospirillum brasilense*," *Molecular and General Genetics*, vol. 264, no. 4, pp. 521–530, 2000.

[22] O. Steenhoudt and J. Vanderleyden, "Azospirillum, a free-living nitrogen-fixing bacterium closely associated with grasses: genetic, biochemical and ecological aspects," *FEMS Microbiology Reviews*, vol. 24, no. 4, pp. 487–506, 2000.

[23] E. Rodrigues, L. Rodrigues, A. de Oliveira et al., "*Azospirillum amazonense* inoculation: effects on growth, yield and N_2 fixation of rice (*Oryza sativa* L.)," *Plant and Soil*, vol. 316, no. 1-2, p. 323, 2009.

[24] B. Bartel, "Auxin biosynthesis," *Annual Review of Plant Biology*, vol. 48, no. 1, pp. 51–66, 1997.

[25] E. E. Idris, D. J. Iglesias, M. Talon, and R. Borriss, "Tryptophan-dependent production of Indole-3-Acetic Acid (IAA) affects level of plant growth promotion by *Bacillus amyloliquefaciens* FZB42," *Molecular Plant-Microbe Interactions*, vol. 20, no. 6, pp. 619–626, 2007.

[26] V. Kriechbaumer, W. J. Park, M. Piotrowski, R. B. Meeley, A. Gierl, and E. Glawischnig, "Maize nitrilases have a dual role in auxin homeostasis and β-cyanoalanine hydrolysis," *Journal of Experimental Botany*, vol. 58, no. 15-16, pp. 4225–4233, 2007.

[27] M. Seo, S. Akaba, T. Oritani et al., "Higher activity of an aldehyde oxidase in the auxin-overproducing superroot1 mutant of arabidopsis thaliana," *Plant Physiology*, vol. 116, no. 2, pp. 687–693, 1998.

[28] S. van Puyvelde, L. Cloots, K. Engelen et al., "Transcriptome analysis of the rhizosphere bacterium *Azospirillum brasilense* reveals an extensive auxin response," *Microbial Ecology*, vol. 61, no. 4, pp. 723–728, 2011.

[29] L. Chen, I. C. Dodd, J. C. Theobald, A. A. Belimov, and W. J. Davies, "The rhizobacterium Variovorax paradoxus 5C-2, containing ACC deaminase, promotes growth and development of Arabidopsis thaliana via an ethylene-dependent pathway," *Journal of Experimental Botany*, vol. 64, no. 6, pp. 1565–1573, 2013.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 689745, 14 pages
http://dx.doi.org/10.1155/2015/689745

Research Article

Shaped Singular Spectrum Analysis for Quantifying Gene Expression, with Application to the Early *Drosophila* Embryo

Alex Shlemov,[1] Nina Golyandina,[1] David Holloway,[2] and Alexander Spirov[3,4]

[1] *Faculty of Mathematics and Mechanics, St. Petersburg State University, Universitetsky Pr. 28, Peterhof, St. Petersburg 198504, Russia*
[2] *Mathematics Department, British Columbia Institute of Technology, 3700 Willingdon Avenue, Burnaby, BC, Canada V5G 3H2*
[3] *Computer Science and CEWIT, SUNY Stony Brook, 1500 Stony Brook Road, Stony Brook, NY 11794, USA*
[4] *The Sechenov Institute of Evolutionary Physiology & Biochemistry, Torez Pr. 44, St. Petersburg 194223, Russia*

Correspondence should be addressed to Alexander Spirov; alexander.spirov@gmail.com

Received 4 July 2014; Revised 10 September 2014; Accepted 10 September 2014

Academic Editor: Hongwei Wang

In recent years, with the development of automated microscopy technologies, the volume and complexity of image data on gene expression have increased tremendously. The only way to analyze quantitatively and comprehensively such biological data is by developing and applying new sophisticated mathematical approaches. Here, we present extensions of 2D singular spectrum analysis (2D-SSA) for application to 2D and 3D datasets of embryo images. These extensions, circular and shaped 2D-SSA, are applied to gene expression in the nuclear layer just under the surface of the *Drosophila* (fruit fly) embryo. We consider the commonly used cylindrical projection of the ellipsoidal *Drosophila* embryo. We demonstrate how circular and shaped versions of 2D-SSA help to decompose expression data into identifiable components (such as trend and noise), as well as separating signals from different genes. Detection and improvement of under- and overcorrection in multichannel imaging is addressed, as well as the extraction and analysis of 3D features in 3D gene expression patterns.

1. Introduction

While the availability of genome sequences has drastically revolutionized biological and biomedical research, our understanding of how genes encode regulatory mechanisms is still limited. Embryonic development depends critically on such regulatory mechanisms in order for cells to differentiate in the correct positions and at the correct times. Global understanding of gene regulation in development requires determining at cellular resolution in vivo when and where each gene is expressed. New dynamic, cellular resolution atlases will address the question of how gene transcription factors influence expression patterning [1].

With the development of automated microscopy technologies in recent years the volume and complexity of image data have increased to the level that it is no longer feasible to extract information without using computational tools. Biologists increasingly rely on computer scientists to come up with new solutions and software [2]. Such computational tools have been essential for processing the images generated by high-throughput microscopy of large numbers and varieties of biological samples under a variety of conditions. Recent advances in labeling, imaging, and computational image analysis are allowing quantitative measurements to be made more readily and in much greater detail in a range of organisms (e.g., *Arabidopsis*, *Ciona*, *Drosophila*, *C. elegans*, mice, *Platynereis*, and zebrafish) [1, 3–6]. In particular, imaging of single intact small organisms, like *Drosophila* and *C. elegans*, is now feasible with high resolution in two dimensions, three dimensions, and across time, resulting in massive image data sets available for comprehensive computational analysis.

These large-scale quantitative data sets provide new insights to address many fundamental questions in developmental biology. The initial inputs for deriving quantitative information of gene expression and embryonic morphology are usually raw image data of stained fluorescent markers in fixed material. These raw image sets are then analyzed by computational algorithms that extract features such as cell location, cell shape, and gene product concentration.

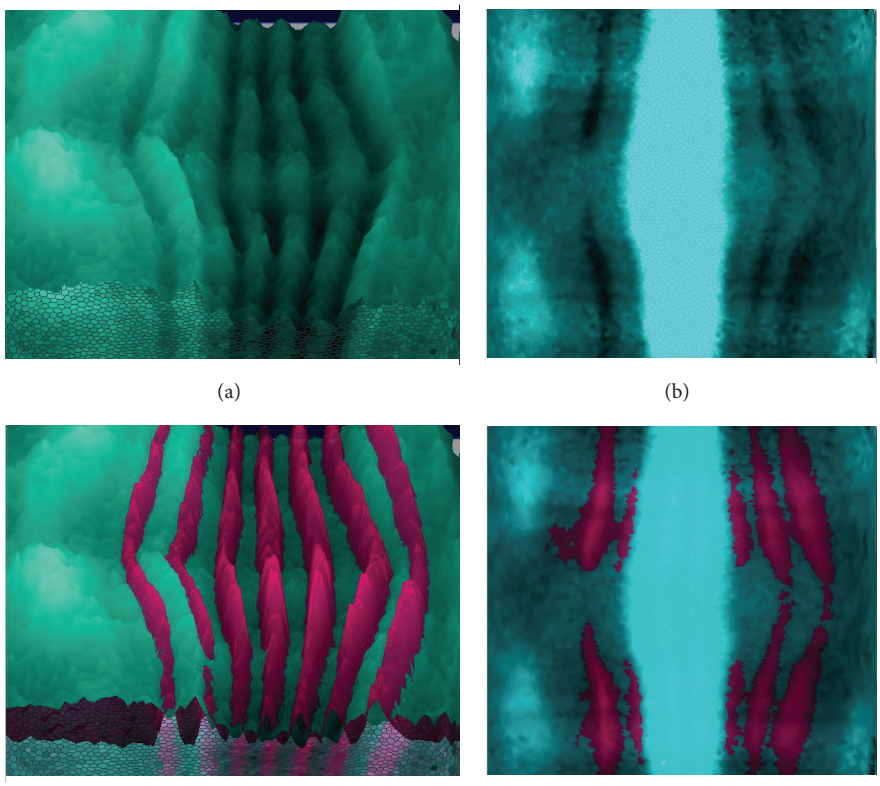


Figure 1: An example of overcorrection in gene expression data causing the subtraction of the reference gene pattern (the seven-striped *ftz* and *eve* patterns; dark magenta) from the pattern under study (*hb* and *Kr* gene products (transcription factors); light blue). Visualization by PointCloudXplore tools [7], BDTNP embryos *hb* "v5-s11512-2oc06-25" ((a) and (c)), *Kr* "v5-s12169-24oc07-22" ((b) and (d)); (c) is the same as (a) with added *ftz*; (d) is the same as (b) with added *eve*.

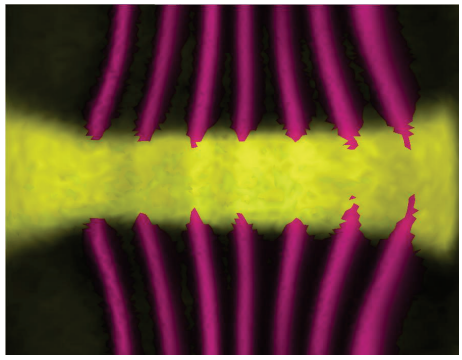


Figure 2: An example of undercorrection, in which the periodic reference gene pattern (*eve*; dark magenta) adds periodicity to the nonperiodic pattern under study (*sna* gene product; yellow). Visualization by PointCloudXplore. Embryo "v5-s10531-28fe05-07."

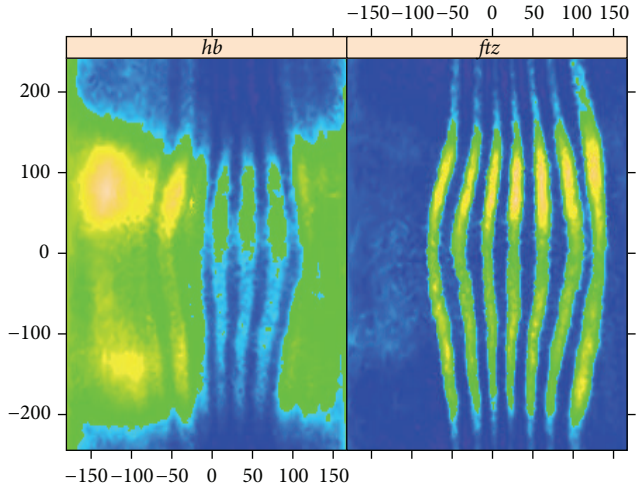


Figure 3: *hb* and *ftz*: original images of the "unrolled" cylindrical surface; the top values are a direct continuation of bottom values.

Ultimately, the most powerful way to analyze 3D spatial data in biology is by developing and applying new sophisticated mathematical approaches, allowing for the rigorous comparison of multiple quantitative features [8, 9].

In this publication, we introduce new computational tools to analyze gene patterning for three spatial dimension datasets, applied to early *Drosophila* embryos. These tools are an extension of two-dimensional singular spectrum analysis (2D-SSA).

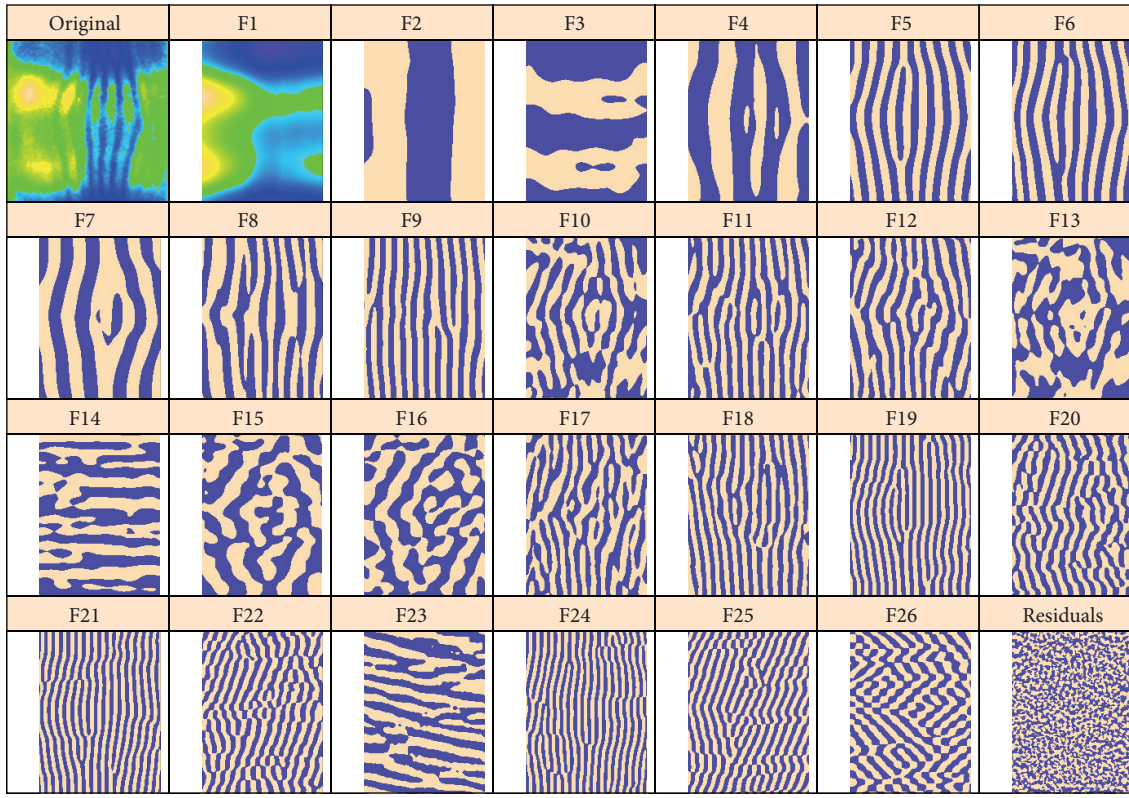


FIGURE 4: *hb*: the original image and the elementary components extracted by circular 2D-SSA. The original image and the leading component (F1) are colour-mapped according to the min and max expression levels. For more contrast, the remaining components are depicted in a binary format, with positive values in beige and negative values in purple.

Introduction to the Method. Singular spectrum analysis [10–15] was originally suggested as a method for decomposition of time series into a sum of identifiable components such as trend (or pattern), oscillations, and noise. One advantage of this method is that it does not need a noise model to be given a priori. We decompose the data series into a set of elementary series, analyze them, choose appropriate components, and finally sum the identifiable components together in classes. As an example, selection of smooth components can produce adaptive smoothing. SSA is very useful for exploratory analysis since the method can deal with modulated noise, that is, noise that can depend on trend values (e.g., has a multiplicative nature).

Recently SSA was extended for analysis of two-dimensional objects (2D-SSA), for example, digital images [16, 17]. Decomposition of images is more complicated compared to time series analysis due to variability of 2D patterns. But methods which are easily controlled and adaptive, such as 2D-SSA, can have broad applicability.

2D-SSA has much in common with the 2D-ESPRIT method (see [18]), which is based on the parametric form of images and has many applications. 2D-SSA and related subspace-based methods are applied in texture analysis [19], seismology [20], spatial gene expression data [21], and medical imaging [22].

The paper [23] applied 2D-SSA to the analysis of digital terrains in geology and demonstrated that 2D-SSA is a useful tool for analyzing different levels of details in surface data. Later, based on the theory given in [17], 2D-SSA was applied to gene expression data to separate nuclear noise from expression trend [21].

The papers [24, 25] present extensions of 2D-SSA which increase the range of SSA applications. In the present paper, we demonstrate how these extensions can be applied to analyzing gene expression data.

This paper is structured as follows. Section 2 describes the data sets which were analyzed. Section 3 describes the new methodology, and Sections 4 and 5 demonstrate the approach on several examples.

The new approaches described here, circular and shaped 2D-SSA, are particulary applicable to cylindrical surfaces (as used for *Drosophila* embryos), to avoid edge effects and patterns of irregular shape. For example, the area of good quality data in an image (e.g., without oversaturation) can be nonrectangular and even have gaps. Also, since the planar projection of a *Drosophila* embryo is nearly elliptical, the ability to analyze nonrectangular shapes can be useful.

Section 4 deals with the problem of detection and improvement of under- and overcorrection in multichannel

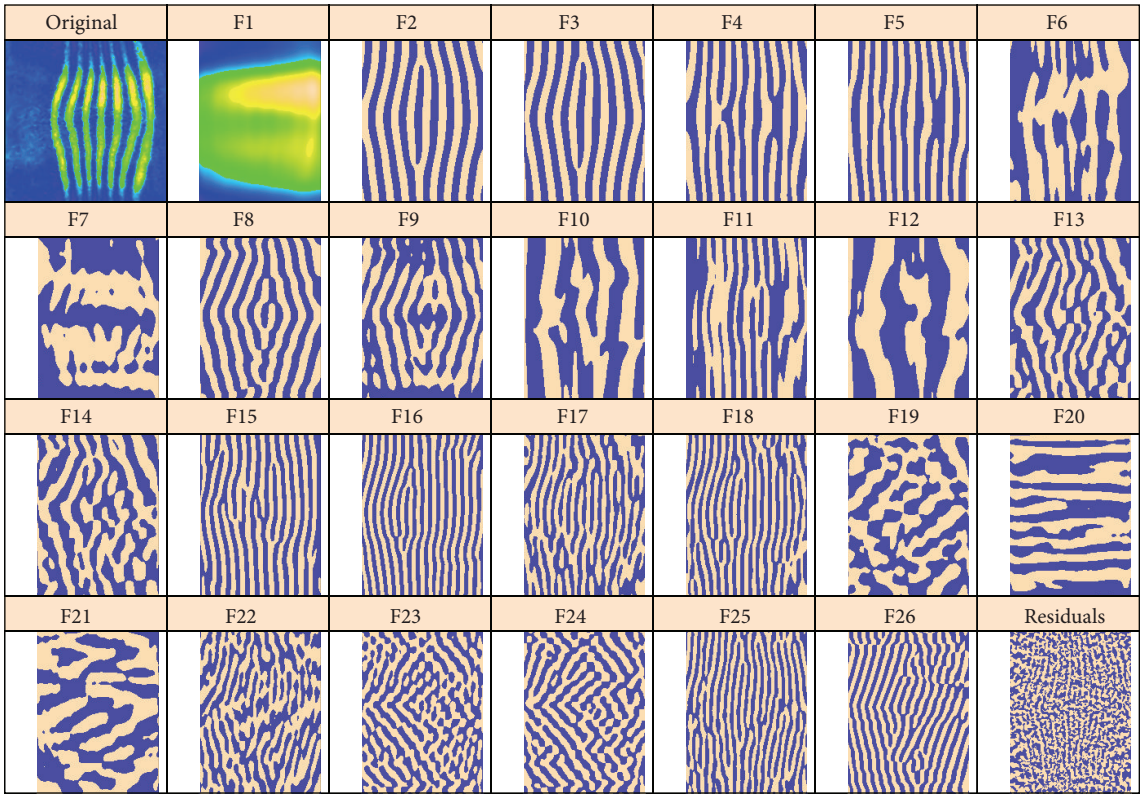


FIGURE 5: *ftz*: original image, F1 with the background; the remaining elementary components are depicted in a binary format.

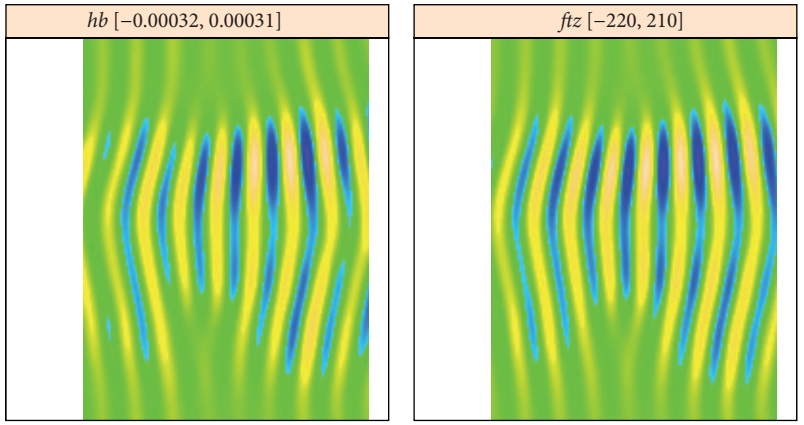


FIGURE 6: *hb* (a) and *ftz* (b): reconstruction from the main striped components 5 and 6 for the *hb* analysis, 2 and 3 for the *ftz* analysis. The stripes are out of phase for *hb* and *ftz*.

imaging, while Section 5 considers the problem of analysis of stripe shapes for the even skipped gene. Section 6 contains a short discussion and conclusions.

2. Materials

Data are taken from the Berkeley Drosophila Transcription Network Project (BDTNP) [4], which contains three-dimensional (3D) measurements of relative mRNA concentration for 95 genes in early development (including *snail* (*sna*)) and the protein expression patterns for four genes (bicoid, giant, hunchback (*hb*), and Krüppel (*Kr*)) during nuclear cleavage cycles 13 (C13) and 14 (C14A). BDTNP Release 2 contains individual datasets (PointCloud files) for 2830 embryos (http://bdtnp.lbl.gov/Fly-Net/bioimaging.jsp). These data were registered to the coordinates of 6078 nuclei on the embryo cortex and presented as an integrated dataset (VirtualEmbryo file, with tools for visualization and analysis).

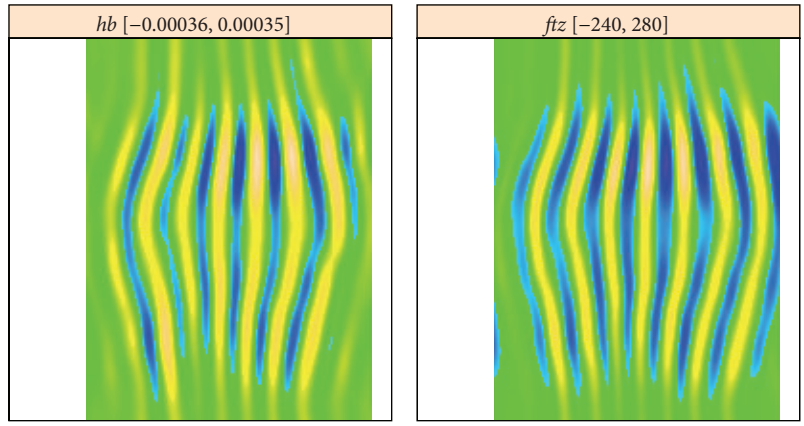


Figure 7: *hb* (a) and *ftz* (b): reconstruction from all striped components.

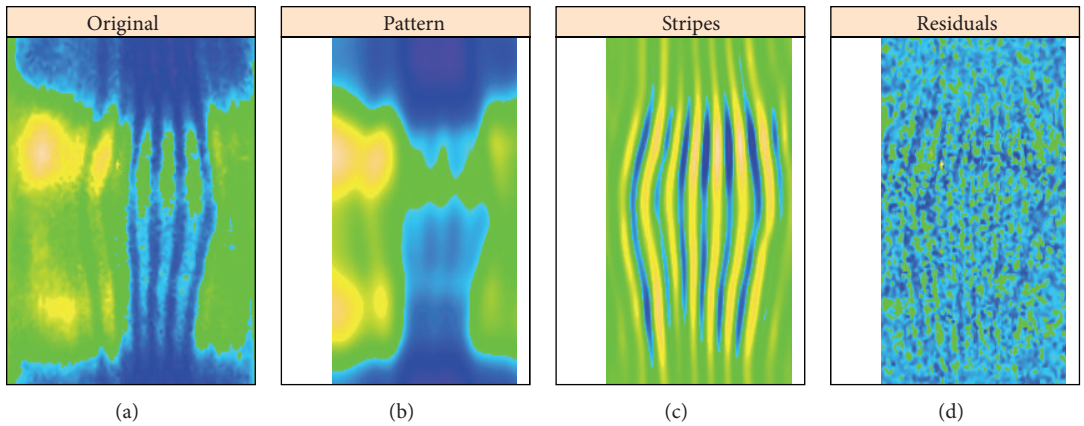


Figure 8: *hb* ((a) to (d)): original image, unstriped pattern, stripes, and residual noise.

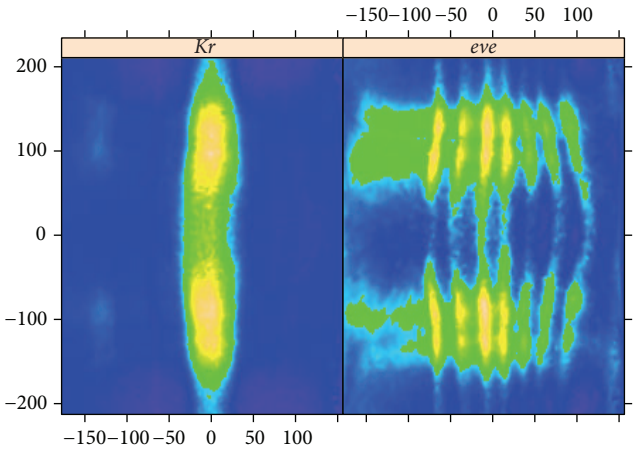


Figure 9: *Kr* and *eve*: original images.

Embryos were fixed and fluorescently stained to label the mRNA expression patterns of two genes plus nuclear DNA. One of the genes stained was either even skipped (*eve*) or *fushi tarazu* (*ftz*), which were used as fiduciary markers for subsequent spatial registration.

3. Methods

3.1. 2D Singular Spectrum Analysis. We will follow the common structure of 2D-SSA algorithms described in [24, 25]. This common structure consists of embedding, decomposition, grouping, and reconstruction steps. Input for a 2D-SSA algorithm consists of an image $\mathbb{X}$ and the shape of a moving window (which is the main algorithm parameter). The output of a 2D-SSA algorithm is the decomposition of $\mathbb{X}$ into identifiable components of the form $\mathbb{X} = \mathbb{X}_1 + \cdots + \mathbb{X}_s$.

Common Scheme of SSA-Like Algorithms

(1) Embedding Step. Construction of the trajectory matrix $\mathbf{X} = \mathcal{T}(\mathbb{X}) \in \mathsf{H}$, where H is a space of structured Hankel-like matrices. The structure of the matrix $\mathbf{X}$ (and the space H) depends on the algorithm modification and on the moving window. Generally speaking, the columns of the trajectory matrix consist of the windows moving along the image, transformed to vectors by a fixed order of window

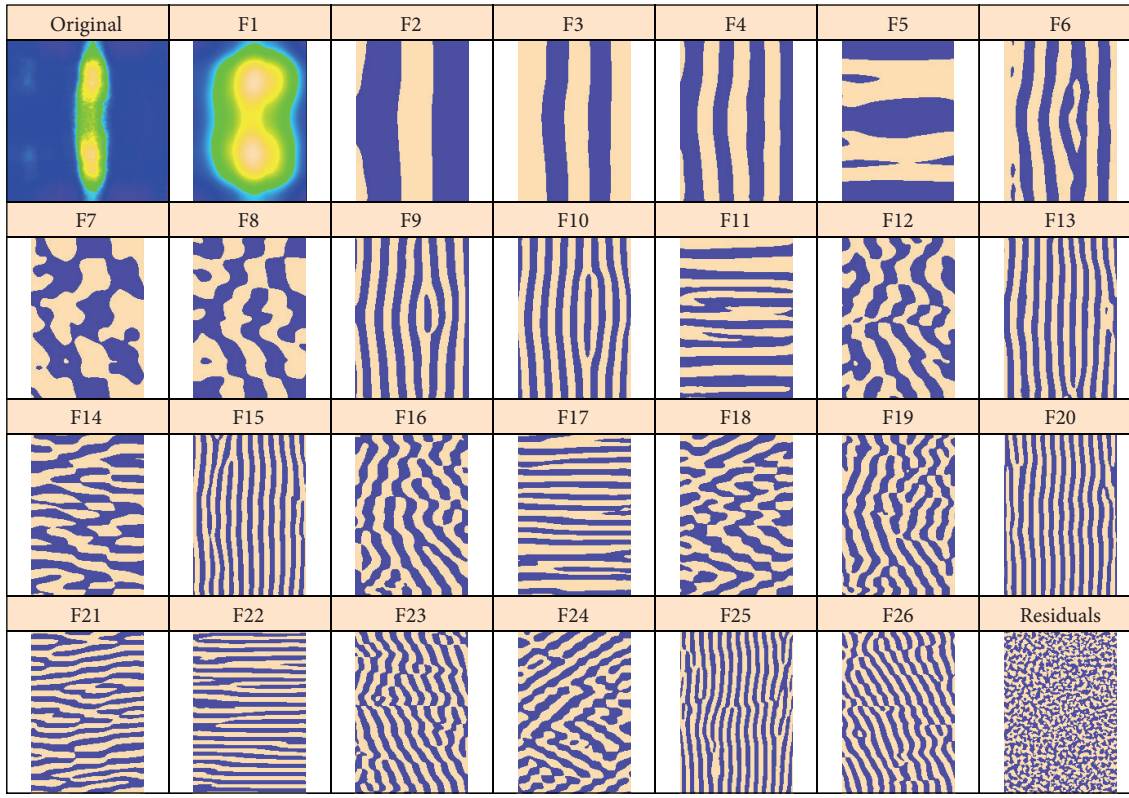


FIGURE 10: *Kr* gene expression, with circular 2D-SSA decomposition: original image and elementary components. As with Figures 11 and 10, the original image and leading component (F1) are colour-mapped according to min and max expression levels. For more contrast, the remaining components are depicted purple and beige.

elements. In a sense, the window size reflects the resolution of the method; that is, larger windows lead to more detailed decompositions.

(2) Decomposition Step. Singular value decomposition (SVD) of the trajectory matrix $\mathbf{X} = \sum_{i=1}^{d} \sqrt{\lambda_i} U_i V_i^{\mathrm{T}} = \sum_{i=1}^{d} \mathbf{X}_i$. Here $(\sqrt{\lambda_i}, U_i, V_i)$ are so-called eigentriples (abbreviated as ET) and consist of singular values, left and right singular vectors of $\mathbf{X}$. The eigenvectors can be transformed back to the window form. This means that we can consider eigenvectors as images and call them eigenimages.

(3) Grouping Step. Partition $\{1, \ldots, d\} = \coprod_{j=1}^{s} I_j$ and grouping of summands in the SVD decomposition to obtain a grouped matrix decomposition $\mathbf{X} = \sum_{j=1}^{s} \mathbf{X}_{I_j}$, where $\mathbf{X}_I = \sum_{k \in I} \mathbf{X}_k$. The grouping with $I_j = \{j\}$ is called elementary. The aim of this step is to group the SVD components to obtain an interpretable decomposition of the initial object. This can be performed by means of analysis of eigentriples.

(4) Reconstruction Step. Decomposition of the initial image $\mathbb{X} = \mathbb{X}_1 + \cdots + \mathbb{X}_s$, where $\mathbb{X}_j = \mathcal{T}^{-1} \mathcal{H}(\mathbf{X}_{I_j})$; $\mathcal{H}$ is the operator of projection on the space H (e.g., hankelization in the 1D case); $\mathcal{H}(\mathbf{X}_I) = \sum_{i \in I} \mathcal{H}(\mathbf{X}_i)$ holds.

Let us explain the sense of the embedding operator $\mathcal{T}$ for the 1D case, since it is simpler and demonstrates the general methodology. For a one-dimensional series $\mathbb{X} = (x_1, \ldots, x_N)$, we take moving 1D windows of length L and construct the columns of the trajectory matrix in the forms $X_1 = (x_1, \ldots, x_L)^{\mathrm{T}}$, $X_2 = (x_2, \ldots, x_{L+1})^{\mathrm{T}}$, and so on. From these $K = N - L + 1$ lagged vectors we gather a Hankel matrix with equal numbers on antidiagonals called the trajectory matrix

$$\mathcal{T}_{\mathrm{SSA}}(\mathbb{X}) = \begin{pmatrix} x_1 & x_2 & x_3 & \cdots & x_K \\ x_2 & x_3 & x_4 & \cdots & x_{K+1} \\ x_3 & x_4 & x_5 & \cdots & x_{K+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_L & x_{L+1} & x_{L+2} & \cdots & x_N \end{pmatrix}. \quad (1)$$

It is well known that Hankel matrices are related to series which consist of sums of products of polynomials, exponentials, and sine waves and the problem is to separate this sum into addends. If we can separate exponential and polynomial approximations from the residual, then we can extract trends and patterns. If we are able to separate sine waves with different frequencies, then we can construct a decomposition on components with different frequency ranges.

The singular value decomposition (SVD) of the trajectory matrix constructs a sequence of elementary matrices, which provides the best approximations of the initial matrix and,

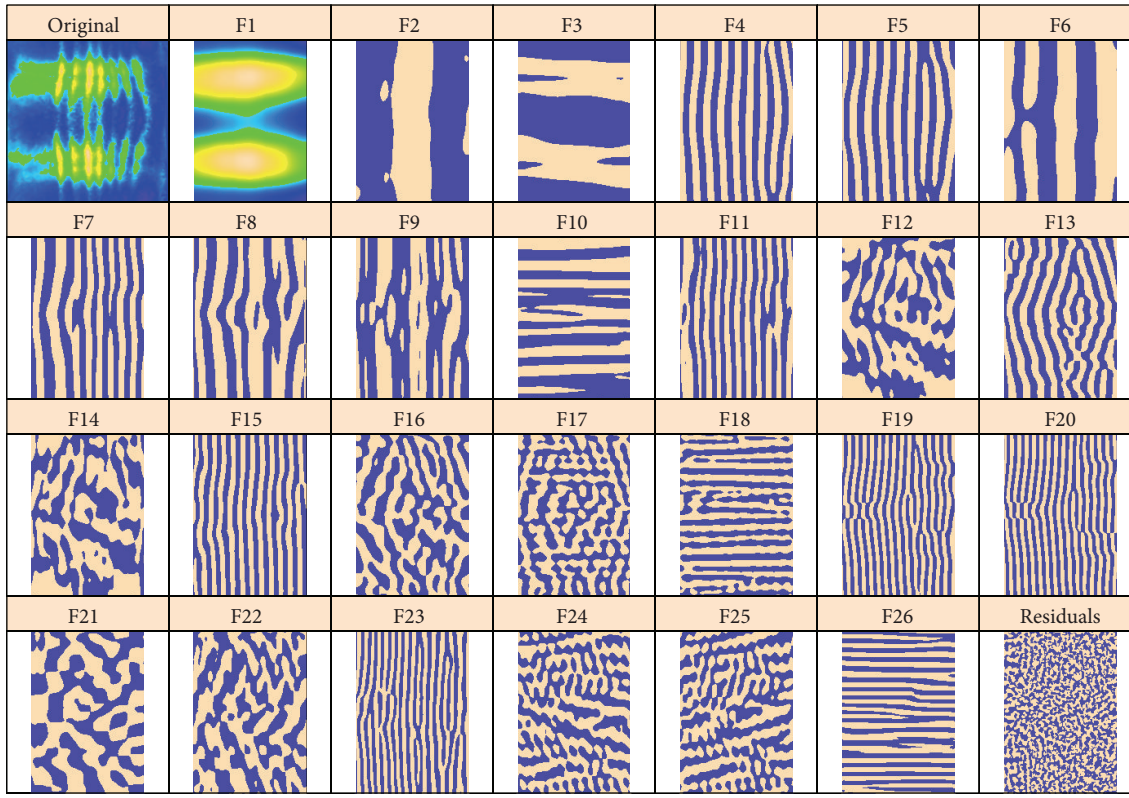


FIGURE 11: *eve* gene expression, with circular 2D-SSA decomposition: original image and elementary components.

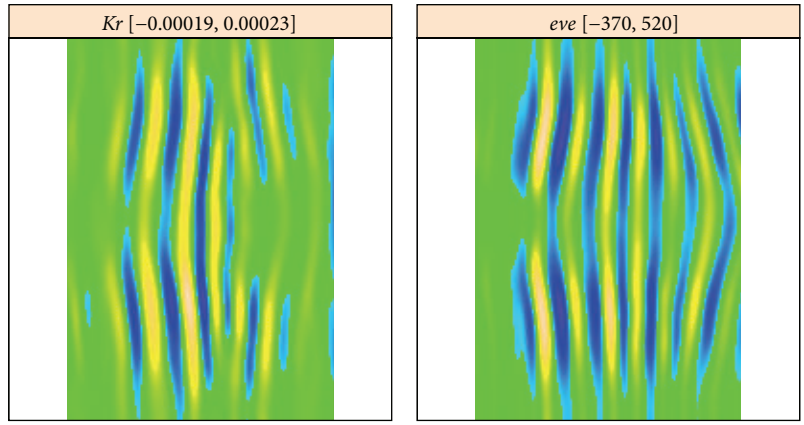


FIGURE 12: *Kr* and *eve*: reconstruction with stripe components, from the *Kr* image (a) and from the *eve* image (b). The frequencies correspond, but are out-of-phase, indicating overcorrection in the unmixing algorithm.

in a sense, of the initial series: $\mathbf{X}_1$, $\mathbf{X}_1 + \mathbf{X}_2$, and so on. Thus, we obtain the optimal decomposition, which is adaptive to the initial series. Note that the maximal number of the decomposition elements is equal to $\min(L, K)$. SSA theory explains why we can group the elementary components in the SVD expansion to solve such problems as, for example, smooth approximation and extraction of regular oscillations.

After a proper grouping, we obtain a matrix $\mathbf{X}_I$, which is close to a Hankel matrix, but not exactly Hankel. We can find the Hankel matrix closest to $\mathbf{X}_I = \{y_{ij}\}$ by hankelization, that is, by averaging values by antidiagonals. Thus, we obtain the series consisting of y_{11}, $(y_{12} + y_{21})/2$, $(y_{13} + y_{22} + y_{31})/3$, and so on. The mth term is determined as $\sum_{i,j \in \mathscr{A}_m} y_{ij}/|\mathscr{A}_m|$, where $\mathscr{A}_m = \{i, j : 1 \le i \le L, 1 \le j \le K, i + j = m + 1\}$.

The role of L is as follows. Small L provides a decomposition to a small number of components, which mostly differ by frequency, and where the leading components present slowly varying series like the trend. Larger L leads to more

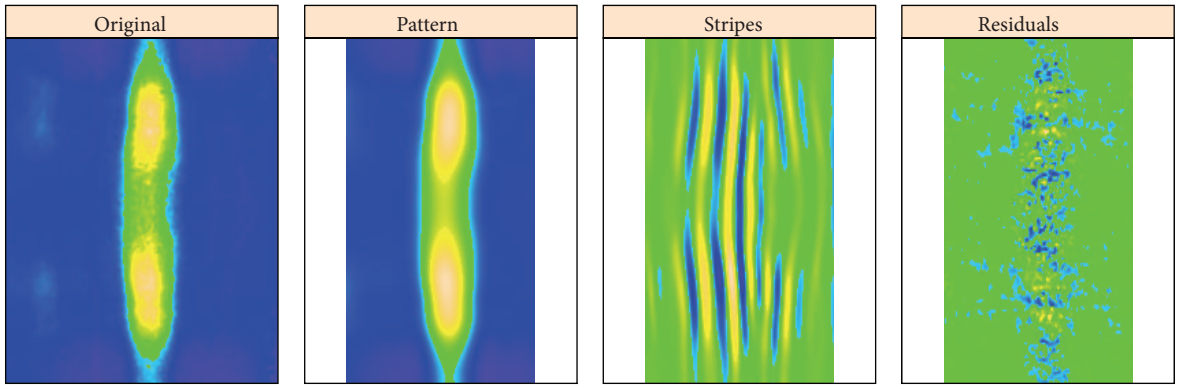


FIGURE 13: *Kr*: processing of the *Kr* expression image by circular 2D-SSA. ((a) to (d)): original image, pattern components (numbers 1–8), stripes (components 9, 10, 13, 15, 20, 25), and residual noise.

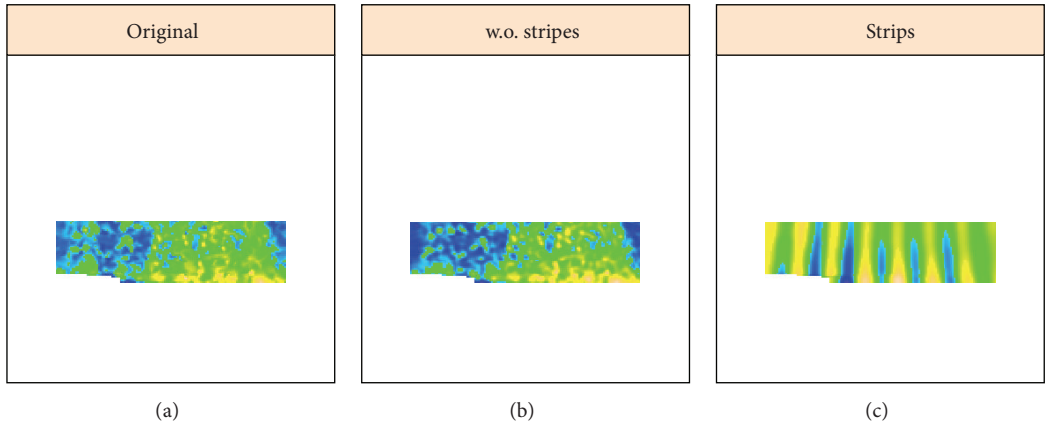


FIGURE 14: *sna* image, area 1, strong expression zone. ((a) to (c)): original image, reconstruction without stripes, and stripe components from the eve marker.

detailed decomposition. This gives more chance to extract a component; however, some components can mix. Therefore, if the data series has a trend with a complex form or has periodicities with complex modulation, then window lengths should be moderate.

These generalities also hold for the case of 2D-SSA. In practice, the difference between 1D and 2D is in the construction of the trajectory matrices, which are quasi-Hankel, in particular Hankel-block-Hankel. The moving window is two-dimensional, for example, a rectangle. In this paper, we introduce circular SSA, for treating rectangles with periodic boundary conditions, for example, data sets on cylindrical geometries. Small window size corresponds to smoothing. We can take into consideration the structure of the image in different directions by choosing different sizes in different directions. The trajectory matrix is constructed from vectorized windows of arbitrary shape moving within the whole image (including circular domains, for periodic boundary conditions).

3.2. Particular Cases. For a rectangular image, with a rectangular window which moves within the image boundaries, we obtain the standard 2D-SSA method. If the image and the window are of arbitrary shape, the shaped version of 2D-SSA is applied [25]. If the window can cross the boundary of the image, we obtain a circular version of 2D-SSA.

For example, let us take an image (a matrix in the mathematical sense)

$$\mathbb{X} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} \tag{2}$$

and the window of size 2×2. Then we have a set of 4 windows in the ordinary version, $\left(\begin{smallmatrix} 1 & 2 \\ 4 & 5 \end{smallmatrix}\right)$, $\left(\begin{smallmatrix} 2 & 3 \\ 5 & 6 \end{smallmatrix}\right)$, $\left(\begin{smallmatrix} 4 & 5 \\ 7 & 8 \end{smallmatrix}\right)$, and $\left(\begin{smallmatrix} 5 & 6 \\ 8 & 9 \end{smallmatrix}\right)$, and two additional windows, $\left(\begin{smallmatrix} 7 & 8 \\ 1 & 2 \end{smallmatrix}\right)$, $\left(\begin{smallmatrix} 8 & 9 \\ 2 & 3 \end{smallmatrix}\right)$, in the circular case. For the circular case, the trajectory matrix will have the form

$$\mathbb{X} = \left(\begin{array}{cc|cc|cc} 1 & 2 & 4 & 5 & 7 & 8 \\ 2 & 3 & 5 & 6 & 8 & 9 \\ \hline 4 & 5 & 7 & 8 & 1 & 2 \\ 5 & 6 & 8 & 9 & 2 & 3 \end{array}\right). \tag{3}$$

One can see that the 2D trajectory matrix consists of trajectory matrices from each matrix's row.

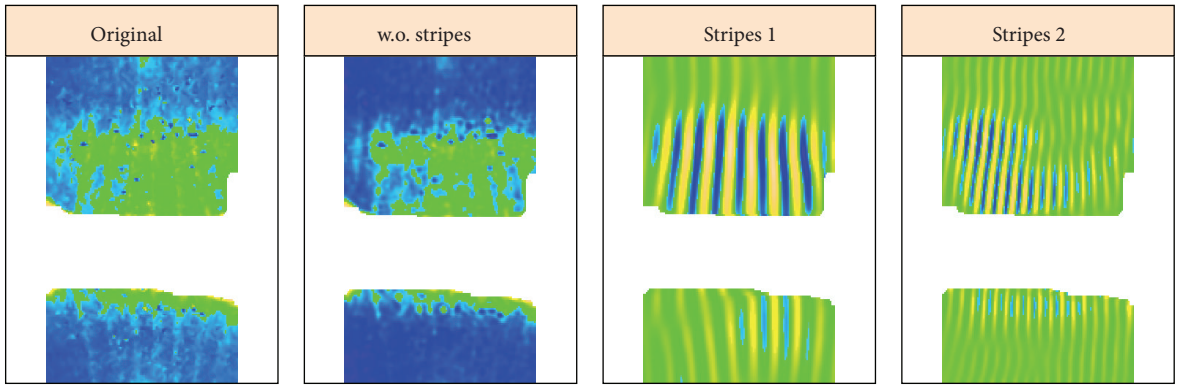


Figure 15: *sna* image, area 2, weak expression zone. ((a) to (d)): original image, reconstruction without stripes, and stripe components.

3.3. Choice of Parameters, Separability, and Component Identification. Approach to the choice of window size for one-dimensional time series is thoroughly described in [13, 26]. Recommendations for 2D objects are more complicated. For extraction of so-called objects of finite rank (sums of products of polynomials, exponentials, and sinusoids), which satisfy linear recurrence relations (LRRs), windows should be large, up to half of the object size. However, real-world patterns usually have complex form and satisfy LRRs only approximately and locally. The window needs to agree with this local character. In particular, sine waves are exactly governed by an LRR. However, if a 2D-sine wave has a slowly changing location, then only its local parts satisfy an LRR. The window sizes need to be in accordance with the scale of this locality. Choice of window size is always a balance between the local and the global scales of the data.

Generally, SSA can separate smooth patterns from noise for a wide variety of patterns. For regular patterns, 2D-SSA can be applied whether the pattern varies smoothly or sharply. However, if the pattern is not regular, variation needs to be smooth in order to use 2D-SSA for signal separation. Irregular pattern with sharp variation is poorly separated by 2D-SSA. If, however, the sharp change occurs in narrow area, this can be cut out, and the remaining data analyzed by shaped SSA, which is a version of 2D-SSA with a nonrectangular shape of the image or the window.

Elementary components are grouped based on their similarity to the data components being extracted. For regular components like sine waves, the number of elementary components can be calculated from theory. Also, patterns usually have a limited frequency range (usually lacking high frequencies). In general, therefore, leading elementary components with the appropriate frequency characteristics are ascribed to pattern.

In this paper we show how 2D-SSA can be used to remove noise, to separate regular oscillations from slowly varying patterns (for correcting erroneous unmixing procedures), and to extract stripes for their further analysis. Shaped SSA allows for the analysis of complex patterns by splitting images into several parts.

Drosophila early gene expression (before the midblastula transition) produces smooth and simple patterns suitable for 2D-SSA processing. A number of web resources have such datasets (BDTNP BID [4], Fly-FISH http://fly-fish.ccbr.utoronto.ca [27], FlyEx http://urchin.spbcas.ru/flyex [28]; see also [29, 30]). Shaped SSA can also be useful for a common subset of this data, in which patterns fall sharply to zero. In these cases, subregions can be excised or analyzed separately from the whole image. The gene *sna* is a typical *Drosophila* example seen in the BDTNP BID; such compact patterns are also seen in other experimental organisms, such as the nine zebrafish genes [31]. We expect 2D-SSA and shaped SSA to therefore have broad applicability to image processing in developmental biology.

The problem of unmixing expression patterns from two different genes in one image [32] requires additional conditions. Specifically, information is needed on the unmixed expression of each gene (i.e., data from one gene in the absence of the other gene). If the two genes have slowly varying patterns, they cannot readily be separated by SSA. In such cases, SSA cannot be used to detect or correct errors in mixed images. However, SSA is an effective unmixing method for cases in which one gene has an approximately regular structure, and this differs from the structure of the other gene. In this paper, we apply SSA to signal unmixing and image correction for such cases from *Drosophila* data.

3.4. Data Preprocessing. Initially, the data for 2D-SSA analysis should be measured on a regular grid. Data for gene expression are measured at nuclei, which are not regularly located on a 3D surface of embryo (which is roughly ellipsoidal in shape). The first step of preprocessing is a cylindrical projection of the data (centred on the major axis of the ellipsoid; the major axis of the embryo is found by principal component analysis). We then interpolate the data to a regular grid on this cylinder. We analyze a central region of the cylinder, in order to avoid corruptions near the poles from the ellipsoid to cylinder transformation. After 2D-SSA decomposition, we interpolated the data back onto the nuclear centers. This

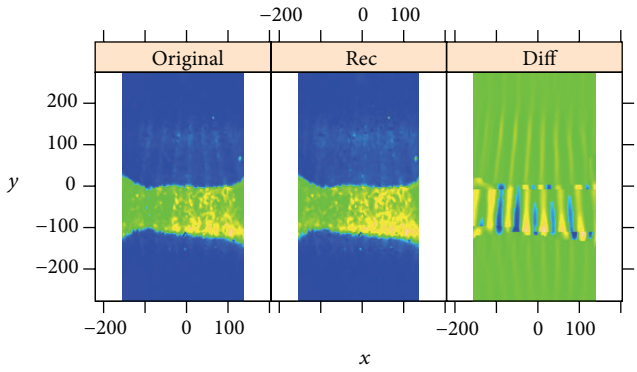


Figure 16: *sna*, combined image (both zones from Figures 14 and 15). ((a) to (d)): original image, reconstruction without stripes, and the difference. BDTNP embryo `v5-s10531-28fe05-07.pce`.

interpolation is performed for smooth components; residuals are calculated as the difference between the initial data and interpolated smooth components.

Interpolation involves Delaunay triangulation followed by linear interpolation of nuclear centers to the triangulation.

3.5. Implementation. The algorithms are implemented in the Rssa and BioSSA packages in *R*. Rssa is a general-purpose package containing effective implementation of singular spectrum analysis and its 2D extensions. 2D-SSA algorithms are time- and memory-consuming and therefore it is very important to have an effective implementation. A description of Rssa with examples can be found in [24, 33]. The *R*-package BioSSA is an addition to Rssa for application to fly embryo gene expressions data and is briefly described at http://biossa.github.io/.

4. Periodic Patterns Produced by Unmixing Algorithms

Different emission spectra for fluorescent probes allows for the simultaneous staining for 3-4 gene products in embryonic tissues. Quantitative imaging projects [4, 30] use the same gene in one of these channels in all embryos, for reliable quantitative comparisons, registration, and so forth. The gene used for this marking in *Drosophila* embryos is commonly one of the pair-rule genes (such as *eve* or *ftz*), which have a characteristic periodic 7-stripe expression pattern.

Multichannel imaging suffers from an inherent problem of overlapping emission spectra (when the fluorescent markers are simultaneously excited (e.g., [34])), where light from more than one fluorescent dye is collected by a given acquisition channel. To computationally reduce this "crosstalk," an automated channel unmixing method was developed and applied to the BDTNP data [32].

The problem with this approach in large scale projects with automatic data processing is that the unmixing parameters can end up being too high or too low. If the parameters are overestimated, unmixing produces an overcorrection, which is manifest as a partial subtraction of the common, reference

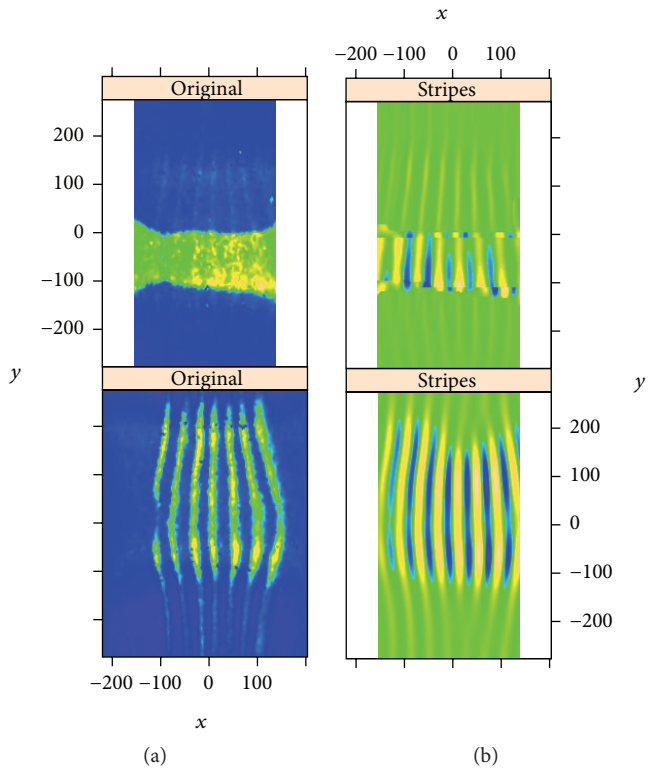


Figure 17: *sna* and *eve*: the original images (a) and the stripes (b), *sna* at top and *eve* at bottom.

pattern from the pattern of the second gene (the gene under study for the embryo). With periodic reference patterns (*eve*, *ftz*), this produces periodic grooves in the "unmixed" pattern. Figure 1 shows the effects of such overcorrection in one of the BDTNP embryos.

On the other hand, if the unmixing parameters are underestimated, unmixing produces an undercorrection, which can be seen as an addition of the common, reference pattern to the pattern of the second gene (that one being studied in the given embryo). Figure 2 shows an example of undercorrection on a BDTNP embryo.

Misestimation of the unmixing parameters can be seen to introduce periodicity in a number of BDTNP embryos from the 7-stripe *eve* or *ftz* reference patterns. The effect is strong enough to be seen in some images integrated from multiple embryos (such as Figure 2).

We now show how decomposition by circular 2D-SSA can be used to estimate and eliminate the periodic components caused by under- or overcorrection, using the examples of the BDTNP images in Figures 1 and 2.

4.1. Circular 2D-SSA, hb Corrupted by ftz, and Strong Overcorrection. Figure 3 shows the original images for *hb* and *ftz* expressions from a BDTNP embryo (ID "`v5-s11512-2oc06-25`"). The natural *hb* trend is of low frequency; the natural pattern of *ftz* is of high frequency; crosstalk, with overcorrection in the unmixing algorithm, "bleeds" the high frequency *ftz* pattern into the *hb* pattern. These images are "unrolled"

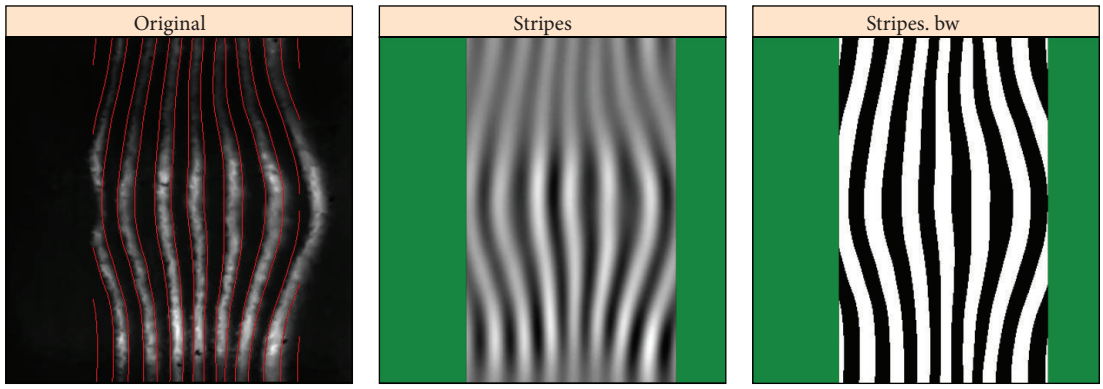


Figure 18: (a) original image, (b) reconstruction of strips, (c) conversion to black and white, according to positive or negative values on the intensity scale; black-white boundaries are shown as red lines on the original image. BDTNP embryo "v5_s10901-20ap06-11s10901."

from the cylindrical projection of the data; therefore, the top and bottom edges connect (periodic boundary conditions).

We preprocess the images by interpolating to a regular grid (step 0.5%) and removing 20% from the left and 5% from the right (to focus on the stripe region). Use of circular 2D-SSA allows us to analyze the cylindrical dataset. We use a rectangular window of 25×10. In consideration of the regular oscillations along the anteroposterior (AP, horizontal) coordinate, the first window dimension, 25, is larger than the second dimension, 10.

Figure 4 presents 2D-SSA decomposition into elementary image components for *hb*; Figure 5 shows this for *ftz* (we depict the 26 largest components; the smaller components were not found to be significant in image reconstruction). Figure 4 contains a number of components with vertical stripes caused by or influenced by the *ftz* channel. If one compares elementary components of the *ftz* decomposition (Figure 5, striped components 2–5, 9–11, and 15–17) with the *hb* decomposition (Figure 4), it appears that *hb* components 1–4 are likely due to expression pattern, while components 5–9, 11, and probably 10, 12 are due to *ftz*-correction.

Figure 6 shows reconstructions from the leading high frequency components for each image, components 5 and 6 from Figure 4, components 2 and 3 from Figure 5. The reconstructions are very similar, but have opposite phases, indicating that the *hb* data was overcorrected. Figure 7 is reconstructed from all striped components for each image; again, the patterns are very similar but of opposite phase.

Simultaneously, with removing stripes, this process also decomposes an image into pattern and noise (residuals): Figure 8 shows reconstruction of *hb* expression from the "unstriped" components 1–4, alongside the striped components (strongly affected by *ftz*) 5–12 and the residuals. Circular 2D-SSA provides a method for removing under- or overcorrection in the unmixing algorithm and therefore of clearing gene patterns from crosstalk effects. For an image without stripes, 2D-SSA produces a direct decomposition into pattern and noise. We show here that SSA decomposition is robust for data with crosstalk stripes.

4.2. Circular 2D-SSA, Kr Corrupted by eve, and Weak Overcorrection. In some cases, crosstalk stripes from the pair-rule reference marker are barely visible in the gene of interest. In these cases, circular 2D-SSA is still effective at removing artefacts from misestimation of the unmixing parameters. Figure 9 shows images from an embryo "v5-s11512-2oc06-25" stained for *Kr* (gene of interest) mRNA and *eve* (reference marker) mRNA. In this case, there is weak overcorrection, with *eve* adding to apparent intensity in the *Kr* image. *Kr*, like *hb* (Figure 3), is a gap gene, with low frequency expression pattern, compared to the high frequency *eve* pair-rule pattern.

We perform the same preprocessing and choose the same method parameters as in Section 4.1. Figure 10 shows circular 2D-SSA (top and bottom edges are contiguous) decomposition into elementary components for *Kr*; Figure 11 shows this for the *eve* image.

The decomposition in Figure 10 shows components with low frequency vertical stripes corresponding to the *Kr* signal, as well as high frequency stripes corresponding to *eve*. These high frequency stripes can be seen in the *eve* decomposition (Figure 11), in particular components 4-5, 7–9, 11, 13, 15, 19, and 20. Conversely, *Kr* crosstalk on the eve image is apparent in Figure 11 in components 9, 10, 13, 15, 20, and 25. Figure 12 shows reconstructions using the stripe components from the images. Again, being a characteristic of overcorrection in the unmixing algorithm, these patterns are of comparable frequency, but of opposite phase.

Figure 13 shows reconstruction of the *Kr* expression pattern from the circular 2D-SSA components. In the analysis of Figure 3, crosstalk overcorrection was strong and evident by eye. In Figure 9, the crosstalk stripes are not as evident by eye, but circular 2D-SSA is still effective for separating signal from the gene of interest (*Kr*) from the striped reference marker. Separation of pattern components leaves residual noise, for studying stochastic effects in gene expression.

4.3. Shaped 2D-SSA, sna Corrupted by eve, and Undercorrection. A number of genes express in patterns which are

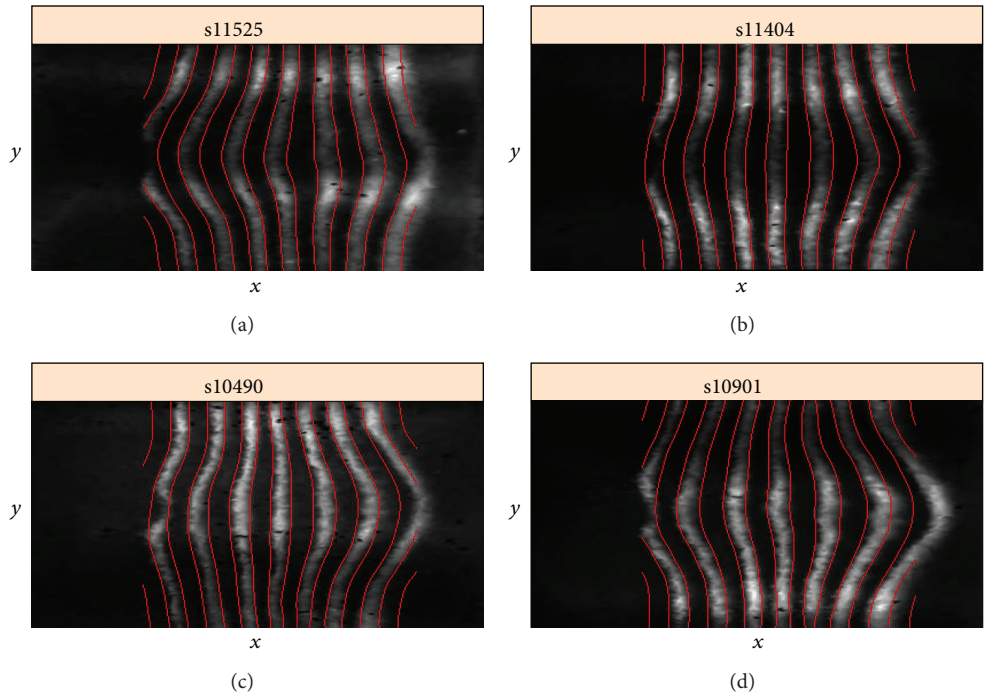


Figure 19: Four cases of the 3D geometry of eve expression stripes. Stripe 4 can be a forward "C"-shape (a), straight (b), a negative "C"-shape (c), or "S"-shaped (d). BDTNP embryo IDs are given on the images.

more complex than the general AP variation seen with gap genes such as *hb* and *Kr*. To analyze crosstalk for such data, we introduce the shaped version of 2D-SSA. As an example, *snail* (*sna*) is expressed in a broad band along the ventral midline of the embryo (Figure 16, v5-s1053l-28fe05-07, cy3_apical). Since *sna* shows a very sharp transition from expressing to nonexpressing regions, we analyzed these separately (Figure 14, expressing; Figure 15, nonexpressing). Analysis was conducted on a regular grid (step 0.5%), clipped 15% from left and right (as for Figure 9). For the central expressing zone (Figure 14), we used a window of 40×10; for the lateral nonexpressing zone (Figure 15), we used a window of 30×10.

Decomposition shows that the elementary components $\{3, 4\}$ (Figure 14) and $\{4, 5, 16, 17\}$ (Figure 15) correspond to stripes, which come from the eve reference marker. Figures 14 and 15 show these stripe components and the effect of removing these stripes to reveal the *sna* signal. Figure 16 shows this for the complete *sna* image (combination of the expressing and nonexpressing zones). In this case, the stripe components from the *sna* image and from the *eve* marker image are in phase, indicating that this is a case of undercorrection in the unmixing algorithm (see Figure 17, where the original images and the stripe reconstructions are put together).

Thus, we have constructed a procedure for removing under- or overcorrections. Note that if an image does not contain stripes, images of elementary components also will not contain stripes and therefore can see if correction is necessary.

5. 3D Geometry of the Early Segmentation Pair-Rule Stripes

As discussed above, the early *Drosophila* embryo is roughly a prolate ellipsoid. Gene expression patterns defining the AP and dorsal-ventral (DV) axes are relatively independent. However, even clearly AP-varying patterns, such as the *eve* and *ftz* pair-rule striped patterns, display some degree of DV variation. This can be affected by deviations from ellipsoidal symmetry (e.g., embryos have a longer ventral surface (or "belly") than dorsal surface) and also from variations in the axial ratio (see [4]).

Embryo-to-embryo variability in *eve* expression in the AP axis has been well documented and discussed in terms of the robustness of the developmental programme. However, such analysis has been in 1D. Analyzing 3D images, for example, with 2D-SSA, reveals new levels of variability.

Figure 18 is an unfolded cylindrical projection of eve expression, showing the DV variation of the 7-stripe pattern, especially as the stripes bend around the ventral "belly" of the embryo (horizontal midline of image). To quantify the stripe geometry, we identify stripe boundaries at the threshold between positive and negative values on the intensity scale.

Using this boundary identification procedure, let us focus on the shape of the central (4th) eve stripe (Figure 19; the 4th stripe has minimal effect from the ellipsoidal to cylindrical projection). Preprocessing included interpolation of the cylindrical projection to a regular grid, clipping 25% of the image on the left and 15% on the right and using a 15×10 window. Applying circular 2D-SSA, we use

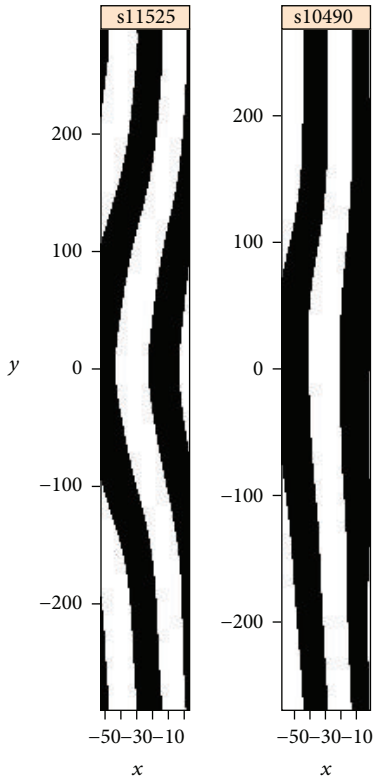


Figure 20: "C" and straight *eve* stripe 4 shapes, shown in black and white. BDTNP embryo IDs given on the images.

components 2 and 3 to represent the striped expression. The 4th stripe is frequently straight across the ventral midline (Figure 19(b)) but can often show curvature as well. Curvature can be "C"-shaped, both forwards (Figure 19(a)) and backwards (Figure 19(c)), or "S"-shaped (Figure 19(d)). For clarity, Figure 20 shows the "C" and straight shapes in black and white and in the original aspect ratio.

Stripe 4 of *eve* is critical for subsequent segmentation events in fly development. These events need to be robust to the curvature variability reported here. It is currently unknown which mechanism might produce this robustness, but it warrants further investigation. For example, what is the correlation between the size of the ventral "belly" of the embryo and stripe 4 curvature? And does this suggest a "shape compensation" such that embryos can develop normally despite variable early geometry? (Systematic analysis should also be done to examine the possible contribution of experimental errors (e.g., fixation procedures) to stripe variability, which may involve comparison with live imaging techniques.)

6. Conclusions

This paper has shown the applicability of our new shaped and circular extensions of 2D-SSA to analyzing embryo images from a quantitative high-throughput project in developmental biology. We have shown that 2D-SSA can decompose images and classify components according to the gene of interest. This is an effective means for reducing the "crosstalk" between gene channels which arises in the imaging technique but can be amplified by the automated postprocessing unmixing algorithm.

Circular 2D-SSA is a critical extension for analyzing cylindrical data projections (accounting for periodic boundaries in "rectangular" images). Shaped 2D-SSA allows for the analysis of subregions of the image, important for analyzing complex expression patterns, complex geometries, and avoiding edge effects.

The procedure is performed under user control and can be adapted to an image's unique structure with a flexible choice of window shapes and sizes. This is currently a manual procedure and future work will focus on reliable automation of the process.

We have demonstrated that 2D-SSA can be used to extract signal and noise from images with both strong and weak over- or undercorrection of crosstalk. This is a significant tool for separating gene expression in multichannel images and for extracting residual noise for studying the stochastic aspects of gene expression. In particular, we have used SSA to separate low frequency genes of interest (the gap genes *hb* and *Kr*, and *sna*) from "bleed-through" crosstalk of the high-frequency pair-rule fiduciary markers (*eve* and *ftz*). In addition, we have shown how SSA components can be used to quantify *eve* stripes (in particular stripe 4) and how this reveals new types of variability in expression, leading to new insights into developmental mechanisms. These are all examples of how 2D-SSA can be applied—we expect them to be broadly generalizable to other cases of multichannel 3D data from *Drosophila* and other organisms.

Conflict of Interests

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

Acknowledgments

This work is supported by the NG13-083 Grant of Dynasty Foundation, US NIH Grant R01-GM072022, and The Russian Foundation for Basic Research Grant 13-04-02137.

References

[1] D. W. Knowles and M. D. Biggin, "Building quantitative, three-dimensional atlases of gene expression and morphology at cellular resolution," *Wiley Interdisciplinary Reviews: Developmental Biology*, vol. 2, no. 6, pp. 767–779, 2013.

[2] J. Schindelin, I. Arganda-Carreras, E. Frise et al., "Fiji: an open-source platform for biological-image analysis," *Nature Methods*, vol. 9, no. 7, pp. 676–682, 2012.

[3] Z. Bao, J. I. Murray, T. Boyle, S. L. Ooi, M. J. Sandel, and R. H. Waterston, "Automated cell lineage tracing in Caenorhabditis elegans," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 8, pp. 2707–2712, 2006.

[4] C. C. Fowlkes, C. L. L. Hendriks, S. V. E. Keränen et al., "A quantitative spatiotemporal atlas of gene expression in the Drosophila blastoderm," *Cell*, vol. 133, no. 2, pp. 364–374, 2008.

[5] X. Liu, F. Long, H. Peng et al., "Analysis of cell fate from single-cell gene expression profiles in C. elegans," *Cell*, vol. 139, no. 3, pp. 623–633, 2009.

[6] J. I. Murray, T. J. Boyle, E. Preston et al., "Multidimensional regulation of gene expression in the C. elegans embryo," *Genome Research*, vol. 22, no. 7, pp. 1282–1294, 2012.

[7] O. Rübel, G. H. Weber, S. V. E. Keränen et al., "Point-cloudxplore: visual analysis of 3d gene expression data using physical views and parallel coordinates," in *Proceedings of the Eurographics/IEEE-VGTC Symposium on Visualization*, pp. 203–210, 2006.

[8] A. Aswani, S. V. E. Keränen, J. Brown et al., "Nonparametric identification of regulatory interactions from spatial and temporal gene expression data," *BMC Bioinformatics*, vol. 11, article 413, 2010.

[9] C. C. Fowlkes, K. B. Eckenrode, M. D. Bragdon et al., "A conserved developmental patterning network produces quantitatively different output in multiple species of *Drosophila*," *PLoS Genetics*, vol. 7, no. 10, Article ID e1002346, 2011.

[10] D. S. Broomhead and G. P. King, "Extracting qualitative dynamics from experimental data," *Physica D: Nonlinear Phenomena*, vol. 20, no. 2-3, pp. 217–236, 1986.

[11] R. Vautard, P. Yiou, and M. Ghil, "Singular-spectrum analysis: a toolkit for short, noisy chaotic signals," *Physica D: Nonlinear Phenomena*, vol. 58, no. 1-4, pp. 95–126, 1992.

[12] J. B. Elsner and A. A. Tsonis, *Singular Spectrum Analysis: A New Tool in Time Series Analysis*, Plenum Press, 1996.

[13] N. Golyandina, V. Nekrutkin, and A. Zhigljavsky, *Analysis of Time Series Structure: SSA and Related Techniques*, Chapman&Hall/CRC, Boca Raton, Fla, USA, 2001.

[14] M. Ghil, M. R. Allen, M. D. Dettinger et al., "Advanced spectral methods for climatic time series," *Reviews of Geophysics*, vol. 40, no. 1, pp. -1–41, 2002.

[15] N. Golyandina and A. Zhigljavsky, *Singular Spectrum Analysis for Time Series*, Springer Briefs in Statistics, Springer, 2013.

[16] D. Danilov and A. Zhigljavsky, *Principal Components of Time Series: The "Cater-pillar", Method*, St.Petersburg Press, 1997 (Russian).

[17] N. E. Golyandina and K. D. Usevich, "2D-extension of singular spectrum analysis: algorithm and elements of theory," in *Matrix Methods: Theory, Algorithms and Applications*, V. Olshevsky and E. Tyrtyshnikov, Eds., pp. 449–473, World Scientific, 2010.

[18] S. Rouquette and M. Najim, "Estimation of frequencies and damping factors by two-dimensional ESPRIT type methods," *IEEE Transactions on Signal Processing*, vol. 49, no. 1, pp. 237–245, 2001.

[19] A. Monadjemi, *Towards efficient texture classification and abnormality detection [Ph.D. thesis]*, University of Bristol, 2004.

[20] S. Trickett, "F-xy cadzow noise suppression," in *Proceedings of the 78th SEG Annual International Meeting*, Expanded Abstracts, pp. 2586–2590, Las Vegas, Nev, USA, November 2008.

[21] D. M. Holloway, F. J. Lopes, L. da Fontoura Costa et al., "Gene expression noise in spatial patterning: *hunchback* promoter structure affects noise amplitude and distribution in *Drosophila* segmentation," *PLoS Computational Biology*, vol. 7, no. 2, Article ID e1001069, 2011.

[22] P. J. Shin, P. E. Z. Larson, M. A. Ohliger et al., "Calibrationless parallel imaging reconstruction based on structured low-rank matrix completion," *Magnetic Resonance in Medicine*, vol. 72, no. 4, pp. 959–970, 2014.

[23] N. Golyandina, I. Florinsky, and K. Usevich, "Filtering of digital terrain models by 2d singular spectrum analysis," *International Journal of Ecology & Development*, vol. 8, no. F07, pp. 81–94, 2007.

[24] N. Golyandina, A. Korobeynikov, A. Shlemov, and K. Usevich, "Multivariate and 2D extensions of singular spectrum analysis with the Rssa package," This paper is accepted by Journal of Statistical Software, http://arxiv.org/abs/1309.5050.

[25] A. Shlemov and N. Golyandina, "Shaped extensions of singular spectrum analysis," in *Proceedings of the 21st International Symposium on Mathematical Theory of Networks and Systems*, pp. 1813–1820, Groningen, The Netherlands, July 2014.

[26] N. Golyandina, "On the choice of parameters in singular spectrum analysis and related subspace-based methods," *Statistics and its Interface*, vol. 3, no. 3, pp. 259–279, 2010.

[27] E. Lécuyer, H. Yoshida, N. Parthasarathy et al., "Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function," *Cell*, vol. 131, no. 1, pp. 174–187, 2007.

[28] A. Pisarev, E. Poustelnikova, M. Samsonova, and J. Reinitz, "FlyEx, the quantitative atlas on segmentation gene expression at cellular resolution," *Nucleic Acids Research*, vol. 37, no. 1, pp. D560–D566, 2009.

[29] J. Jaeger, D. H. Sharp, and J. Reinitz, "Known maternal gradients are not sufficient for the establishment of gap domains in *Drosophila melanogaster*," *Mechanisms of Development*, vol. 124, no. 2, pp. 108–128, 2007.

[30] S. Surkova, D. Kosman, K. Kozlov et al., "Characterization of the *Drosophila segment* determination morphome," *Developmental Biology*, vol. 313, no. 2, pp. 844–862, 2008.

[31] C. Castro-Gonzlez, M. Luengo-Oroz, L. Duloquin et al., "A digital framework to build, visualize and analyze a gene expression atlas with cellular resolution in Zebrafish early embryogenesis," *PLoS Computational Biology*, vol. 10, no. 6, Article ID e1003670, 2014.

[32] C. L. Luengo Hendriks, S. V. E. Keränen, M. D. Biggin, and D. W. Knowles, "Automatic channel unmixing for high-throughput quantitative analysis of fluorescence images," *Optics Express*, vol. 15, no. 19, pp. 12306–12317, 2007.

[33] N. Golyandina and A. Korobeynikov, "Basic singular spectrum analysis and forecasting with R," *Computational Statistics and Data Analysis*, vol. 71, pp. 934–954, 2014.

[34] C. L. L. Hendriks, S. V. E. Keränen, C. C. Fowlkes et al., "Three-dimensional morphology and gene expression in the *Drosophila* blastoderm at cellular resolution I: data acquisition pipeline," *Genome Biology*, vol. 7, no. 12, p. R123, 2006.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 308475, 7 pages
http://dx.doi.org/10.1155/2015/308475

Research Article

Effect of Celastrol on Growth Inhibition of Prostate Cancer Cells through the Regulation of hERG Channel *In Vitro*

Nan Ji,[1,2] Jinjun Li,[1] Zexiong Wei,[3] Fanhu Kong,[3] Hongyan Jin,[3] Xiaoya Chen,[1] Yan Li,[4] and Youping Deng[1]

[1] *Medical College, Wuhan University of Science and Technology, Wuhan 430065, China*
[2] *The Outpatient Department, The Second Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou 310009, China*
[3] *Puren Hospital Affiliated to Wuhan University of Science and Technology, Wuhan 430081, China*
[4] *Department of Internal Medicine, Rush University Medical Center, Chicago, IL 60121, USA*

Correspondence should be addressed to Jinjun Li; entry2003@126.com and Youping Deng; youpingd@gmail.com

Received 30 June 2014; Accepted 28 July 2014

Academic Editor: Hongwei Wang

Objective. To explore the antiprostate cancer effects of Celastrol on prostate cancer cells' proliferation, apoptosis, and cell cycle distribution, as well as the correlation to the regulation of hERG. *Methods.* DU145 cells were treated with various concentrations of Celastrol (0.25–16.0 μmol/L) for 0–72 hours. MTT assay was used to evaluate the inhibition effect of Celastrol on the growth of DU145 cells. Cell apoptosis was detected through both Annexin-V FITC/PI double-labeled cytometry and Hoechst 33258. Cell cycle regulation was examined by a propidium iodide method. Western blot and RT-PCR technologies were applied to assess the expression level of hERG in DU145 cells. *Results.* Celastrol presented striking growth inhibition and apoptosis induction potency on DU145 cells *in vitro* in a time- and dose-dependent manner. The IC_{50} value of Celastrol for 24 hours was 2.349 ± 0.213 μmol/L. Moreover, Celastrol induced DU145 cell apoptosis in a cell cycle-dependent manner, which means Celastrol could arrest DU145 cells in G_0/G_1 phase; accordingly, cells in S phase decreased gradually and no obvious changes were found in G_2/M phase cells. Through transmission electron microscope, apoptotic bodies containing nuclear fragments were found in Celastrol-treated DU145 cells. Overexpression of hERG channel was found in DU145 cells, while Celastrol could downregulate it at both protein and mRNA level in a dose-dependent manner ($P < 0.01$). *Conclusions.* Celastrol exhibits its antiprostate cancer effects partially through the downregulation of the expression level of hERG channel in DU145 cells, suggesting that Celastrol may be a potential agent against prostate cancer with a mechanism of blocking the hERG channel.

1. Foreword

Ion channels exist widely in all cells in a variety of physiological functions. Cancer is also associated with ion channel dysfunction. Research [1] recently suggested that some potassium channels (voltage-gated potassium channels, KV) are related to the occurrence and development of malignant tumors, and the relationship between voltage-gated potassium channels and tumor has become a research hotspot. The human EAG gene (human ether-à-go-go related gene, HERG) encodes the HERG protein A subunit of delayed rectifier potassium channel. Researches have found that [2] high expression of HERG protein in tumor cells has a widespread impact on the biological behavior of tumors and is closely related to the differentiation and invasion of tumor cell proliferation and apoptosis [3–5]. There are reports that hERG protein can affect the tumor cell membrane potential in the depolarized state, which is conducive to tumor cell survival, proliferation, and invasion [5]. Therefore, hERG potassium channel will become a promising target for cancer therapy in the selection of specific molecular targeted agents that play an important role in the process. Celastrol (CSL) is one of the main active components extracted from the traditional Chinese medicine *Tripterygium wilfordii*. A general and three-terpene pigment monomer, Celastrol, is found early to have anti-inflammatory and antitumor pharmacological effects on liver cancer, colon cancer, lung cancer, leukemia, cancer of the esophagus, brain cancer, bladder cancer, and so

on. Much tumor cell growth and proliferation were inhibited [6, 7]. But its antitumor mechanism has only in recent years seen sporadic reports. Its role in the regulation of hERG potassium channel protein has not been reported. Therefore in this paper, with hERG potassium channel as the molecular target, the effect of different concentration of tripterine on its regulation and tripterine antitumor effect correlation will be investigated.

2. Materials and Methods

2.1. Drugs and Reagents. Tripterine, molecular formula (C29-H38O4), a molecular weight of 450, and purity more than 95%, was purchased from USA Calbiochem company, dissolved in two dimethyl sulfoxide (DMSO), and kept at −20°C, before using DMEM culture medium diluted to final concentration. DMEM medium and fetal calf serum were from the USA Gibco BRL company. Annexin-V/PI apoptosis kits were purchased from Wuhan Boster Engineering Co., Ltd. Rabbit anti-human hERG1 monoclonal antibody was purchased from Sigma company, HRP standard. Sheep anti free, two anti purchased from USA Santa Cruz products. Bio-Rad protein assay kit, TRIzol kit, and ECL lighting kits originated from Sweden Amersham company. RT-PCR kit was purchased from Fermentas company. Primers were synthesized by Shanghai Sangon company.

2.2. Cell Lines and Cell Culture. Prostate cancer cell DU145 was obtained from Tongji Medical College, Department of Immunology. Medium contained 10% fetal bovine serum, penicillin 100 IU/mL, and streptomycin 100 g/mL DMEM, at 37°C, 5% CO_2, and water saturated humidity condition. Every 1~2 days, for a fluid passage, the logarithm growth period of cell activity of more than 98% cells was used in this study.

2.3. Effects of Tripterygium wilfordii Red MTT Method to Detect the Proliferation of DU145 Cells. Logarithmic growth phase DU145 cell experiments, cells per hole 2×10^5/mL cells were seeded in 96-well plate, adding different concentrations of celastrol (0.25–16.0 μmol/L), with the culture solution containing equal volume DMSO as blank control, each drug concentration group with 3 holes, each hole total volume 200 μL. For 5% CO_2, 37°C incubator culture 24–72 h, each hole with 5 mg/mL MTT 20 μL reagent, 37°C incubate for 4 h, carefully to absorb the air in culture supernatant, DMSO solution was added to each well of 150 μL, 10 min oscillation, crystallize fully dissolved in the Bio-Rad M450 enzyme labeled each hole optical density were measured in 492 nm wavelength 1.25 meter value (OD), with each experiment repeated 1.273 times, to calculate the inhibitory rate of cell proliferation. Proliferation inhibition rate (%) = (1 − experimental group, control group od/OD) × 100%.

2.4. Annexin-V/PI Staining to Detect Apoptosis. Operate according to kit, divided into the experimental group and the control group with single standard. A collection of different concentration of tripterine (1, 2 and 4 μmol/L) DU145 cells and blank treated cells in control group, with 4°C ice cold PBS wash 2 times, and then to 1×10^6/mL cell density weight suspended from 100 μL binding buffer, adding 5 μL Annexin-V and 10 μL PI dye solution, mix gently, light reaction temperature 15 min, add 300 μL of the buffer solution, detection of it within 1 h.

2.5. Hoechst 33258 Staining for Detection of Cell Morphology of Apoptosis. The logarithmic growth phase DU145 cell was 5×10^5 cells, join the 6 hole plate with cover glass. Blank group and tripterine IC_{50} 100 μmol/L dosing experiment group, after 24 h incubation, washed two times with PBS, then with fixed liquid (methanol : acetic acid = 3 : 1) fixed 15 min, washed two times with PBS; 37°C Hoechst 33258 staining solution (5 mg/L) staining of 15~30 min, PBS wash two times; neutral resin sheet, fluorescent cell morphology was observed under microscope and photographed. Mirror were found 5 does not repeat, apoptotic cell count per 200 cells, the percentage of apoptotic cells is the apoptosis rate.

2.6. The Cell Cycle Was Detected by Flow Cytometry. Collection of the treated DU145 cell was 1×10^6, washing with PBS buffer 2 times, with 70% cold ethanol at 4°C fixed overnight, centrifugation, washing 1 time with PBS, adding 20 μL RNase A at 37°C water bath for 30 min and then adding 300~500 μL PI dye solution mixing and placing 4°C under dark condition for 30 min; the cell cycle was detected by flow cytometry, with red fluorescent wavelength at 488 nm.

2.7. Effects of Tripterygium wilfordii by Semiquantitative RT-PCR Detection of Red Pigment on the Expression of HERG Gene in DU145 Cells. TRIzol kit was used to extract to-tal cellular RNA synthesis of cDNA, according to the instructions. In the first chain cDNA cells were used as template, PCR reaction. PCR primer was synthesized by Sangon company in Shanghai, of which hERG gene upstream primer was 5′-CAGCGGCTGTACTCGGGCACAG-3′, downstream primer was 5′-CAGAAGTGGTCGGAGAA-CTC-3′, amplified fragment is 345 bp; 3-glyceraldehyde phosphate dehydrogenase gene (GAPDH) upstream primer is 5′-GATTTGGTCGTATTGGGGCGC-3′, downstream primer is 5′-CAGAGATGACCCTTTTGGCTCC-3′, amplified fragment is 136 bp. The PCR amplification conditions were 95°C denaturing 5 min, 94°C 1 min, 55°C 50 s, 72°C 1 min, cycle 35, 72°C 10 min end reaction. PCR products were detected by 1.5% agarose gel electrophoresis, UV photography, and scanning analysis, the hERG/GAPDH expression of hERG semiquantitative analysis of the level of.

2.8. Detection of Western Blot Methods. Different concentrations of tripterine treated DU145 cells and control cells, with cell lysate 100 μL pre-cooling (configuration according to molecular cloning method) the ice cracking 30 min, extraction of total cellular protein, quantitative protein by Lowry method. Conventional adhesive preparation, sampling, protein electrophoresis, and then transferring to the membrane were done. Rabbit anti-human hERG1 monoclonal antibodies were added (1 : 1500), 4°C overnight incubation. Rinse after adding HRP-labeled goat anti-rabbit IgG (1 : 2000), at 37°C

with shaking and incubation for 1 h. Finally, ECL chemiluminescence reagent, X-ray exposure and development, analysis of computer software. Each concentration was repeated 3 times, taking the mean measurement results.

2.9. Statistical Analysis. Experimental data $\overline{X} \pm S$, among groups, were compared using F test, SPSS 11.5 statistical software analysis.

3. Results

3.1. Effects of Tripterine on Proliferation of DU145 Cells. It can be seen from Figure 1, respectively, by 0.25, 0.5, 1, 2, 4, 8, and 16 μmol/L tripterine in DU145 cells after 24~72 h, that cell proliferation activity in different cell groups was lower than that of control cells, but in a concentration less than 1 μmol/L, proliferation effects of tripterine on DU145 cells are small, and when the tripterine concentration reached 1 μmol/L, the proliferation inhibition activity was significantly increased, with significant difference ($P < 0.05$). And, with the increase of tripterine drug concentration and action time, the inhibitory effects of proliferation were enhanced, and an obvious time dose effect relationship is apparent. The 24 h IC_{50} value was 2.349 ± 0.213 μmol/L.

3.2. Effects of Tripterine on Apoptosis of DU145 Cells. The Hoechst 33258 results can be seen. DU145 cells of 2 μmol/L tripterine acid treatment in typical apoptosis morphological changes are as follows: cytoplasmic density, chromosome condensation, marginalization, nuclear condensation, and formation of apoptotic bodies increased (Figure 2(a)). Then the Annexin-V/PI double staining (Figure 2(b)) quantitative detection results show that, with the increase of tripterine concentration, apoptosis of DU145 cells gradually increases the proportion, respectively, (6.57 ± 0.11)%, (11.02 ± 3.10)%, and (23.23 ± 1.56)% and the control group (2.24 ± 1.08)% in comparison and the difference had statistical significance.

3.3. Effects of Celastrol on the Cell Cycle of DU145 Cells. With different concentrations of tripterine and the role of DU-145 cells after 24 h, the cell cycle distribution has also been changed, as shown in Figure 3. With the increase of tripterine concentration, the percentage of G_0/G_1 phase cells increased gradually, followed by (39.95 ± 1.88)%, (40.48 ± 2.34)%, (51.40 ± 1.96)%, and (68.58 ± 2.89)%, while the percentage of cells in S phase was dose-dependently reduced, followed by (49.45 ± 1.67)%, (46.19 ± 1.86)%, (37.67 ± 2.03)%, and (23.29 ± 1.52)%; in contrast, Celastrol effects on G_2/M cells were not significant (Table 1). This indicates that Celastrol induced DU145 cell cycle arrest occurs in G_0/G_1.

3.4. Regulatory Effect of Tripterine on DU145 Cells of hERG Potassium Channel Protein. Compared with normal mononuclear cells, the presence of hERG potassium channel protein expression levels was higher in DU145 cells and the 0.5~2.0 μmol/L tripterine was treated for 24 h. The protein expression had a concentration dependent decline, with a statistically significant difference ($P < 0.05$). In order to further clarify the role of *Tripterygium wilfordii* red on the hERG protein, we examined the changes of hERG protein and mRNA content in the level of gene transcription. Similarly, hERG potassium channel protein level of mRNA was dose-dependently downregulated and obviously higher than the mononuclear cells of normal hERG gene expression level (Figure 4).

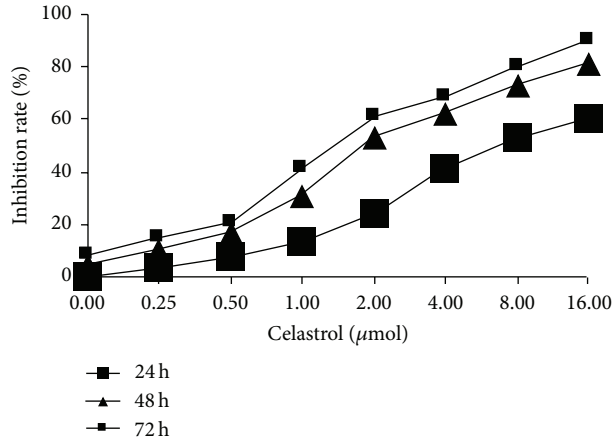
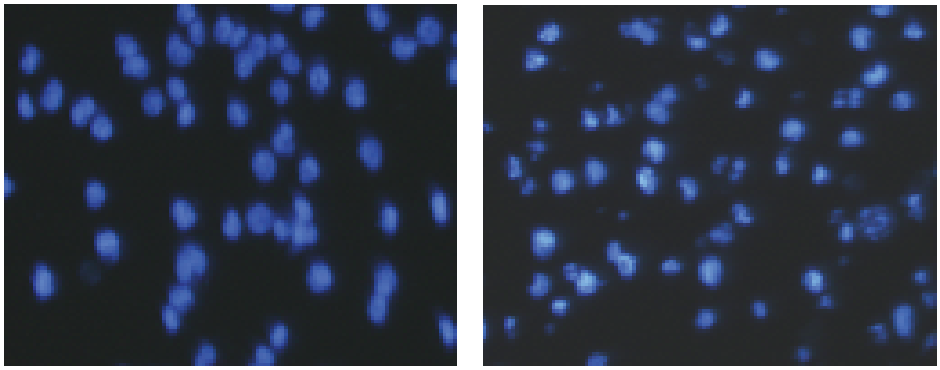


Figure 1: Effects of Celastrol on the proliferation of DU145 cells.

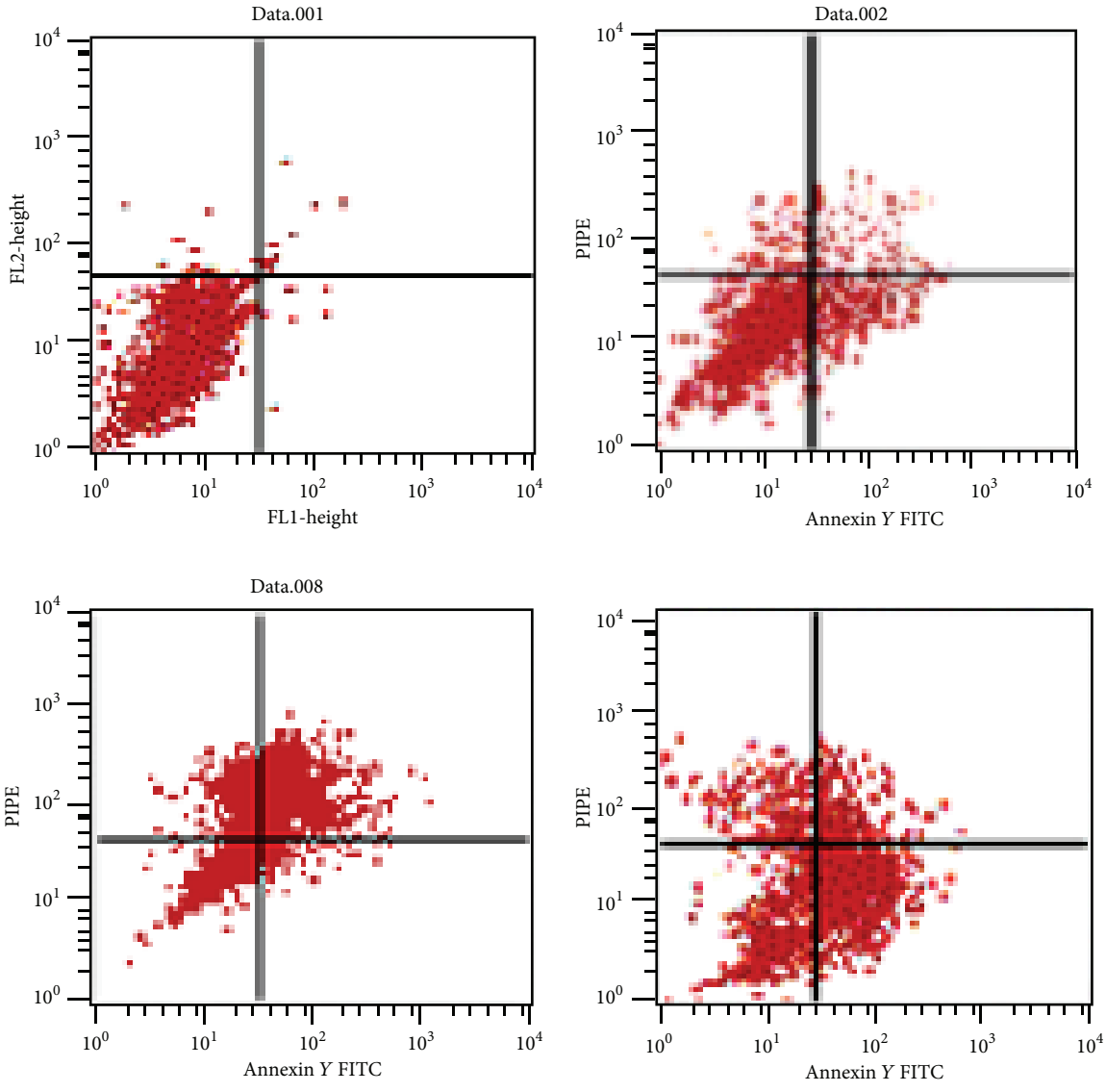
4. Discussion

People have found that many natural preparations, especially in plants and food components, have significant antitumor activity *in vitro* and *in vivo*. In the early 1970s, there were reports of tripterine having anti-inflammatory, analgesic, antioxidant, and antiviral effects, and inducing apoptosis of tumor cells is more positive. Further, to determine the effective components of gambogic acid differently from the general characteristics of anticancer drugs, it can selectively kill tumor cells, but normal hematopoietic cells and heart, liver, kidney, and other organs showed no obvious damage, so it is considered to be a safe and effective antitumor drug [8, 9] for long-term continuous use. The antitumor mechanism has had some scattered reports, but studies in prostate cancer are very rare. In this experiment, cultured prostate cancer cell line DU145 is used as the research object to observe the effect of tripterine on DU145 cell growth inhibition and apoptosis induction and to explore its possible molecular mechanism.

The results show that Celastrol can inhibit the proliferation of DU145 cells and the inhibition is associated with the duration of drug action and drug concentration. At the same time, Celastrol can induce apoptosis of DU145 cells through strong 0.5~2.0 μmol/L tripterine treated for 24 h, the apoptosis rate of DU145 cells increased significantly, and the typical morphological changes of cell apoptosis emerged. The Celastrol-induced apoptosis of cycle arrest effect may be related to induction of closely related events. With the increase of tripterine concentration, the percentage of G_0/G_1 phase cells increased gradually and the percentage of cells in S phase decreased gradually and the tripterine had little effect on the percentage of cells in G_2/M phase. That Celastrol

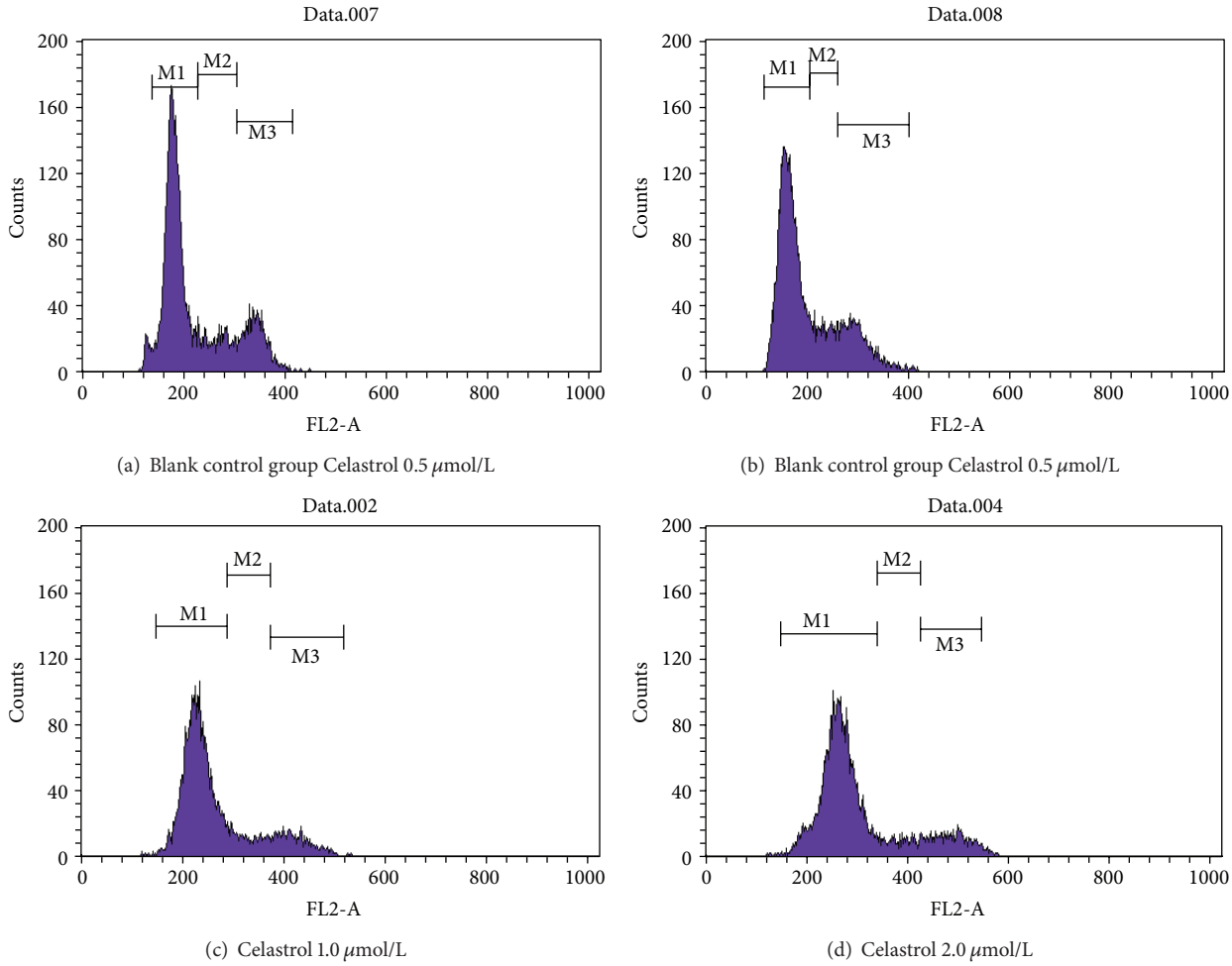


(a) Blank control group Celastrol 2.0 μmol/L



(b) Blank control group Celastrol 0.5 μmol/L

Figure 2: (a) Apoptosis morphological changes of DU145 cells induced by Celastrol with Hoechst 33258 (24 h). (b) Effects of Celastrol on cell apoptosis with Annexin-V FITC/PI assay. Cells were treated with various concentrations of Celastrol for 24 h.



(a) Blank control group Celastrol 0.5 μmol/L

(b) Blank control group Celastrol 0.5 μmol/L

(c) Celastrol 1.0 μmol/L

(d) Celastrol 2.0 μmol/L

Figure 3: Effects of Celastrol on cell cycle distribution.

Table 1: Effects of Celastrol on cell cycle distribution and early apoptosis ($n = 3$, ±S).

Celastrol	Cell cycle (%)			Apoptosis rate (%)
(μmol/L)	G_1/G_0	S	G_2/M	Sub-G_1
0 (control)	39.95 ± 1.88	49.45 ± 1.67	10.60 ± 1.33	2.88 ± 0.33
0.5	40.48 ± 2.34	46.19 ± 1.86	13.33 ± 1.84	1.69 ± 0.24
1.0	51.40 ± 1.96**	37.67 ± 2.03**	10.93 ± 0.98	3.25 ± 0.78
2.0	68.58 ± 2.89**	23.29 ± 1.52**	8.13 ± 1.02	13.77 ± 2.15**

** $P < 0.01$ versus control group.

operating mainly by blocking DU145 cells at G_0/G_1 phase to have an apoptosis inducing effect has been reported in the literature [10].

Celastrol has become a hot spot of oncology field expression of ion channel proteins in normal and tumor cells. The ion channel proteins in numerous studies, tumor cell lines by hERG potassium channel protein encoded by the hERG gene as selective surface in different tissues and primary tumor cells, and tumor cell proliferation, differentiation, apoptosis, invasion, and sensitivity to chemotherapy are closely related and are considered to be the molecular target in cancer cells more specifically. In general, the hERG gene was only expressed in the early stages of embryonic development and followed by the inward rectifier potassium channel current it was replaced by [11]. HERG potassium channels are voltage gated ion channels typical in mammals. hERG due to the voltage dependence of fast inactivation exhibited strong inward rectification activities and played an important role in maintaining the differentiation and physiological function of normal heart rhythm and neurons [12, 13]. In addition, hERG potassium channel is involved in a variety of ventricular arrhythmias and may be the cause of congenital long QT syndrome (LQTS) which is one of the major virulence factors [14]. As mentioned before, the hERG potassium

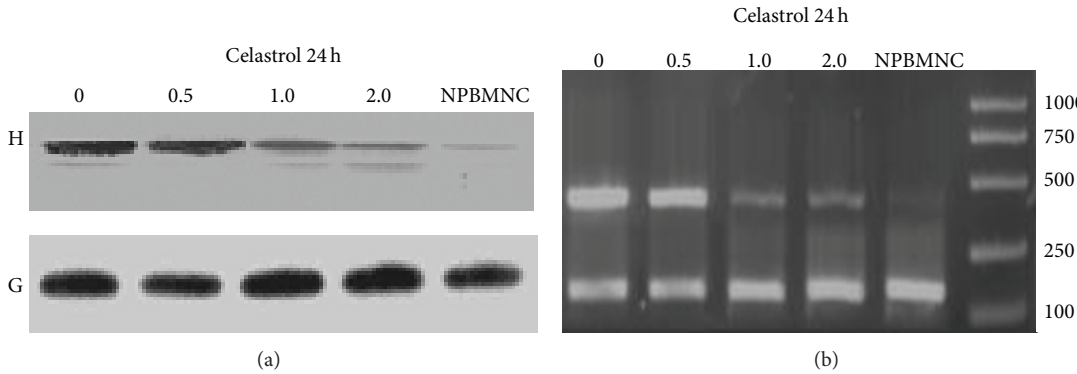


Figure 4: (a) Effects of Celastrol on the expression of hERG protein in DU145 cells and normal mononuclear cells with various concentrations for 24 h. (b) Effects of Celastrol on the expression of hERG mRNA in DU145 cells and normal mononuclear cells with various concentrations for 24 h.

channel protein is crucial in maintaining the cell resting membrane electric's localization in the polarization state, but new research shows that the proliferation characteristics and cell membrane limited tumor cell depolarization are closely related to the status, suggesting the protein and tumor cell proliferation activity of hERG potassium channel. And successive studies have confirmed that hERG protein was highly expressed in tumor cells and primary cells of various tissues of endometrial cancer, colon cancer, and neuroblastoma derived, while in the normal tissues or cells in the corresponding source, no expression or low expression of [4, 5] was found. In addition, a variety of tumor cells, which include the hERG potassium channel proteins, are also present in prostate cancer cells with high expression [3]. Blockade of the hERG potassium channel proteins by specific agents can inhibit the proliferation and metastasis of tumor cells and the corresponding [15, 16], increasing the sensitivity of tumor cells to chemotherapeutic drugs. At present, hERG potassium channel protein with tumor necrosis factor, integrin receptor, VEGF protein interaction, and active cancer protein [17–20] have been studied. It is not difficult to see that hERG potassium channel is a promising target for cancer therapy. Therefore, in this experiment, hERG gene as a target observe the red Chinese medicine *Tripterygium wilfordii* on DU145 cells of hERG potassium channel protein regulation. The results show that, compared with the mononuclear cells of normal control, the expression of hERG potassium channel protein levels rises to higher DU145 cell memory, while the normal mononuclear cells hardly show expression. For the effect of tripterine intervention, the protein and gene expression levels were concentration-dependent underground tune. The effect of tripterine on proliferation of DU145 cells and the intracellular expression of hERG potassium channel are closely related. Although confirmed in tumor cells, inhibition of hERG potassium channel protein expression can inhibit the proliferation of tumor cells, inducing tumor cell apoptosis. Tripterine is expected to become the hERG potassium channel protein inhibitor of a new generation.

Effects of hERG potassium channel protein on the biological behavior of the tumor that can inhibit the growth of tumor cells by inhibiting the expression or channel current of IhERG and promote tumor cell differentiation or apoptosis, reduce its invasiveness. There lays a good foundation for tumor targeting therapy and drug screening.

Conflict of Interests

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

Acknowledgment

This work was supported by research grants from the Natural Science Foundation of Hubei Province of China (2011-CDB236) to Youping Deng.

References

[1] W. Ping and W. Bo, "Study on the relationship between potassium channels and tumor progression of," *International Journal of Pathology and Clinical Medicine*, vol. 25, no. 6, pp. 488–491, 2005.

[2] W. Ping and W. Bo, "TASK-3 potassium ion channels of," *International Journal of Pathology and Clinical Medicine*, vol. 26, no. 3, pp. 225–231, 2006.

[3] S. Pillozzi, M. F. Brizzi, M. Balzi et al., "HERG potassium channels are constitutively expressed in primary human acute myeloid leukemias and regulate cell proliferation of normal and leukemic hemopoietic progenitors," *Leukemia*, vol. 16, no. 9, pp. 1791–1798, 2002.

[4] E. Afrasiabi, M. Hietamäki, T. Viitanen et al., "Expression and significance of HERG (KCNH2) potassium channels in the regulation of MDA-MB-435S melanoma cell proliferation and migration," *Cellular Signalling*, vol. 22, no. 1, pp. 57–64, 2010.

[5] I. Staudacher, L. Wang, X. Wan et al., "hERG K^+ channel-associated cardiac effects of the antidepressant drug desipramine," *Naunyn-Schmiedeberg's Archives of Pharmacology*, vol. 383, no. 2, pp. 119–139, 2011.

[6] H. Li, Y. Y. Zhang, X. Y. Huang et al., "Benificia1 effect of tripterine on systemic lupus erythmatosus induced by active chromatin in BALB/c mice," *European Journal of Pharmacology*, vol. 11, no. 2, pp. 231–237, 2005.

[7] S. Abbas, A. Bhoumik, R. Dahl et al., "Preclinical studies of celastrol and acetyl isogambogic acid in melanoma," *Clinical Cancer Research*, vol. 13, no. 22, pp. 6769–6778, 2007.

[8] Y. Dai, J. T. DeSano, Y. Meng et al., "Celastrol potentiates radiotherapy by impairment of DNA damage processing in human prostate cancer," *International Journal of Radiation Oncology Biology Physics*, vol. 74, no. 4, pp. 1217–1225, 2009.

[9] I. Dogan, A. Cumaoglu, A. Aricioglu, and A. Ekmekci, "Inhibition of ErbB2 by Herceptin reduces viability and survival, induces apoptosis and oxidative stress in Calu-3 cell line," *Molecular and Cellular Biochemistry*, vol. 347, no. 1-2, pp. 41–51, 2011.

[10] A. Arcangeli, O. Crociani, E. Lastraioli, A. Masi, S. Pillozzi, and A. Becchetti, "Targeting ion channels in cancer: a novel frontier in antineoplastic therapy," *Current Medicinal Chemistry*, vol. 16, no. 1, pp. 66–93, 2009.

[11] C. Gessner, R. Macl-klene, J. C. Starkus et al., "The amiodarone derivative KBl 30015 activates hERGl potassium channels via a navel mechanism," *The European Journal of Pharmacology*, vol. 632, no. 1–3, pp. 52–59, 2010.

[12] P. L. Smith, T. Baukrowitz, and G. Yellen, "The inward rectification mechanism of the HERG cardiac potassium channel," *Nature*, vol. 379, no. 6568, pp. 833–836, 1996.

[13] C. Jiang, D. Atkinson, J. A. Towbin et al., "Two long QT syndrome loci map to chromosomes 3 and 7 with evidence for further heterogeneity," *Nature Genetics*, vol. 8, no. 2, pp. 141–147, 1994.

[14] M. C. Sanguinetti and M. Tristani-Firouzi, "hERG potassium channels and cardiac arrhythmia," *Nature*, vol. 440, no. 7083, pp. 463–469, 2006.

[15] S. Z. Chen, S. H. Zhang, J. H. Gong, and Y. S. Zhen, "Erythromycin inhibits the proliferation of HERG K^+ channel highly expressing cancer cells and shows synergy with anticancer drugs," *Zhonghua Yi Xue Za Zhi*, vol. 86, no. 47, pp. 3353–3357, 2006.

[16] S. Z. Chen, M. Jiang, and Y. S. Zhen, "HERG K+ channel expression-related chemosensitivity in cancer cells and its modulation by erythromycin," *Cancer Chemotherapy and Pharmacology*, vol. 56, no. 2, pp. 212–220, 2005.

[17] H. Lin, J. Xiao, X. Luo et al., "Overexpression HERG K+ channel gene mediates cell-growth signals on activation of oncoproteins SP1 and NF-κB and inactivation of tumor suppressor Nkx3.1," *Journal of Cellular Physiology*, vol. 212, no. 1, pp. 137–147, 2007.

[18] A. Cherubini, G. Hofmann, S. Pillozzi et al., "Human ether-a-go-go-related gene 1 channels are physically linked to β 1 integrins and modulate adhesion-dependent signaling," *Molecular Biology of the Cell*, vol. 16, no. 6, pp. 2972–2983, 2005.

[19] A. Masi, A. Becchetti, R. Restano-Cassulini et al., "hERG1 channels are overexpressed in glioblastoma multiforme and modulate VEGF secretion in glioblastoma cell lines," *British Journal of Cancer*, vol. 93, no. 7, pp. 781–792, 2005.

[20] S. Pillozzi, M. F. Brizzi, P. A. Bernabei et al., "VEGFR-1 (FLT-1), β1 integrin, and hERG K^+ channel for a macromolecular signaling complex in acute myeloid leukemia: role in cell migration and clinical outcome," *Blood*, vol. 110, no. 4, pp. 1238–1250, 2007.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 207347, 8 pages
http://dx.doi.org/10.1155/2015/207347

Research Article

The Expression and Distributions of ANP32A in the Developing Brain

Shanshan Wang,[1] Yunliang Wang,[2] Qingshan Lu,[3] Xinshan Liu,[1] Fuyu Wang,[4] Xiaodong Ma,[4] Chunping Cui,[5] Chenghe Shi,[5] Jinfeng Li,[2] and Dajin Zhang[5]

[1] *Weifang Medical University, Weifang 261042, China*
[2] *Department of Neurology, The 148th Hospital, Zibo 255300, China*
[3] *Zhengzhou University, Zhengzhou 450001, China*
[4] *Department of Neurosurgery, PLA 301 Hospital, Beijing 100853, China*
[5] *Center for Basic Medical Sciences, Navy General Hospital of Chinese PLA, Beijing 100048, China*

Correspondence should be addressed to Jinfeng Li; lijinfeng148@163.com and Dajin Zhang; dajinzhang@sina.com

Received 9 July 2014; Revised 20 August 2014; Accepted 20 August 2014

Academic Editor: Hongwei Wang

Acidic (leucine-rich) nuclear phosphoprotein 32 family, member A (ANP32A), has multiple functions involved in neuritogenesis, transcriptional regulation, and apoptosis. However, whether ANP32A has an effect on the mammalian developing brain is still in question. In this study, it was shown that brain was the organ that expressed the most abundant ANP32A by human multiple tissue expression (MTE) array. The distribution of ANP32A in the different adult brain areas was diverse dramatically, with high expression in cerebellum, temporal lobe, and cerebral cortex and with low expression in pons, medulla oblongata, and spinal cord. The expression of ANP32A was higher in the adult brain than in the fetal brain of not only humans but also mice in a time-dependent manner. ANP32A signals were dispersed accordantly in embryonic mouse brain. However, ANP32A was abundant in the granular layer of the cerebellum and the cerebral cortex when the mice were growing up, as well as in the Purkinje cells of the cerebellum. The variation of expression levels and distribution of ANP32A in the developing brain would imply that ANP32A may play an important role in mammalian brain development, especially in the differentiation and function of neurons in the cerebellum and the cerebral cortex.

1. Introduction

ANP32A is a member of acidic nuclear phosphoprotein 32 kDa (ANP32) family [1, 2]. The nomenclature of ANP32A family members is confusing because the same protein has been given more than one name based on the context of isolation. The family members are comprised of ANP32A (also known as PP32, LANP, HPPCn, I1PP2A, MAPM, or PHAPIa), ANP32B (PAL31, APRIL, or PHAPIb) [3], ANP32C (PP32r1) [4], ANP32D (PP32r2) [5], and ANP32E (Cpd1, LANP-L, or PHAPIII) [6].

All ANP32 proteins share two highly conserved regions: the N-terminal leucine-rich repeats (LRRs) sequence and the C-terminal acidic tail [7]. For ANP32A, the hydrophobic LRRs shape a globular head domain. LRRs, belonging to a superfamily with diverse bioactivity, may potentially function in mediating protein-protein interaction [8]. The extended and hydrophilic C-terminal domain is highly unusual in its amino acid composition, containing abundant aspartic and glutamic acid residues, and a putative nuclear localization signal (NLS) [9]. According to the structure characteristic, it is not surprising to find ANP32 proteins involved in a variety of cellular processes in both nucleus and cytoplasm, including signaling, apoptosis, protein degradation, and morphogenesis.

More studies have focused on ANP32A, the founding member of the ANP32 family. ANP32A (PP32) was originally found as a tumor suppressor [10], associated with

cancer cell survival and drug efficacy. Meanwhile its closely related homologue PP32r1 is oncogenic and is overexpressed in breast cancer and prostate cancer [11]. Family member ANP32B was indicated as a potential prognostic marker of human breast cancer [2]. In fact, ANP32A has also been identified having a potential oncogenic and drug-resistant function in hepatocellular carcinoma [12], colorectal cancer [13], and pancreatic tumor [14]. The confused role in tumorigenesis of ANP32A may be due to its functions and subcellular localizations being almost bewildering in variety.

ANP32A is known to be a key component of the inhibitor of acetyltransferase (INHAT) complex in the nucleus, involved in regulating chromatin remodeling or transcription initiation [15]. ANP32A forms a multisubunit heterocomplex with HuR, regulating the nucleocytoplasmic shuttling of HuR, which is essential for RNA stability and transport [16]. In the cytoplasm, ANP32A (mapmodulin) is positioned to microtubule associated proteins (MAPs), involved in regulating microtubule function and microtubule-based vesicular trafficking [17]. ANP32A may control enzymatic activities by inhibition of protein phosphatase 2A (PP2A) or activation of caspases [18]. ANP32A (LANP) and SET were observed at the inner surface of the plasma membrane of lymphocytes and supposedly play a role in signal transduction [19]. ANP32A (HPPCn) was the first factor that it could transport to the extracellular space and act as an autocrine factor to promote DNA synthesis and suppress apoptosis by upregulating myeloid cell leukemia-1 [20]. These functions taken together suggest that ANP32A could therefore be an important regulator of cellular homeostasis.

ANP32A plays essential roles in a variety of neural pathophysiology processes. The level of ANP32A (I1PP2A) is increased in Alzheimer's disease (AD) and may be involved in regulatory mechanism of affecting Tau phosphorylation and impairing the microtubule network and neurite outgrowth [21]. ANP32A (LANP) regulates neuronal differentiation by epigenetic modulating expression of the neurofilament light chain, an important neuron-specific cytoskeletal gene [22]. ANP32A can interact with the retinoblastoma protein Rb in both young and mature neurons and is implicated in the regulation of neuronal survival by CXCL12/CXCR4 [23]. The decreased levels of ANP32A, as a potent and selective PP2A inhibitor, may contribute to abnormal neuritic morphology in a dominantly inherited neurodegenerative disorder of the spinocerebellar ataxia type 1 [24].

The expression characteristic of ANP32A in the developing brain, especially details on the expression and distributions of ANP32A in the human brain, had rarely been reported [25]. In this study, the distribution of ANP32A in different human brain areas, as well as ANP32A abundance in the human fetal and adult brain, was identified. For more details of the expression and localization of ANP32A in the developing brain, a series of different time point mouse brains from embryonic stage to adult stage was harvested and analyzed with ANP32A specific primers and antibodies. To explore ANP32A in the development of the nervous system, it could provide crucial information about pivotal roles of the protein in morphogenetic process and regulating mechanisms.

2. Materials and Methods

2.1. Animals. Adult C57 BL/6 mice were kept with free accesses to food and water. The day of insemination was designated as embryonic day 0 (E0). The day of birth was designated as postnatal day 0 (P0). Brains from different embryonic period (E12 and E16), early time points after birth (P0, P5, and P12), pubescent male mice (approximately 5-6 weeks old), and adult male mice (approximately 8–10 weeks old) were frozen quickly and stored at −80°C until required for experiments. E12 and E16 brains were collected under the dissection microscope and the mesenchymal tissues were removed with fine forceps as much as possible. For immunohistochemistry preparation, brains were fixed in formalin (4% formaldehyde in 1× PBS, pH7.4) for 24 hours and then embedded in paraffin. Serial sections (4 μm) were mounted onto silane-coated slides (Dako, Denmark).

2.2. Human Multiple Tissue Expression (MTE) Array Analysis. The human MTE array (BD, America) was a positively charged nylon membrane to which poly A^+ RNAs from different human tissues had been normalized and immobilized in separate dots, along with several controls. The MTE array made it possible to determine the relative expression levels of a target mRNA in different tissues and developmental stages [26]. To this end, poly A^+ RNA samples on each MTE array had been normalized to the mRNA expression levels of eight different "housekeeping" genes, which minimized the small tissue-specific variations in expression of any single housekeeping gene [27]. The human MTE array gave a relative convenient and convincing way to demonstrate the levels of ANP32A mRNA in 20 different areas of the human brain, as well as adult whole brain and fetal brain.

A 750 bp *ANP32A* probe was amplified from human ANP32A gene (Gene ID: 8125) in the pET-24a(+) plasmid (kept by our laboratory) with primers 5′-CGGGATCCATGGAGATGGGCAGACGGATT-3′ (forward) and 5′-AACTGCAGGTCAT-CATCTTCTCCCTCATC-3′ (reverse). Probe used for the human MTE assay was radioactively labeled with α-[^{32}P]dCTP using the DNA labeling kit (Promega, America) as described by the operation manual. Prehybridization of the MTE array was performed at 68°C for 2 hours in ExpressHyb hybridization (Clontech, America) with sheared salmon testis DNA added to a final concentration of 0.1 mg/mL. The denatured radiolabeled probe was mixed directly into the prehybridization solution and hybridized overnight at 68°C. After hybridization, the array was washed three times at 37°C in solution 2× SSC with 0.1% SDS for 10 min each time, repeated at 65°C in solution 0.1× SSC with 0.1% SDS for 20 min each time. Array was exposed to X-ray film at −70°C for 48 hours. ANP32A mRNA levels were determined by densitometric scanning of autoradiographs.

2.3. Western Blotting. Mouse whole brain tissues were homogenized in RIPA buffer (Sigma, America) on ice, lysates were centrifuged at 12,000 ×g, 4°C for 15 minutes, and supernatants were collected for western blot analysis. Protein

concentration of samples was determined by the Bio-Rad protein Assay (Bio-Rad, America). Protein extracts (20 μg) were fractionated by 12% SDS-PAGE and then transferred to PVDF membranes by the Trans-Blot semidry transfer cell (Bio-Rad, America). Membranes were subsequently blocked with 10% skim milk in 1× PBS and then incubated with a rabbit polyclonal antibody against ANP32A (1 μg/mL, ab 5991, dilution 1 : 1000) (Abcam, British) or β-actin (loading control, number 4970, dilution 1 : 1000) (CST, America). Following washing with PBST buffer, membranes were incubated with secondary antibody at appropriate dilution in 5 mL 10% blocking buffer. Protein bands on blots were detected by enhanced chemiluminescence (Applygen Technologies Inc., China) and visualized by LAS-4000 (GE, America).

2.4. Quantitative Real-Time RT-PCR. Total RNA was extracted from mouse whole brains using a Trizol (Invitrogen, America). 1 μg of total RNA was converted to cDNA using FastQuant RT kit (with gDNase) (TIANGEN BIOTECH, China). cDNA samples were then used as templates for quantitative real-time PCR by the SuperReal Premix Plus (SYBR Green) (TIANGEN BIOTECH, China) with a Bio-Rad Chromo 4 real-time PCR detector (Bio-Rad, CA, USA). The mouse ANP32A gene (157 base pairs [bp]) specific primers were 5′-CAGGGGACCTGGAAGTATTGG-3′ (forward) and 5′-TTCAGGTTGGTCACCTCACAG-3′ (reverse). Mouse β-actin (263 bp) primers were 5′-GAGACCTTCAACACCCCAGC-3′ (forward) and 5′-ATGTCACGCACGATTTCCC-3′ (reverse). Amplifications were carried out in 20 μL reaction mixture containing 20 ng cDNA samples, and the final concentration of 0.5 μmol/L of each primer pair was added in a program comprising 10 minutes at 95°C, followed by 40 cycles consisting of 95°C for 10 s, 55°C for 30 s, and 72°C for 30 s. Data was analyzed using the Bio-Rad CFX Manger. Statistical significance of differences was assessed by Student's t-test.

2.5. Immunohistochemistry. Slides were deparaffinized in xylene and rehydrated in different concentrations of ethanol (100%, 95%, 90%, 80%, and 70%), and antigen retrieval performed using a citrate buffer. Slides were blocked (37°C, 2.5 h) with 10% fetal bovine serum and then incubated with anti-ANP32A antibody (2 μg/mL) for 12 hours at 4°C. IgG-purified normal rabbit serum (2 μg/mL) (I5006, Sigma, SF, USA) was used as a control. After washing in phosphate buffered saline, sections were incubated with polyperoxidase-anti-rabbit IgG for 30 min at 37°C. Signals were visualized by DAB oxidation and observed by Ti-s microscope (Nikon, Japan).

3. Results

3.1. The Distribution of ANP32A in Different Areas of Human Central Nervous System. The differences of the levels of ANP32A mRNA between various areas of human brain were tested by the spot hybridized with the human MTE assay. The MTE array provides a fast way to simultaneously compare the relative abundance of ANP32A mRNA in a wide array of tissues, normalized and immobilized in separate dots, along with several controls. It was shown that brain was the organ that expressed the most abundant ANP32A, followed with heart, liver, and kidney. As shown with blue column in Figure 1, the expression levels of ANP32A mRNA were fluctuated in different areas of human brain. Among 20 different brain tissues, cerebellum right is the area with most abundant ANP32A; next is temporal lobe followed with cerebellum left, nucleus accumbens, substantia nigra, and cerebral cortex. Although ANP32A was abundant in the most brain areas, it was hardly detected in the pons. ANP32A could be slightly detected in medulla oblongata and spinal cord, a little higher than amygdala and parietal lobe.

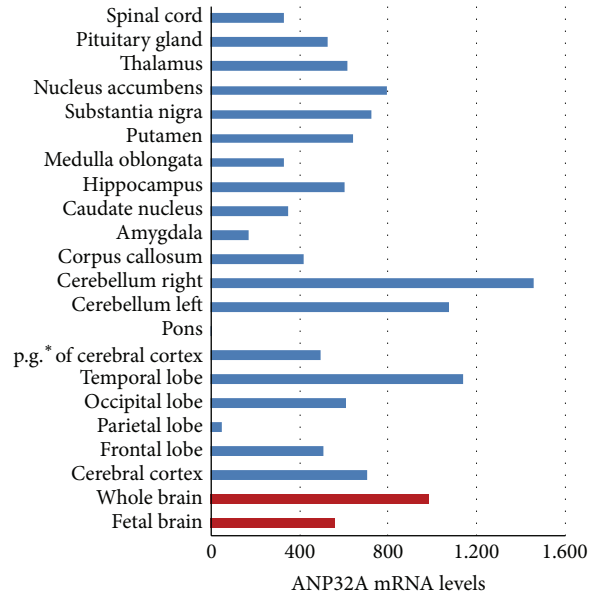


Figure 1: Distribution of human ANP32A transcripts in brain. Human MTE array was probed with a 750 bp human ANP32A radiolabeled probe as described under "experimental procedures." mRNA levels were determined by densitometric scanning of autoradiographs. Whole brain and fetal brain are shown as red column, and different anatomical region is shown as blue column.

3.2. ANP32A Was More Abundant in Adult Than Embryonic Brain of Both Human and Mouse. The difference between the levels of ANP32A mRNA in adult human brain and fetal human brain was also analyzed by MTE array. Brain in embryonic stage was still the organ with most abundant ANP32A, compared with 7 important human fetal organs including heart, liver, and kidney. The level of ANP32A mRNA in the adult whole brain was about 1.5-fold that in the fetal brain (as shown with red column in Figure 1). Then the expression of ANP32A gene in a different developmental stage of C57 BL/6 brain was studied. It was shown that the expression of ANP32A was higher in adult brain than that in embryonic brain of C57 BL/6 by both western blotting and qPCR. Although it is not too significant, the rising of the levels of ANP32A protein was confirmatory during the growing of the mice in a time-dependent manner (Figure 2(a)). As for the levels of ANP32A mRNA in mouse

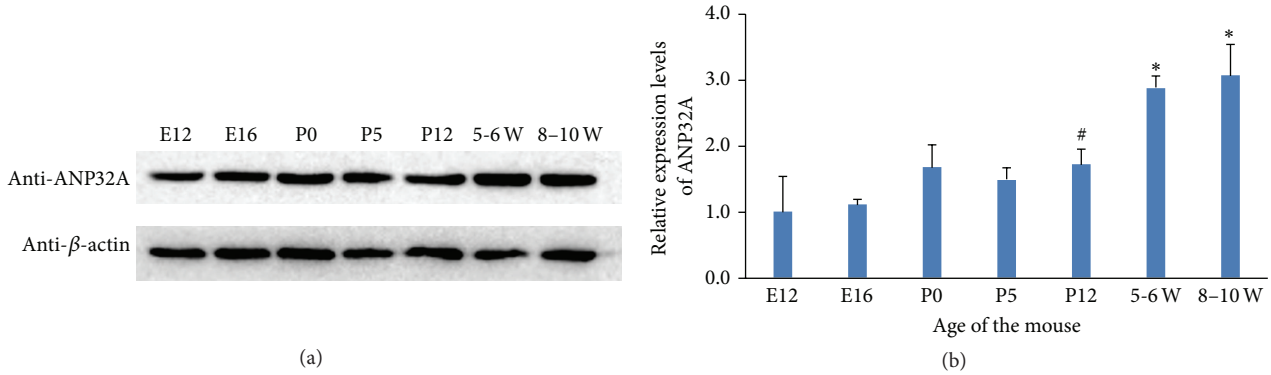


Figure 2: Expression of ANP32A gene in mouse developing brain. (a). Total proteins from the developing C57 BL/6 whole brains were analyzed by western blotting; (b). Total RNAs from the developing C57 BL/6 whole brains were analyzed by qPCR. E12 and E16, embryonic days 12 and 16, respectively; P0, P5, and P12, postnatal days 0, 5, and 12, respectively; 5-6 W and 8–10 W, adult brain from 5-6 weeks and 8–10 weeks old C57 BL/6. ∗: compared to E12 and E16, $P < 0.01$; #: compared to E12 and E16, $P < 0.05$.

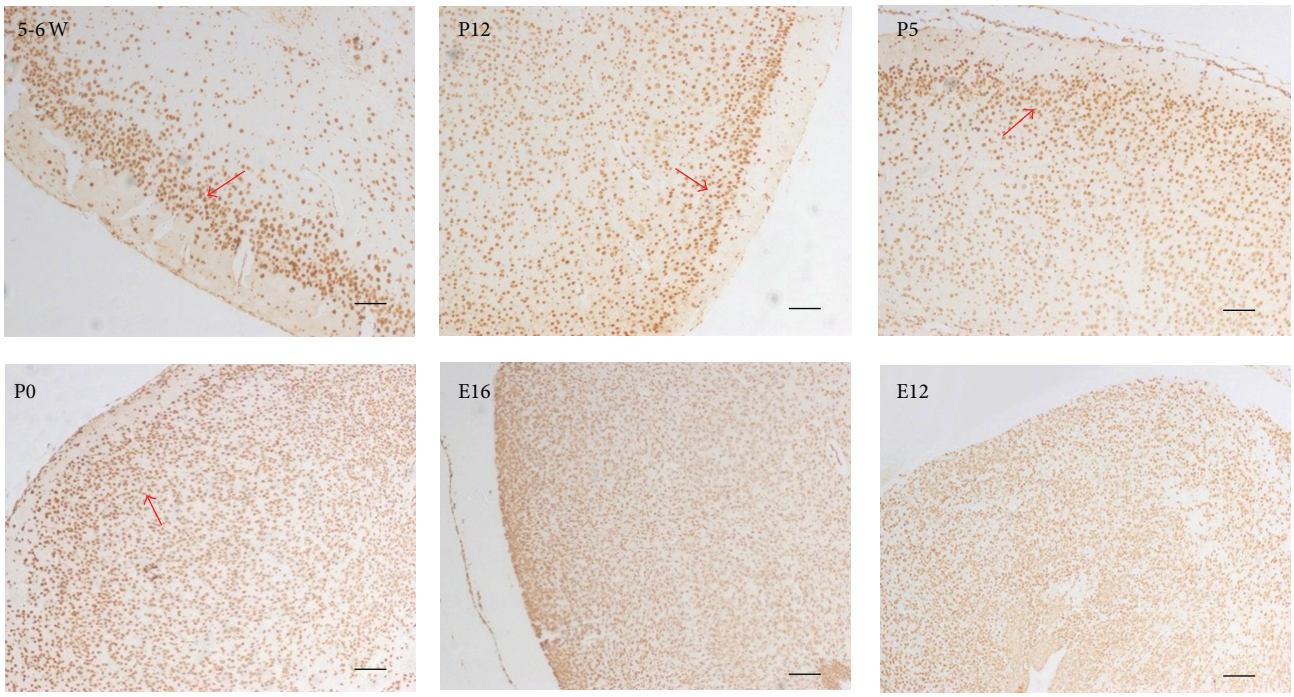


Figure 3: Immunohistochemical study of ANP32A in mouse cerebral cortex. ANP32A signal was visualized with DAB in C57 BL/6 brain. E12 and E16, embryonic days 12 and 16, respectively; P0, P5, and P12, postnatal days 0, 5, and 12, respectively; 5-6 W, adult brain from 5-6 weeks old C57 BL/6. Scale bar = 200 μm. Granule cells were marked by red arrow.

brain analyzed by qPCR, it tended to increase after birth and was apparently raised in postnatal day 12 (P12), about 1.5-fold compared to embryonic day 12 and 16 (E12 and E16). When the mice were 5-6 weeks to 8–10 weeks old, the levels of ANP32A mRNA in the brain were increased significantly, reaching about 3-fold than that in E12 and E16 (Figure 2(b)). There was no significant difference in the interior-group.

3.3. The Expression Characteristics of ANP32A in the Developing Mouse Cerebral Cortex. Compared with the embryonic mouse, ANP32A protein was expressed along with the differential cerebral cortex's layer, more and more significantly, after birth. The changes of morphology of the positive stained cells were also coincidental to the development of the neuron in the nervous system. In the embryonic period, the layer of cerebral cortex is not clear. All positive staining cells are small and numerous, and the expression of ANP32A seemed to be nothing special. When the mice were birthed and growing, the positive staining cells were bigger and more strongly stained in external granular layer of the cerebral cortex (Figure 3).

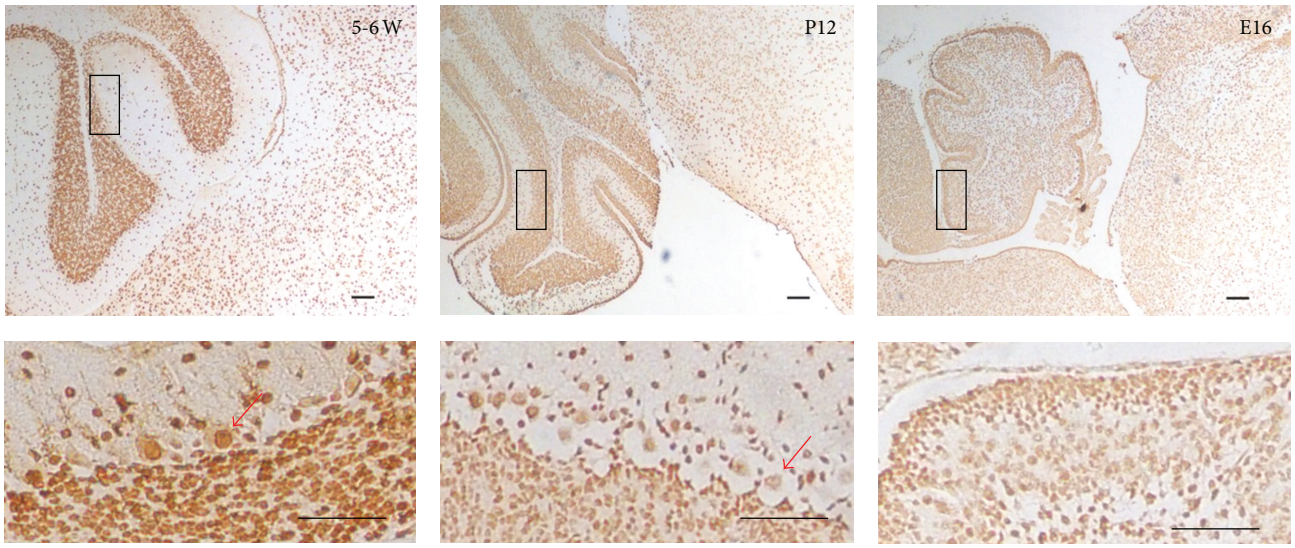


Figure 4: Immunohistochemical study of ANP32A in mouse cerebellum. ANP32A signal was visualized with DAB in C57 BL/6 brain. E16, embryonic day 16; P12, postnatal day 12; 5-6 W, adult brain from 5-6 weeks old C57 BL/6. In each image, the upper rectangle inset is magnified and displayed in the corresponding lower image. Purkinje cells were marked by red arrow. Scale size = 100 μm.

In detail, from the day of birth (P0), the molecular layer of cerebral cortex appeared, and the staining of the ANP32A in the molecular layer was initially decreased than other partitions of cerebral cortex. On the P5 and P12, the positive stain cells were stratification and conversely in the molecular layer and external granular layer. The difference of staining between the molecular layer and external granular layer in 5-6 weeks old mouse brain was much more significant. The expression of ANP32A was fairly abundant in the external granular layer, localized in the nucleus of the neurons, while being apparently lower in molecular layer.

3.4. The Cellular Localization of ANP32A in the Granule Cells and the Purkinje Cells in the Developing Mouse Cerebellum. The distribution of ANP32A changed with the migration of external granule cells in the developing mouse cerebellum. In the embryonic period, the ANP32A was expressed moderately in the nucleus of the internal granule cells and strongly in the external granule cells. In the postnatal day 12, the signals in internal granular layer became stronger, while in the 5-6 W mouse brain the ANP32A's expression in the internal granule cells became much stronger which may be attributed in large part to migration of the external granule cells to internal granular layer.

The amount and cellular localization of the expression of ANP32A in Purkinje cells seemed to be associated with the mice age and the position in the adult mouse gyrus cerebelli. In the cerebellum P12, ANP32A was expressed moderately in only the nucleus of the Purkinje cells, which scattered evenly between the molecular layer and the granular layer. In the cerebellum 5-6W, it was absorbing that ANP32A was localized in both the nucleus and the cytoplasm, as well as dendrites arborization of Purkinje cells (Figure 4). The signals became weaker in the Purkinje cells in the root of the gyrus cerebelli. On the other hand, the signals of ANP32A became stronger in the Purkinje cells in the head of the gyrus cerebelli. It is implied that ANP32A may associate with the differentiation and function of Purkinje cells.

4. Discussion

ANP32 family members had been thought all functionally redundant in vivo. Because loss-of-function mutants for ANP32 family members include two independently targeted ANP32A-deficient mice [2, 28], an ANP32E-deficient mouse [2, 29] was viable and fertile. No obvious abnormalities in any of the major organ systems, including the nervous system, could be observed. However, a recent prominent finding is that mice carrying ANP32B mutations are sensitized to loss of ANP32A. The study revealed previously hidden roles for ANP32A in mouse development by compound mutants lacking ANP32A, ANP32B, and/or ANP32E [2]. Since ANP32 family members are not completely redundant in mammals, it is reasonable to presume that they may engage in regulating mechanism in the mammal development in a hierarchical and successive manner.

In view of a broad array of physiological activities [15–20] and roles in nervous system disease [21–24] of ANP32A, the description of the expression and distributions characteristics of ANP32A in the developing brain would provide strong clues for the regulating mechanism of ANP32A in the nervous system development and disease. So in this study, the distribution of ANP32A in different areas of human central nervous system, as well as the expression levels in the fetal and adult brain, was analyzed by a human multiple tissue expression array. The cellular localization and expression levels fluctuation of ANP32A were detected in the developing mouse brain. Our results indicated that ANP32A may play an important role in human nervous system development and differentiation.

MTE array showed that the expression of ANP32A was higher in the human adult brain than the fetal brain. The similar evidences have been collected by the analysis of ANP32A abundance in a series of different time point mouse brain from embryonic stage to adult stage. Both ANP32A mRNA and proteins were elevated in a time-dependent manner in the developing mouse brain. Mutai et al. had reported that expression of PAL31/ANP32B mRNA and protein in the rat brain was high during the fetal period and decreased after birth [30]. Collecting these evidences, it implied that ANP32A and ANP32B are not completely functionally redundant along with observation of compound mutants lacking ANP32A and/or ANP32B [2]. ANP32B may mainly function in fetal period of mammal, while ANP32A may primarily play roles in adult brain.

In the embryonic mouse brain, all positive staining cells were distributed homogeneously, small and numerous. The expression of ANP32A seemed to be nothing special except that the staining in the external granule cells of cerebellum was a little stronger. However, in postnatal day 12, the signals in internal granular layer became stronger. ANP32A was significantly abundant in the granular layer of the cerebellum, and the cerebral cortex when the mice were 5-6 weeks old, as well as in the Purkinje cells of the cerebellum. It may be attributed in large part of migration of the external granule cells to internal granular layer [31].

The amount and cellular localization of the expression of ANP32A in Purkinje cells seemed to be associated with the mice age and the position. In the cerebellum P12, ANP32A was expressed moderately in the nucleus of the Purkinje cells, similar to some ANP32 proteins [25, 28], which scattered evenly between the molecular layer and the granular layer. However, in the cerebellum 5-6 W, it was observed that ANP32A was localized in both the nucleus and the cytoplasm, as well as dendrites arborization of Purkinje cells. And the abundance of ANP32A in the Purkinje cells was varied according to the site of gyrus cerebella, more strongly stained in the head and less stained in the root of the adult mouse gyrus cerebella by immunohistochemistry. It indicated that ANP32A may associate with the differentiation and function of Purkinje cells [32–34].

The distribution of ANP32A in the different adult brain areas was dramatically diverse. Strongly stained nuclei were observed in the external granular layer of cerebral cortex and the granule cells in the cerebellum. MTE array showed that ANP32A was abundant in the human nervous system with high expression in cerebellum, temporal lobe, nucleus accumbens, substantia nigra, and cerebral cortex.

The cerebellum is a region of the brain that plays an important role in motor control. It may also be involved in some cognitive functions such as attention and language, and in regulating fear and pleasure responses. Its movement-related functions are the most solidly established. Learning how to ride a bicycle is an example of a type of neural plasticity that may take place largely within the cerebellum [31]. The cerebral cortex plays a key role in memory, attention, perceptual awareness, thought, language, and consciousness. The temporal lobe is one of the four major lobes of the cerebral cortex in the brain of mammals. It is involved in the retention of visual memories, processing sensory input, and comprehending language [35]. Research has indicated the nucleus accumbens has an important role in pleasure including laughter, reward, and reinforcement learning [36]. The substantia nigra plays an important role in reward, addiction, and movement. Altogether, it implied that ANP32A may have important functions in neural plasticity for essential acquired ability and participate in advancing nervous activity, such as language, emotion, learning, and memory.

On the other hand, ANP32A was lowly expressed in pons, medulla oblongata, and spinal cord. These areas of nervous system function primarily in vital activity, such as control sleep, respiration, swallowing, and motor organization. ANP32A should have tiny effects on these functional areas. In this context, it is not surprising that loss-of-function mutants for ANP32A are not fatal.

In conclusion, ANP32A was abundant in the central nervous system. The expression of ANP32A in the developing brain was raised in a time-dependent manner. And the distribution of ANP32A changed dramatically in different brain areas and layer of cerebellum or cerebral cortex, which implied the roles of ANP32A involved in differentiation and specific functional regulation of neurons. Potential mechanisms of ANP32A in the development and differentiation of nervous system may be involved in neuritogenesis modulating, apoptosis regulating, and transcription control, according to the protein localization in and out of the neurons [22, 37, 38].

Ethical Approval

All experiments were reviewed by the Ethics Committee of the Navy General Hospital of Chinese PLA.

Conflict of Interests

There is no conflict of interests for any authors.

Authors' Contribution

Shanshan Wang and Yunliang Wang contributed equally to this study.

Acknowledgments

This study was supported in part by the National Natural Science Foundation of China (nos. 31071256, 81272700, and 81472350) and the Innovation Fund of the Navy General Hospital of Chinese PLA (no. CX200904).

References

[1] A. Matilla and M. Radrizzani, "The Anp32 family of proteins containing leucine-rich repeats," *Cerebellum*, vol. 4, no. 1, pp. 7–18, 2005.

[2] P. T. Reilly, S. Afzal, C. Gorrini et al., "Acidic nuclear phosphoprotein 32kDa (ANP32)B-deficient mouse reveals a hierarchy of ANP32 importance in mammalian development," *Proceedings*

of the National Academy of Sciences of the United States of America, vol. 108, no. 25, pp. 10243–10248, 2011.

[3] L.-F. Shen, H. Cheng, M.-C. Tsai, H.-S. Kuo, and K.-F. Chak, "PAL31 may play an important role as inflammatory modulator in the repair process of the spinal cord injury rat," *Journal of Neurochemistry*, vol. 108, no. 5, pp. 1187–1197, 2009.

[4] K. Imamachi, F. Higashino, T. Kitamura et al., "pp32r1 controls the decay of the RNA-binding protein HuR," *Oncology Reports*, vol. 31, no. 3, pp. 1103–1108, 2014.

[5] S. S. Kadkol, G. A. E. Naga, J. R. Brody et al., "Expression of pp32 gene family members in breast cancer," *Breast Cancer Research and Treatment*, vol. 68, no. 1, pp. 65–73, 2001.

[6] T. A. Santa-Coloma, "Anp32e (Cpd1) and related protein phosphatase 2 inhibitors," *Cerebellum*, vol. 2, no. 4, pp. 310–320, 2003.

[7] T. Huyton and C. Wolberger, "The crystal structure of the tumor suppressor protein pp32 (Anp32a): structural insights into Anp32 family of proteins," *Protein Science*, vol. 16, no. 7, pp. 1308–1315, 2007.

[8] B. Kobe and J. Deisenhofer, "A structural basis of the interactions between leucine-rich repeats and protein ligands," *Nature*, vol. 374, no. 6518, pp. 183–186, 1995.

[9] S.-B. Seo, T. Macfarlan, P. McNamara et al., "Regulation of histone acetylation and transcription by nuclear protein pp32, a subunit of the INHAT complex," *The Journal of Biological Chemistry*, vol. 277, no. 16, pp. 14005–14010, 2002.

[10] S. Hoffarth, A. Zitzer, R. Wiewrodt et al., "pp32/PHAPI determines the apoptosis response of non-small-cell lung cancer," *Cell Death & Differentiation*, vol. 15, no. 1, pp. 161–170, 2008.

[11] S. Buddaseth, W. Göttmann, R. Blasczyk, T. Huyton, and W. Göttmann, "Overexpression of the pp32r1 (ANP32C) oncogene or its functional mutant pp32r1Y140H confers enhanced resistance to FTY720 (Finguimod)," *Cancer Biology & Therapy*, vol. 15, no. 3, pp. 289–296, 2014.

[12] C. Li, H.-Q. Ruan, Y.-S. Liu et al., "Quantitative proteomics reveal up-regulated protein expression of the SET complex associated with hepatocellular carcinoma," *Journal of Proteome Research*, vol. 11, no. 2, pp. 871–885, 2012.

[13] H. Shi, K. A. Hood, M. T. Hayes, and R. S. Stubbs, "Proteomic analysis of advanced colorectal cancer by laser capture microdissection and two-dimensional difference gel electrophoresis," *Journal of Proteomics*, vol. 75, no. 2, pp. 339–351, 2011.

[14] T. K. Williams, C. L. Costantino, N. A. Bildzukewicz et al., "pp32 (ANP32A) expression inhibits pancreatic cancer cell growth and induces gemcitabine resistance by disrupting HuR binding to mRNAs," *PLoS ONE*, vol. 5, no. 11, Article ID e15455, 2010.

[15] S. Kadota and K. Nagata, "pp32, an INHAT component, is a transcription machinery recruiter for maximal induction of IFN-stimulated genes," *Journal of Cell Science*, vol. 124, no. 6, pp. 892–899, 2011.

[16] T. Kuroshima, M. Aoyagi, M. Yasuda et al., "Viral-mediated stabilization of AU-rich element containing mRNA contributes to cell transformation," *Oncogene*, vol. 30, no. 26, pp. 2912–2920, 2011.

[17] P. Opal, J. J. Garcia, F. Propst, A. Matilla, H. T. Orr, and H. Y. Zoghbi, "Mapmodulin/leucine-rich acidic nuclear protein binds the light chain of microtubule-associated protein 1B and modulates neuritogenesis," *The Journal of Biological Chemistry*, vol. 278, no. 36, pp. 34691–34699, 2003.

[18] C. Habrukowich, D. K. Han, A. Le et al., "Sphingosine interaction with acidic leucine-rich nuclear phosphoprotein-32A (ANP32A) regulates PP2A activity and cyclooxygenase (COX)-2 expression in human endothelial cells," *The Journal of Biological Chemistry*, vol. 285, no. 35, pp. 26825–26831, 2010.

[19] M. Vaesen, S. Barnikol-Watanabe, H. Gotz et al., "Purification and characterization of two putative HLA class II associated proteins: PHAPI and PHAPII," *Biological Chemistry Hoppe-Seyler*, vol. 375, no. 2, pp. 113–126, 1994.

[20] J. Chang, Y. Liu, D.-D. Zhang et al., "Hepatopoietin Cn suppresses apoptosis of human hepatocellular carcinoma cells by up-regulating myeloid cell leukemia-1," *World Journal of Gastroenterology*, vol. 16, no. 2, pp. 193–200, 2010.

[21] S. Chen, B. Li, I. Grundke-Iqbal, and K. Iqbal, "I1PP2A affects Tau phosphorylation via association with the catalytic subunit of protein phosphatase 2A," *Journal of Biological Chemistry*, vol. 283, no. 16, pp. 10513–10521, 2008.

[22] R. K. Kular, M. Cvetanovic, S. Siferd, A. R. Kini, and P. Opal, "Neuronal differentiation is regulated by leucine-rich acidic nuclear protein (LANP), a member of the inhibitor of histone acetyltransferase complex," *The Journal of Biological Chemistry*, vol. 284, no. 12, pp. 7783–7792, 2009.

[23] M. Z. Khan, A. Vaidya, and O. Meucci, "CXCL12-mediated regulation of ANP32A/Lanp, a component of the inhibitor of histone acetyl transferase (INHAT) complex, in cortical neurons," *Journal of Neuroimmune Pharmacology*, vol. 6, no. 1, pp. 163–170, 2011.

[24] I. Sánchez, P. Pinol, M. Corral-Juan, M. Pandolfo, and A. Matilla-Dueñas, "A novel function of ataxin-1 in the modulation of PP2A activity is dysregulated in the spinocerebellar ataxia type 1," *Human Molecular Genetics*, vol. 22, no. 17, pp. 3425–3437, 2013.

[25] K. Matsuoka, M. Taoka, N. Satozawa et al., "A nuclear factor containing the leucine-rich repeats expressed in murine cerebellar neurons," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 91, no. 21, pp. 9670–9674, 1994.

[26] E. Spanakis and D. Brouty-Boyè, "Evaluation of quantitative variation in gene expression," *Nucleic Acids Research*, vol. 22, no. 5, pp. 799–806, 1994.

[27] E. Spanakis, "Problems related to the interpretation of autoradiographic data on gene expression using common constitutive transcripts as controls," *Nucleic Acids Research*, vol. 21, no. 16, pp. 3809–3819, 1993.

[28] P. Opal, J. J. Garcia, A. E. McCall et al., "Generation and characterization of LANP/pp32 null mice," *Molecular and Cellular Biology*, vol. 24, no. 8, pp. 3140–3149, 2004.

[29] P. T. Reilly, S. Afzal, A. Wakeham et al., "Generation and characterization of the Anp32e-deficient mouse," *PLoS ONE*, vol. 5, no. 10, Article ID e13597, 2010.

[30] H. Mutai, Y. Toyoshima, W. Sun, N. Hattori, S. Tanaka, and K. Shiota, "PAL31, a novel nuclear protein, expressed in the developing brain," *Biochemical and Biophysical Research Communications*, vol. 274, no. 2, pp. 427–433, 2000.

[31] E. M. Hamilton, E. Polder, A. Vanderver et al., "Hypomyelination with atrophy of the basal ganglia and cerebellum: further delineation of the phenotype and genotype-phenotype correlation," *Brain*, vol. 137, part 6, pp. 1921–1930, 2014.

[32] J. T. Fleming, W. He, C. Hao et al., "The Purkinje neuron acts as a central regulator of spatially and functionally distinct cerebellar precursors," *Developmental Cell*, vol. 27, no. 3, pp. 278–292, 2013.

[33] R. Ohashi, S. Sakata, A. Naito, N. Hirashima, and M. Tanaka, "Dendritic differentiation of cerebellar Purkinje cells is promoted by ryanodine receptors expressed by Purkinje and

granule cells," *Developmental Neurobiology*, vol. 74, no. 4, pp. 467–680, 2014.

[34] T. Nakatani, Y. Minaki, M. Kumai, C. Nitta, and Y. Ono, "The c-Ski family member and transcriptional regulator Corl2/Skor2 promotes early differentiation of cerebellar Purkinje cells," *Developmental Biology*, vol. 388, no. 1, pp. 68–80, 2014.

[35] G. Rees, "Neural correlates of consciousness," *Annals of the New York Academy of Sciences*, vol. 1296, no. 1, pp. 4–10, 2013.

[36] A. Rinaldi, A. Oliverio, and A. Mele, "Spatial memory, plasticity and nucleus accumbens," *Reviews in the Neurosciences*, vol. 23, no. 5-6, pp. 527–541, 2012.

[37] C. S. Hunter, R. E. Malik, F. A. Witzmann, and S. J. Rhodes, "LHX3 interacts with inhibitor of histone acetyltransferase complex subunits LANP and TAF-1β to modulate pituitary gene regulation," *PLoS ONE*, vol. 8, no. 7, Article ID e68898, 2013.

[38] W. Pan, L. S. de Graca, Y. Shao, Q. Yin, H. Wu, and X. Jiang, "PHAPI/pp32 suppresses tumorigenesis by stimulating Apoptosis," *The Journal of Biological Chemistry*, vol. 284, no. 11, pp. 6946–6954, 2009.

Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 124721, 8 pages
http://dx.doi.org/10.1155/2015/124721

Research Article

Protecting Intestinal Epithelial Cell Number 6 against Fission Neutron Irradiation through NF-κB Signaling Pathway

Gong-Min Chang,[1,2] Ya-Bing Gao,[1] Shui-Ming Wang,[1] Xin-Ping Xu,[1] Li Zhao,[1] Jing Zhang,[1] Jin-Feng Li,[3] Yun-Liang Wang,[3] and Rui-Yun Peng[1]

[1]*Department of Experimental Pathology, Beijing Institute of Radiation Medicine, 27 Taiping Road, Haidian District 100850, China*
[2]*Department of Medical Oncology, Air Force PLA General Hospital, Fucheng Road No. 30, Haidian District, Beijing 100142, China*
[3]*The Neurology Department of the 148th Hospital, 20 Zhanbei Road, Zibo 255300, China*

Correspondence should be addressed to Yun-Liang Wang; wangyunliang81@163.com and Rui-Yun Peng; pengry@nic.bmi.ac.cn

Received 28 June 2014; Accepted 26 July 2014

Academic Editor: Hongwei Wang

The purpose of this paper is to explore the change of NF-κB signaling pathway in intestinal epithelial cell induced by fission neutron irradiation and the influence of the PI3K/Akt pathway inhibitor LY294002. Three groups of IEC-6 cell lines were given: control group, neutron irradiation of 4Gy group, and neutron irradiation of 4Gy with LY294002 treatment group. Except the control group, the other groups were irradiated by neutron of 4Gy. LY294002 was given before 24 hours of neutron irradiation. At 6 h and 24 h after neutron irradiation, the morphologic changes, proliferation ability, apoptosis, and necrosis rates of the IEC-6 cell lines were assayed and the changes of NF-κB and PI3K/Akt pathway were detected. At 6 h and 24 h after neutron irradiation of 4Gy, the proliferation ability of the IEC-6 cells decreased and lots of apoptotic and necrotic cells were found. The injuries in LY294002 treatment and neutron irradiation group were more serious than those in control and neutron irradiation groups. The results suggest that IEC-6 cells were obviously damaged and induced serious apoptosis and necrosis by neutron irradiation of 4Gy; the NF-κB signaling pathway in IEC-6 was activated by neutron irradiation which could protect IEC-6 against injury by neutron irradiation; LY294002 could inhibit the activity of IEC-6 cells.

1. Introduction

As we all know, the pathological change of intestine induced by neutron irradiation is on the whole elucidated, while the mechanisms of injury were not elucidated completely. The transcription factor nuclear factor-kappa B (NF-κB) plays a pivotal role in the cellular response to various kinds of stress situations. Exposure to extracellular stimuli, such as microbial products and proinflammatory cytokines, as well as internally initiated stress signals derived from reactive oxygen species, hypoxia, or endoplasmatic reticulum stress, initiates signaling pathways that activate the gene expression-inducing capacity of NF-κB [1]. On the one hand, NF-κB plays an important role in inflammatory reaction; on the other hand, NF-κB can protect intestinal epithelial cell against damage [2]. NF-κB is composed of p65 and p50 subunit. NF-κB combines together with the inhibitor of kappa B (IκB) family in the cytoplasm which is not activity. Many extracellular stimulation can cause a series of cascade reactions of intracell. IκB kinase (IKK) family proteins activate IκB family proteins segregating NF-κB and IκB family proteins; then, NF-κB is activated and controls many genes transcription [3, 4]. LY294002 is the classic inhibitor of phosphatidylinositol-3-kinase (PI3K). PI3K and Akt can induce NF-κB activation. LY294002 inhibits the activation of PI3K/Akt which can inhibit the activation of NF-κB at the same time [5]. We established intestinal epithelial cell model injured by neutron irradiation of 4Gy, to study the

protection of NF-κB signaling pathway on intestinal epithelial cell injured by neutron irradiation and explore how PI3K/Akt regulate NF-κB signaling pathway. Our studies might provide important theoretical and practical evidence about intestine injured by neutron irradiation.

2. Materials and Methods

2.1. Cell Culture and Reagents. Intestinal epithelial cell number 6 (IEC-6) cell lines (origin of SD rat) were kindly provided by Professor Qingliang Luo. IEC-6 cells were inoculated in Dulbecco's modified eagle medium (DMEM) (Sigma-Aldrich Company, New Jersey, USA). The media were supplemented with 10% fetal bovine serum (FBS) (Yuan Heng Sheng Ma Biology Technology Research Institute, Beijing, China). The culture solutions were replaced every two days and IEC-6 cells were subcultured by trypsinization every three days. Three groups were randomly given: control group, neutron irradiation group, and LY294002 (Cell Signaling Technology Company, Beverly, MA, USA) treatment group. LY294002 was purchased from Cell Signaling Technology and 40 millimolar (mM) stock solutions of LY294002 in dimethyl sulfoxide (DMSO) (Beijing Chemical Reagent Factory, China) were stored at −20°C. LY294002 was added into the culture solutions 24 hours before neutron irradiation (final LY294002 concentration was 10 μM). The control groups received isovolumic DMSO. Annexin V fluorescein isothiocyanate (Annexin V-FITC) kit was from Beijing Bao Sai Biotech of China. Anti-NF-κB and anti-phosphor-NF-κB, anti-IKKα/β and anti-phosphor-IKKα/β, and anti-IκBα and anti-phosphor-IκBα antibodies were purchased from Cell Signaling Technology Company of America (NF-κB Pathway Sampler Kit, including primary antibodies and secondary antibodies, all antibodies were stored at–20°C). Anti-PI3K and anti-phosphor-PI3K, anti-Akt and anti-phosphor-Akt antibodies were also purchased from Cell Signaling Technology Company of America. All primary antibodies were diluted to 1 : 1000 in 0.1% Tween in *tris*-buffered saline (TBS-T). Secondary antibodies were diluted to 1 : 5000 in 0.1% TBS-T. Anti-glyceraldehyde-3-phosphate dehydrogenase (GAPDH) was gotten from KangChen Biotechnology of China (stored at 4°C). Anti-GAPDH was diluted to 1 : 10 000 in 0.1% TBS-T.

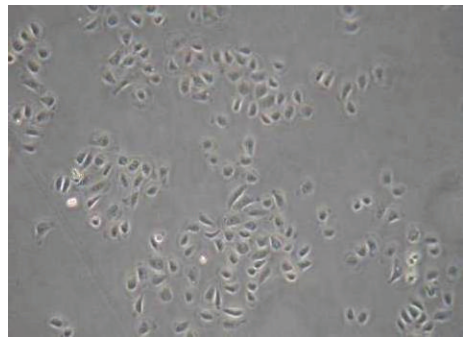
2.2. Neutron Irradiation. Fission neutron source was provided by Nuclear Energy Technology Design Academy of Tsinghua University, Beijing, China. The power of the reactor is 50 kilowatt (kW). The rate of neutron and γ ray is 9 : 1 (neutron occupies 90%). The average energy of neutron was 1.33 MeV. The dose rate of neutron was 39.04cGy/min and its absorbed dose was 4Gy. The IEC-6 cell culture bottles were fixed on a plastic disc. When the IEC-6 cell culture bottles were radiated by neutron, the plastic disc was rotating (rotation speed of 10 cycles per minute) in order to assure the IEC-6 cells received even neutron irradiation.

2.3. Inverted Phase Contrast Microscope (IPCM) Assay. The morphologic changes of the IEC-6 cells were observed by IPCM (Olympus, Japan) at 6 h and 24 h after neutron irradiation.

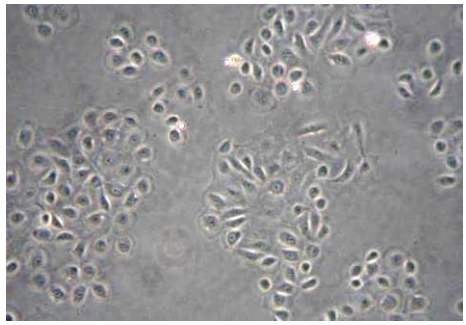
2.4. 3-(4,5-Dimethylthiazol-2-yl)-2,5-diphenyltetrazolium Bromide (MTT) Assay. IEC-6 cells were inoculated in 96-pore plate (cell density $3 \sim 5 \times 10^4$/mL, 200 μL/pore). The IEC-6 cells were assayed by MTT (Gibco, USA) colorimetry at 6 h and 24 h after neutron irradiation of 4Gy. Experiment procedure: (1) add 20 μL MTT solution into each pore of the 96-pore plate; (2) put the 96-pore plate into 37°C attemperator for 4 h; (3) suck and discard the supernatant and add 200 μL DMSO into each pore of the 96-pore plate; (4) fix the 96-pore plate onto a shaker and shake it thoroughly so that the crystallizations were dissolved completely; (5) assay the optical density (OD) value of each pore by means of an enzyme linked immunosorbent assay detector (wavelength of 570 nm).

2.5. Flow Cytometry (FCM) Assay. The apoptosis and necrosis rates of the IEC-6 cells were assayed at 6 h and 24 h after neutron irradiation of 4Gy. Experiment procedure according to Annexin V-FITC kit (Bao Sai Biology Technology Company, Beijing, China) instruction: (1) trypsinize the IEC-6 cells by 0.25% trypsin and wash them at 4°C 0.1% phosphate buffered solution (PBS) by centrifuge; (2) modulate the IEC-6 cells concentration to $5 \times 10^5 \sim 1 \times 10^6$/mL and wash them at 4°C 0.1% PBS by centrifuge again; (3) suspend the IEC-6 cells in 200 μL buffer solution; (4) add 10 mL Annexin V-FITC and 5 mL propidium iodide (PI) into the buffer solution; (5) admix above-mentioned solution uniformly and leave the solution to react at the room or 4°C temperature in the dark; (6) add 300 μL binding buffer into above-mentioned solution and then assay the apoptosis and necrosis rates of the IEC-6 cells by FCM (B-D, America).

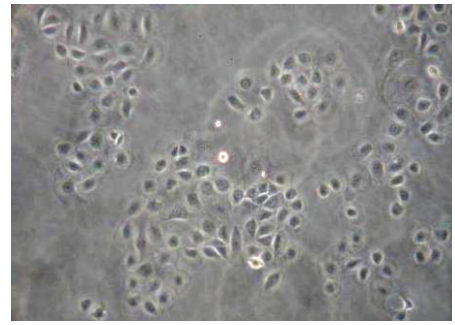
2.6. Western Blotting Assay. IEC-6 cells were collected and the total proteins were extracted at 6 h and 24 h after neutron irradiation of 4Gy. Proteins were extracted using a Whole Cell Extraction Kit (EMD Millipore Corporation, Billerica, MA, USA), according to the instructions. The proteins were separated in the sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) and transferred to the polyvinylidene difluoride (PVDF) membrane (EMD Millipore Corporation, Billerica, MA, USA). We judged the interest protein straps according to the protein marker. The PVDF membranes were blocked in 5% nonfat milk (diluted in 0.1% TBS-T) for 2 h at room temperature. The PVDF membranes were probed with primary antibodies on the rocking bed overnight at 4°C, washed in 0.1% TBS-T three times of 10 minutes, and incubated with appropriate secondary antibodies on the rocking bed for 1 h at room temperature. Antibody binding was detected using enhanced chemiluminescence (ECL) Pro-Light horseradish peroxidase (HRP) kit (Tiangen Biotechnology Company, Beijing, China) and photos were taken by means of FluorChem FC2 imaging system (Nature Gene, America). Signals were quantified using CMIAS-II image analysis system (Beijing University of Aeronautics & Astronautics, Beijing,



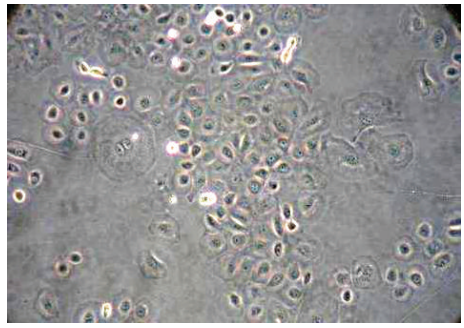
(a) Control



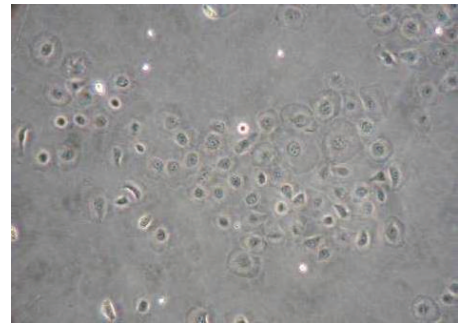
(b) Six h after neutron radiation of 4Gy



(c) Twenty-four h after neutron radiation of 4Gy



(d) Six h after neutron radiation of 4Gy with LY294002 treatment



(e) Twenty-four h after neutron radiation of 4Gy with LY294002 treatment

Figure 1: The changes of IEC-6 appearance (IPCM, magnification of 200x). (a) Control, the IEC-6 cells looked like applanatus polygon or fusiform shape and aggregated together shape of chrysanthemum thyse appearance. (b) Six h after neutron radiation of 4Gy, the IEC-6 cells swelled and became approximately round shaped and lots of dead cells floated on the culture solution. (c) Twenty-four h after neutron radiation of 4Gy, the cells were injured more seriously than those of 6 h after neutron irradiation. ((d), (e)) Neutron irradiation of 4Gy with LY294002 treatment group cells, the cells were injured more seriously than those of the neutron irradiation group.

China) and compared with the integral optical density (IOD) values.

2.7. Statistical Analysis. The data were analyzed by one way ANOVA using of the statistical package for the social sciences (SPSS) 13.0 statistical software. The data were described using mean and standard deviation ($\overline{X} \pm s$). Comparing with control group, $*$ was $P < 0.05$, $**$ was $P < 0.01$; comparing with neutron irradiation group, # was $P < 0.05$, ## was $P < 0.01$. Values of $P < 0.05$ were considered statistically significant.

3. Results

3.1. Morphologic Changes of IEC-6. The control group IEC-6 cells grew side by side tightly adhering to the culture flask. The IEC-6 cells aggregated together and looked like cluster or chrysanthemum thyse appearance. The IEC-6 cells looked like applanate polygon or fusiform shape (Figure 1(a)). At 6 h and 24 h after neutron irradiation of 4Gy, the IEC-6 cells swelled and became approximately round shaped and lots of dead cells floated on the culture solution (Figures 1(b) and 1(c)). The LY294002 treatment group IEC-6 cells were injured more seriously than neutron irradiation group (Figures 1(d) and 1(e)).

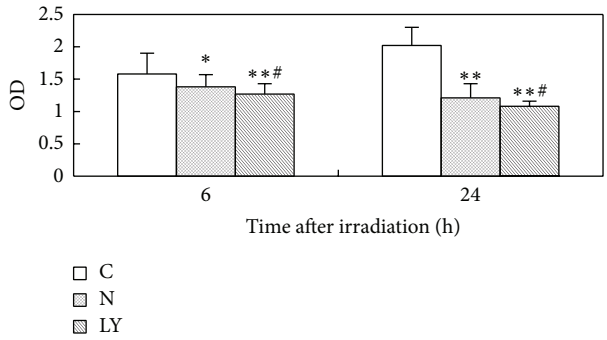


Figure 2: The proliferation ability changes of IEC-6 cells after neutron irradiation of 4Gy and using of LY294002 (C: control group; N: neutron irradiation group; LY: neutron irradiation of 4Gy with LY294002 treatment group).

3.2. Proliferation Ability of IEC-6. The proliferation ability of the control group IEC-6 cells was increased gradually in 24 h. The proliferation ability of the neutron irradiation group IEC-6 cells was decreased obviously in 24 h of neutron irradiation of 4Gy. The descent tendency of the LY294002 treatment group cells was more obvious than that of the neutron irradiation group. Figure 2 showed the results of statistical analysis.

3.3. Apoptosis and Necrosis Rates of IEC-6. The apoptosis and necrosis rates of the IEC-6 cells were increased obviously at 6 h and 24 h after neutron irradiation of 4Gy. The apoptosis rates of the IEC-6 cells were the peak value at 6 h after neutron irradiation, while most of the cells appeared necrosis at 24 h after neutron irradiation. The apoptosis and necrosis rates of the LY294002 treatment group IEC-6 cells were higher than those of the neutron irradiation group cells. Figure 3(a) (A, B, C, D, E, F) showed the scatterplot of apoptosis and necrosis of IEC-6 exposed to neutron irradiation of 4Gy and treated by LY294002. Figures 3(b) and 3(c) showed the results of statistical analysis.

3.4. Expressions of the Key Signaling Molecule of NF-κB Signaling Pathway in IEC-6 (Figure 4)

3.4.1. NF-κB (p65) and Phosphor-NF-κB. The expressions of NF-κB (p65) in the IEC-6 cells were upregulation at 6 h and 24 h after neutron irradiation of 4Gy. The expressions of phosphor-NF-κB in the IEC-6 cells were upregulation from 30 min to 6 h (reaching peak at 6 h) after neutron irradiation of 4Gy, while no expressions were assayed at 12 h and 24 h after neutron irradiation of 4Gy. The expressions of NF-κB (p65) and phosphor-NF-κB were inhibited by LY294002.

3.4.2. IKKα, IKKβ, and Phosphor-IKKα/β. The expressions of IKKα and IKKβ in the IEC-6 cells were upregulation at 6 h and 24 h after neutron irradiation of 4Gy. There were two subunits in IKKs (α/β), so two electrophoresis strips were assayed for phosphor-IKKα/β. The expressions of phosphor-IKKα/β in the IEC-6 cells were upregulation from 30 min to 6 h (reaching peak at 6 h) after neutron irradiation of 4Gy, while no expressions were assayed at 12 h and 24 h after neutron irradiation of 4Gy. The expressions of IKKα, IKKβ, and phosphor-IKKα/β were inhibited by LY294002.

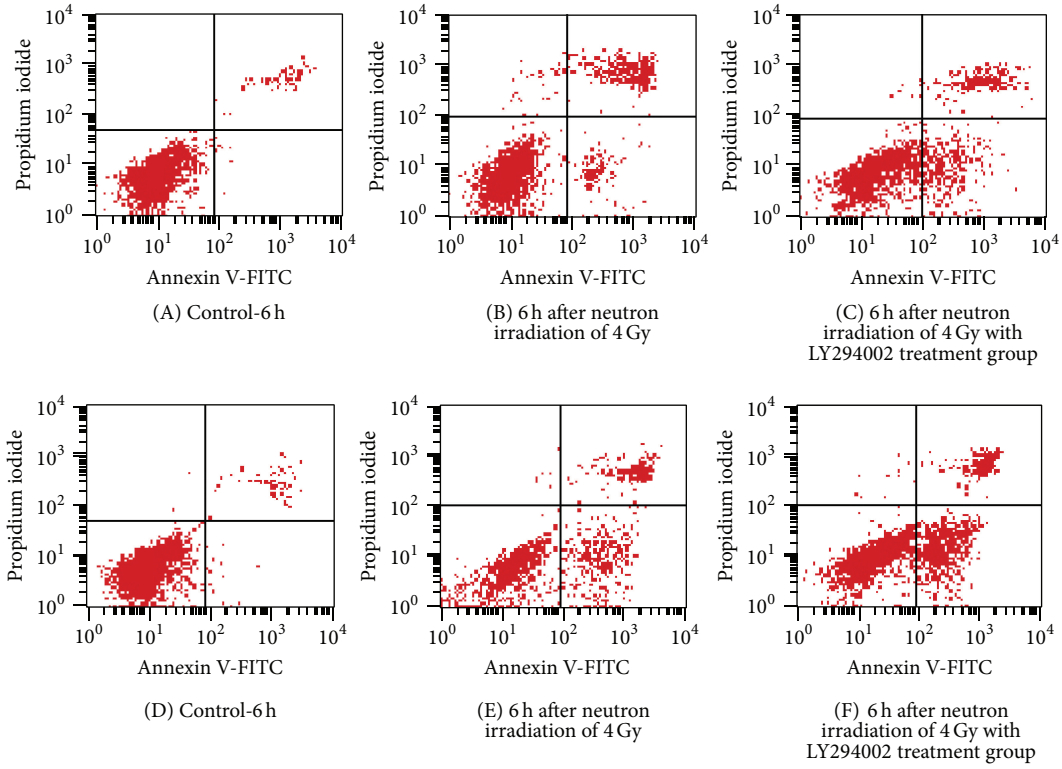
3.4.3. IκBα and Phosphor-IκBα. The expressions of IκBα in the IEC-6 cells were downregulation at 6 h and 24 h after neutron irradiation of 4Gy. No expressions of phosphor-IκBα were assayed at 30 min and 2 h after neutron irradiation of 4Gy, while the expressions of phosphor-IκBα reached peak at 6 h after neutron irradiation of 4Gy. No expressions were assayed at 12 h and 24 h after neutron irradiation of 4Gy. The expressions of IκBα and phosphor-IκBα were increased by LY294002.

3.5. Expressions of PI3K and Phosphor-PI3K, Akt and Phosphor-Akt in IEC-6. The expressions of PI3K in the IEC-6 cells were upregulation at 6 h and 24 h after neutron irradiation of 4Gy. The expressions of phosphor-PI3K and phosphor-Akt in the IEC-6 cells were upregulation from 30 min to 6 h (reaching peak at 6 h) after neutron irradiation of 4Gy, while no expressions were assayed at 12 h and 24 h after neutron irradiation of 4Gy. The expressions of PI3K and phosphor-PI3K, Akt and phosphor-Akt were inhibited by LY294002. Figures 5 and 6 showed the results of Western blotting.

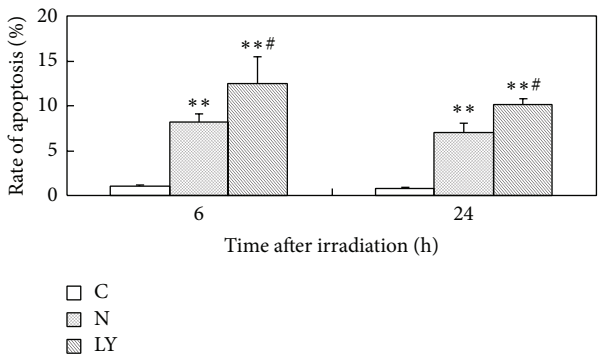
4. Discussion

As we all know, neutron is high lineal energy transfer (LET) ionizing irradiation which can cause bodies more severe damage than γ rays. The intestine is highly sensitive to neutron irradiation, severely injured by neutron irradiation, and hard to recover. Unfortunately, there is still no effective therapeutic measure so far. IEC-6 cells coming from the normal SD rat jejunum crypt epithelial cells can reflect the characteristics of the intestine epithelial cells and can be generally used in the study on the in vitro model of intestine diseases. IEC-6 cells are highly sensitive to ionizing radiation. When the cell lines were irradiated by ionizing radiation, their proliferation activity decreased seriously and there was obvious dose-effect relationship [6]. Therefore, in this study, intestinal epithelial cell (IEC) model was made which was injured by neutron irradiation of 4Gy. We would investigate NF-κB signaling pathway in the regulation of IEC damaged by neutron irradiation. This could be sought to elucidate the molecular mechanism of neutron irradiation-induced intestinal injury, which might help to find new potential therapies. At the same time, we would study how PI3K (the upstream signaling molecule of NF-κB) regulates the NF-κB signaling pathway.

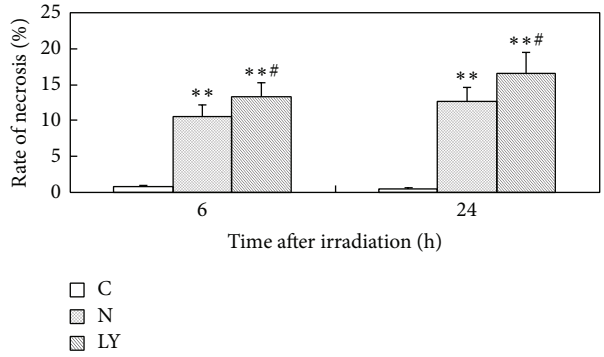
NF-κB is an important nuclear factor which resides in cells widely. It is composed of p65 and p50 which are the two important subunits. NF-κB combines IκB (repressor of NF-κB) which forms an unreactive trimer in the quiescent condition cytoplasm. Lots of extracellular harmful factors such as tumour necrosis factor-α, lymphotoxin-β, and irradiation



(a) The scatterplots of apoptosis and necrosis of IEC-6 exposed neutron irradiation of 4Gy and using of LY294002



(b) The apoptosis rates of IEC-6 cells after neutron irradiation of 4Gy and using of LY294002

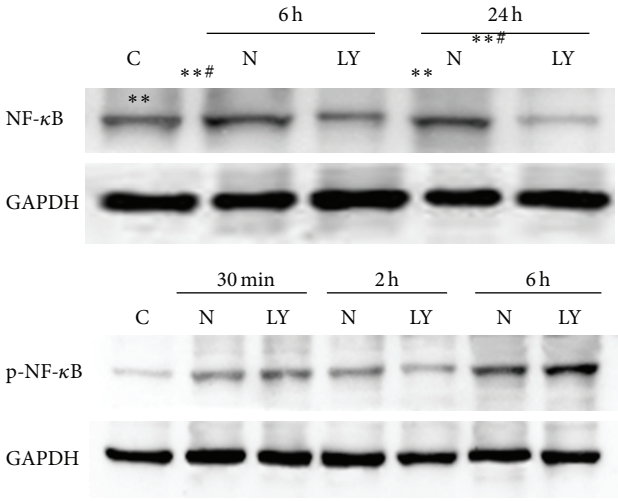


(c) The necrosis rates of IEC-6 cells after neutron irradiation of 4Gy and using of LY294002

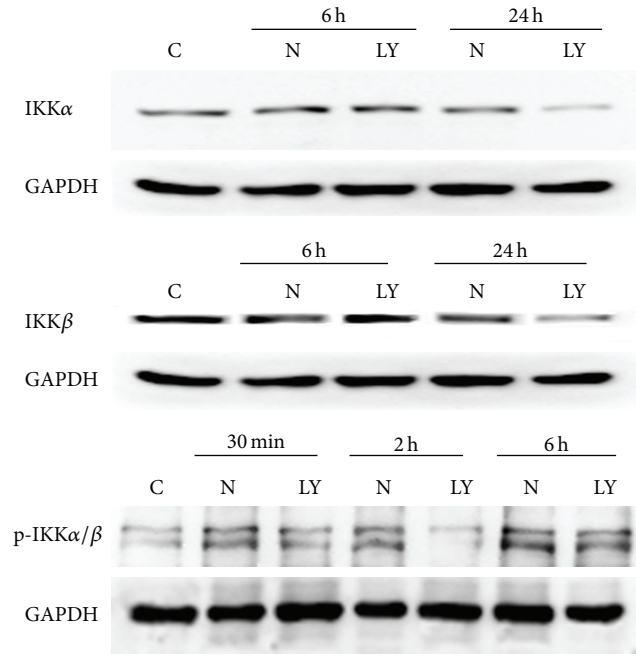
FIGURE 3: The apoptosis and necrosis rates of IEC-6 cells after neutron irradiation of 4Gy and using of LY294002 (C: control group; N: neutron irradiation group; LY: neutron irradiation of 4Gy with LY294002 treatment group).

can activate IκB kinase (IKK) which can lead to IκB phosphorylation and ubiquitination, degradation. At the same time, as soon as NF-κBp50/65 subunits are liberated in the cytoplasm, NF-κB is then free to translocate to the nucleus and bind DNA leading to the activation of target genes [6–8]. In the previous reports, many studies showed NF-κB regulates the target genes correlation with immune and inflammatory reaction. However, now, more and more studies showed NF-κB also could regulate some genes correlation with cell proliferation, apoptosis, and differentiation [9, 10].

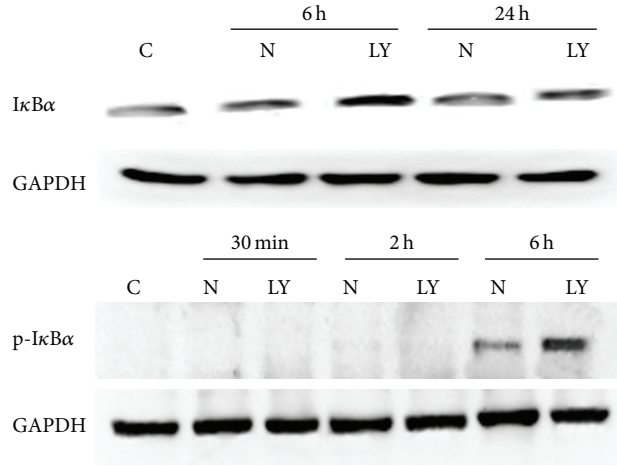
PI3K is the upstream molecule of NF-κB signaling pathway. NF-κB signaling pathway activation depends on PI3K/Akt activating in some cell injury models. Activating NF-κB signaling pathway can negatively regulate apoptosis and improve the cells proliferation [11, 12]. LY294002 is the classic inhibitor of PI3K; it can specifically inhibit the activation of PI3K which can lead to inhibiting the activation of PI3K/Akt signaling pathway [13]. Some extracellular harmful factors such as virus, interferon, and irradiation can activate PI3K/Akt signaling pathway which can induce NF-κB signaling pathway activation [14, 15], while up to now, there are no studies on the function mechanism of NF-κB signaling pathway in intestinal epithelial cell injured by neutron irradiation and how



(a) The expressions of NF-κB and p-NF-κB in IEC-6 cells after neutron irradiation of 4Gy and using of LY294002 (C: control group; N: neutron irradiation group; LY: neutron irradiation of 4Gy with LY294002 treatment group)



(b) The expressions of IKKα, IKKβ, and p-IKKα/β in IEC-6 cells after neutron irradiation of 4Gy and using of LY294002 (C: control group; N: neutron irradiation group; LY: neutron irradiation of 4Gy with LY294002 treatment group)



(c) The expressions of IκBα and p-IκBα in IEC-6 cells after neutron irradiation of 4Gy and using of LY294002 (C: control group; N: neutron irradiation group; LY: neutron irradiation of 4Gy with LY294002 treatment group)

Figure 4: The expressions of key molecules of NF-κB signaling pathway in IEC-6 cell lines after neutron irradiation of 4Gy and using of LY294002 (C: control group; N: neutron irradiation group; LY: neutron irradiation of 4Gy with LY294002 treatment group).

PI3K/Akt signaling pathway regulates NF-κB signaling pathway.

In this study, we found IEC-6 cell lines showed serious apoptosis and necrosis and cell proliferation activity depression after being irradiated by neutron of 4Gy, while the IEC-6 cells treated by LY294002 (inhibitor of PI3K) and exposed to neutron irradiation of 4Gy showed higher apoptosis and necrosis rates than the neutron irradiation group cells. The expression of the critical signaling molecules of NF-κB signaling pathway in the IEC-6 cells such as NF-κB (p65) and IKKα/β was upregulated after being irradiated by neutron of 4Gy, while IκBα (the repressor of NF-κB) was downregulated. The expression of PI3K in the IEC-6 cell lines was upregulated after neutron irradiation of 4Gy, while using LY294002 could inhibit the expression of PI3K. We also found LY294002 could downregulate the expression of NF-κB (p65) and IKKα/β. These results showed neutron irradiation could activate the NF-κB signaling pathway in

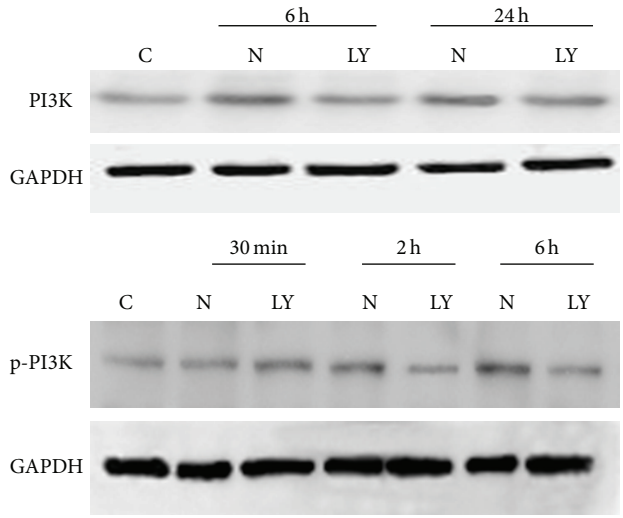


Figure 5: The expressions of PI3K and p-PI3K in IEC-6 cells after neutron irradiation of 4Gy and using of LY294002 (C: control group; N: neutron irradiation group; LY: neutron irradiation of 4Gy with LY294002 treatment group).

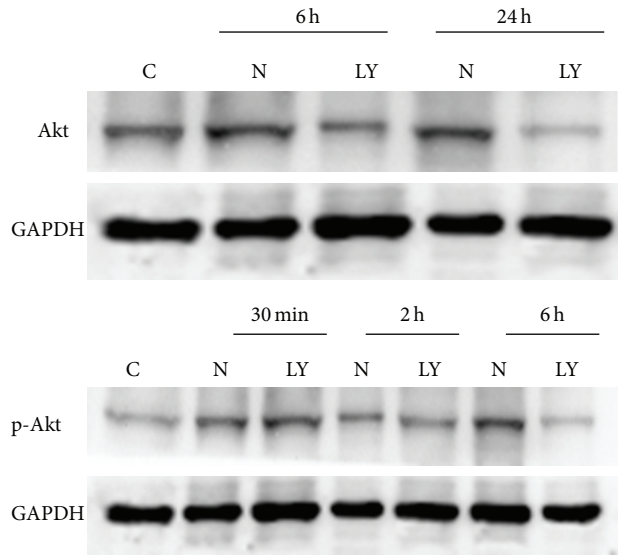


Figure 6: The expressions of Akt and p-Akt in IEC-6 cells after neutron irradiation of 4Gy and using of LY294002 (C: control group; N: neutron irradiation group; LY: neutron irradiation of 4Gy with LY294002 treatment group).

IEC-6 cells. Moreover, PI3K could positively regulate NF-κB signaling pathway in IEC-6 cells, while using the inhibitor of PI3K could also inhibit the activation of NF-κB pathway which led to aggravating the IEC-6 cells damage.

In this study, we explored the function of NF-κB signaling pathway in the IEC-6 injured by neutron irradiation. We found neutron irradiation could activate NF-κB signaling pathway and PI3K positively regulated NF-κB signaling pathway in IEC-6 cells, while LY294002, inhibitor of PI3K, could also inhibit the activation of NF-κB signaling pathway in IEC-6 cells irradiated by neutron of 4Gy which led to aggravating the IEC-6 cells injury. On the contrary, this result indicated that activating NF-κB signaling pathway could protect the IEC-6 cells injured by neutron irradiation. This discovery provided theoretical evidence and foundation for elucidating the molecule mechanism of intestine damaged by neutron irradiation and finding new therapy target.

5. Conclusions

Results of this study suggest that IEC-6 cells were obviously damaged and induced serious apoptosis and necrosis by neutron irradiation of 4Gy; the NF-κB signaling pathway in IEC-6 was activated by neutron irradiation which could protect IEC-6 against injuries by neutron irradiation; LY294002 could inhibit the proliferation activity of IEC-6 cells.

Conflict of Interests

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

Acknowledgment

This paper is supported by the National Natural Science Foundation of China (Grant nos. 81202146, 30870735).

References

[1] T. D. Gilmore, "Introduction to NF-κB: players, pathways, perspectives," *Oncogene*, vol. 25, no. 51, pp. 6680–6684, 2006.

[2] M. Pasparakis, T. Luedde, and M. Schmidt-Supprian, "Dissection of the NF-κB signalling cascade in transgenic and knockout mice," *Cell Death and Differentiation*, vol. 13, no. 5, pp. 861–872, 2006.

[3] J. Kucharczak, M. J. Simmons, Y. Fan, and C. Gélinas, "To be, or not to be: NF-κB is the answer: role of Rel/NF-κB in the regulation of apoptosis," *Oncogene*, vol. 22, no. 56, pp. 8961–8982, 2003.

[4] Z. Peng, L. Peng, Y. Fan, E. Zandi, H. G. Shertzer, and Y. Xia, "A critical role for IκB kinase β in metallothionein-1 expression and protection against arsenic toxicity," *Journal of Biological Chemistry*, vol. 282, no. 29, pp. 21487–21496, 2007.

[5] N. M. Dagia, G. Agarwal, D. V. Kamath et al., "A preferential p110α/γ PI3K inhibitor attenuates experimental inflammation by suppressing the production of proinflammatory mediators in a NF-κB-dependent manner," *American Journal of Physiology: Cell Physiology*, vol. 298, no. 4, pp. C929–C941, 2010.

[6] X. Z. Ran, Y. P. Su, T. M. Cheng et al., "A simple method for the culture of rat small intestinal epithelial cells," *China Journal of Modern Medicine*, vol. 101, part 1, pp. 8–12, 1997.

[7] S. G. Pereira and F. Oaklev, "Nuclear factor-kappaB1: regualation and function," *The International Journal of Biochemistry & Cell Biology*, vol. 40, no. 8, pp. 1425–1430, 2008.

[8] A. Lavorgna, R. De Filippi, S. Formisano, and A. Leonardi, "TNF receptor-associated factor 1 is a positive regulator of the NF-κB alternative pathway," *Molecular Immunology*, vol. 46, no. 16, pp. 3278–3282, 2009.

[9] N. Magné, C. Didelot, R.-A. Toillon, P. van Houtte, and J.-F. Peyron, "Biomodulation of transcriptional factor NF-κB by

ionizing radiation," *Cancer/Radiothérapie*, vol. 8, no. 5, pp. 315–321, 2004.

[10] J.-L. Luo, H. Kamata, and M. Karin, "IKK/NF-κB signaling: Balancing life and death—a new approach to cancer therapy," *Journal of Clinical Investigation*, vol. 115, no. 10, pp. 2625–2632, 2005.

[11] M. Esfandiarei, S. Boroomand, A. Suarez, X. Si, M. Rahmani, and B. McManus, "Coxsackievirus B3 activates nuclear factor κ B transcription factor via a phosphatidylinositol-3 kinase/protein kinase B-dependent pathway to improve host cell viability," *Cellular Microbiology*, vol. 9, no. 10, pp. 2358–2371, 2007.

[12] X. Zhang, B. Jin, and C. Huang, "The PI3K/Akt pathway and its downstream transcriptional factors as targets for chemoprevention," *Current Cancer Drug Targets*, vol. 7, no. 4, pp. 305–316, 2007.

[13] A. K. Gupta, G. J. Cerniglia, R. Mick et al., "Radiation sensitization of human cancer cells in vivo by inhibiting the activity of PI3K using LY294002," *International Journal of Radiation Oncology* Biology* Physics*, vol. 56, no. 3, pp. 846–853, 2003.

[14] H.-G. Yu, Y.-W. Ai, L.-L. Yu et al., "Phosphoinositide 3-kinase/Akt pathway plays an important role in chemoresistance of gastric cancer cells against etoposide and doxorubicin induced cell death," *International Journal of Cancer*, vol. 122, no. 2, pp. 433–443, 2008.

[15] X. Chao, J. Zao, G. Xiao-Yi, M. Li-Jun, and S. Tao, "Blocking of PI3K/AKT induces apoptosis by its effect on NF-κB activity in gastric carcinoma cell line SGC7901," *Biomedicine & Pharmacotherapy*, vol. 64, no. 9, pp. 600–604, 2010.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 909657, 7 pages
http://dx.doi.org/10.1155/2014/909657

Research Article

Human Umbilical Cord Mesenchymal Stem Cells Infected with Adenovirus Expressing *HGF* Promote Regeneration of Damaged Neuron Cells in a Parkinson's Disease Model

Xin-Shan Liu,[1] Jin-Feng Li,[2] Shan-Shan Wang,[1] Yu-Tong Wang,[3] Yu-Zhen Zhang,[2] Hong-Lei Yin,[2] Shuang Geng,[2] Hui-Cui Gong,[2] Bing Han,[2] and Yun-Liang Wang[2]

[1] *Weifang Medical University, 7166 Baotong Road, Weifang 261053, China*
[2] *The Neurology Department, The 148th Hospital, 20 Zhanbei Road, Zibo 255300, China*
[3] *Medical College of Henan University, 357 Jinming Road, Kaifeng 475001, China*

Correspondence should be addressed to Bing Han; icecold148@163.com and Yun-Liang Wang; wangyunliang81@163.com

Received 28 June 2014; Revised 22 July 2014; Accepted 5 August 2014; Published 3 September 2014

Academic Editor: Shiwei Duan

Parkinson's disease (PD) is a neurodegenerative movement disorder that is characterized by the progressive degeneration of the dopaminergic (DA) pathway. Mesenchymal stem cells derived from human umbilical cord (hUC-MSCs) have great potential for developing a therapeutic agent as such. HGF is a multifunctional mediator originally identified in hepatocytes and has recently been reported to possess various neuroprotective properties. This study was designed to investigate the protective effect of hUC-MSCs infected by an adenovirus carrying the *HGF* gene on the PD cell model induced by MPP+ on human bone marrow neuroblastoma cells. Our results provide evidence that the cultural supernatant from hUC-MSCs expressing HGF could promote regeneration of damaged PD cells at higher efficacy than the supernatant from hUC-MSCs alone. And intracellular free Ca^{2+} obviously decreased after treatment with cultural supernatant from hUC-MSCs expressing HGF, while the expression of CaBP-D28k, an intracellular calcium binding protein, increased. Therefore our study clearly demonstrated that cultural supernatant of MSC overexpressing *HGF* was capable of eliciting regeneration of damaged PD model cells. This effect was probably achieved through the regulation of intracellular Ca^{2+} levels by modulating of CaBP-D28k expression.

1. Introduction

Parkinson's disease (PD) was first described by British physician James Parkinson in 1817, which is a progressive degenerative disease of the central nervous system of the elderly. The clinical symptoms include tremor, muscle rigidity, bradykinesia, and postural instability, which seriously affect the quality of life. The main pathological changes are believed to be the degeneration of dopaminergic neurons in the substantia nigra pars compacta (SNpc) and their terminals in the striatum [1, 2]. Current treatment for PD relies on medicine, such as levodopa, that alleviates early symptoms but failed to prevent disease progression. In recent years, cell and gene therapies have gained traction in the treatment of PD with the focus on the regeneration of dopamine (DA) producing neurons [3, 4]. Mesenchymal stem cells are multipotent that can be differentiated into various cells of mesodermal lineage. They can be obtained from several sources including bone marrow, adipose tissue, placenta, umbilical cord, and cord blood. Human umbilical cord mesenchymal stem cells (hUC-MSC) due to its availability through noninvasive procedure have demonstrated advantages that become one of the top choices for repairing the damaged neurons [5]. Hepatocyte growth factor (HGF) is a multifunctional growth factor produced by stromal cells. It activates the signal transduction cascade through tyrosine phosphorylating its protooncogenic c-Met receptor. Although it was discovered originally as a growth factor for hepatocytes, it has been demonstrated to be involved in differentiation, proliferation, and regeneration of variety of cells [6–8]. Expression of HGF is found in human brain tissue and is believed to be a survival factor for motor and sensory neurons [9–11]. Salehi and Rajaei reported that

HGF could be involved in the pathogenesis of Parkinson's disease [12]. One of the sources for HGF is MSC itself, and it has been shown that HGF signaling plays a critical role during organogenesis [13].

Our previous study has shown that hUC-MSC being infected by an adenovirus carrying the *HGF* gene (Ad-*HGF*) can express dopaminergic neuron specific marker tyrosine hydroxylase and dopamine transporter; in addition, the dopamine levels in the cultural medium of these cells increase significantly [14]. Our finding indicated that hUC-MSC when overexpressing HGF has the differentiation potential for dopaminergic neuron. We hypothesized that the supernatant of hUC-MSC-HGF may contain important factors for neuron cell differentiation and damage repair. To verify that, we designed the present study to treat PD model cells with cell culture supernatant of hUC-MSC overexpressing HGF, and the recovery of cell viability was observed and mechanisms were investigated.

2. Materials and Methods

2.1. Reagents, Antibodies, and the Expression Vectors. Adenovirus expressing green fluorescent protein gene (Ad-*GFP*) was provided by Beckman Medical Instruments, USA. Recombinant adenovirus carrying *HGF* gene (Ad-*HGF*) was constructed in our lab. Umbilical cord tissue was obtained from the Gynecology Department of the 148 Hospital. All donors have signed informed consent and the study was approved by the ethics committee of the 148 Hospital. MPTP, CaBP-D28k antibody, and Fluo-3-AM were purchased from Sigma (USA).

2.2. hUC-MSC Preparation. Isolation and verification of hUC-MSC were carried out based on previously described protocols, and determination on Ad-*HGF* optimal transfection efficiency was performed based on protocols described previously [14].

2.3. Preparation of Conditioned Medium. The hUC-MSC cells were infected by Ad-*HGF* at 200 MOI for 48 hours, and the supernatant (CM-HGF) was centrifuged using ultrafiltration tubes at 3000 r/min at 4°C for 1.5 h. The centrifugation was repeated three times until the original culture supernatant was concentrated for 7.5 folds and subsequently was stored at −80°C. Supernatant from hUC-MSC cells culture medium without adenovirus infection (CM-MSC) was also centrifuged in the same manner.

2.4. Analysis of HGF Protein Level by ELISA. The hUC-MSCs were transfected with Ad-*HGF* at a MOI of 200 in serum-free F12 for 2 h. The supernatants were harvested at different time-points after transduction (24 h, 48 h, and 72 h). Concentrations of immunoreactive HGF in the supernatant were measured by enzyme linked immunosorbent assay (ELISA). ELISA plates (R&D system) were coated overnight at room temperature with 100 μL of a 12 μg/mL solution of affinity purified anti-HGF diluted in 1X antibody coating buffer. Following 2 washes with 1X wash buffer, the plate was blocked with 300 μL of 1X General Blocker Buffer for 3–6 h at room temperature. Blocking solution was removed, and the plate firmly was tapped on absorbent paper to remove excess liquid and was used immediately. Ninety-five μL of general assay dilutant was added to each well, followed by 5 μL of standard blank (purified HGF) or serum samples. The plate was then sealed and incubated overnight at 4°C. Following 5 washes with 1X washing buffer, 100 μL of HRP-conjugated secondary antibody was added to each well and the plate was incubated again for 1 h at room temperature. The plate was washed 5 additional times. After complete removal of excess solution, 100 μL of TMB substrate was added to each well. Following a 15-minute incubation period at room temperature, 100 μL of stop solution was added and the absorbance at 450 nm was read using a plate reader (Bio-Rad). Concentrations of HGF in the samples were calculated relative to the exponential standard curve obtained from the standard included in each assay.

2.5. Preparation of PD Cell Model. SH-SY5Y cells in logarithmic growth phase were incubated in 96 well plates at a density of 5×10^4/mL and 100 μL/well for 12 h and then treated with MPP+ of various concentrations for 24 h. Cell viability was assessed by adding 10 μL of CCK-8 (Dojindo Molecular Technologies) to the culture and continuing incubation for 2 h. Cell viability was measured by spectrometry with 450 nm wavelength. The concentration of MPP+ for desired cell damage was determined for PD cell model based on cell viability.

2.6. SH-SY5Y Cell Proliferation Measurement. SH-SY5Y cells are divided into 4 groups: normal cells (control group), cells treated with MPP+ (model group), PD model treated with CM-MSC for 24 or 48 h (CM-MSC group), and PD model treated with CM-HGF for 24 or 48 h (CM-HGF group). Cell viability was assessed by CCK-8 assay as described above.

2.7. Observation on Ca^{2+} Changes in SH-SY5Y Cells by Confocal Microscope. SH-SY5Y cells were washed 3 times with PBS after various treatments and incubated with Fluo-3-AM (5 μM) at 37°C or room temperature for 0.5 to 1 h. Cells were rinsed 2~3 times and observed under confocal microscope. Cell images were analyzed by LaserSharp 2000 image analysis software.

2.8. Western Blot. Cells were collected and lysed after various treatments indicated. The protein concentration was determined using the Bio-Rad protein assay kit (Bio-Rad, Hercules, CA, USA). The samples were separated by 12% SDS-PAGE and transferred to PVDF membranes. CaBP-D28k expression was detected with rabbit anti-CaBP-D28k antibody (Sigma) and subsequent HRP-conjugated secondary antibody. Image was analyzed for densitometry with BIO-RAD Quantity One software. The blots were probed with an anti-β-actin antibody (Sigma) for loading quantity.

2.9. Statistical Analysis. Statistical analysis was performed using SPSS10.0 software. One-way ANOVA was used for

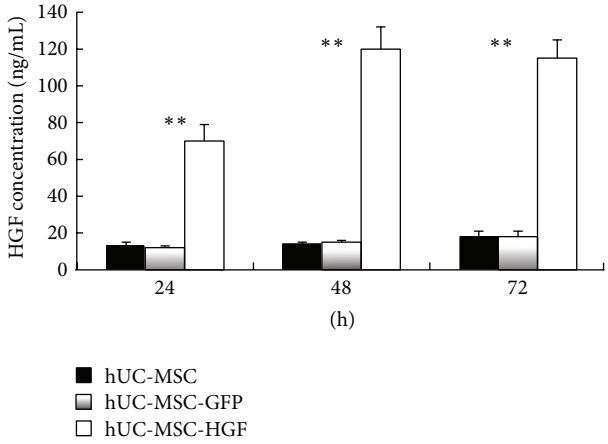


Figure 1: The concentration of HGF in cells supernatant was detected by ELISA. After the hUC-MSCs were transduced with Ad-HGF, HGF gradually accumulated in hUC-MSCs supernatant, peaking at about 120 ng/mL at 48 h. Levels remained stable at about 15 ng/mL in the Ad-GFP and blank control group. Indeed, statistically significant differences emerged between the control and Ad-HGF groups upon analysis ($^{**}P < 0.01$).

comparison between two groups. All data were presented as mean ± standard deviation ($x \pm s$), and $P < 0.05$ was considered statistically significant.

3. Results

3.1. hUC-MSC Isolation. We have previously shown that hUC-MSC can be successfully isolated from human umbilical cord, which expresses CD29, CD44, and CD105, the known surface markers for mesenchymal stem cells, but not hematopoietic stem cell marker CD45 or epithelia cell marker CD31 [14]. The maximum infection efficiency can be achieved when the hUC-MSC cells were infected by adenovirus at the m.o.i. of 200, as shown by flow cytometry and fluorescence microscopy (data not shown).

3.2. The Concentration of HGF Increased in hUC-MSCs Supernatant after Transduction with Ad-HGF. As shown in Figure 1, HGF accumulated to about 75 ng/mL in the supernatant at 24 h in Ad-HGF groups. It gradually increased, peaking at 120 ng/mL at 48 d. At 72 d, concentration of HGF declined slightly to approximately 115 ng/mL. In blank control and Ad-GFP groups, HGF concentration remained stable at about 15 ng/mL (Figure 1). It is clear that HGF protein level obviously increased after infection with Ad-HGF than hUC-MSC cells alone at different time-point ($^{**}P < 0.01$).

3.3. Establishment of PD Model with MPP+ Treated SH-SY5Y Cells. Human neuroblastoma cell line SH-SY5Y is one of the widely used cell lines for studying the neurodegeneration and neurotoxicity related to PD. It has been shown that undifferentiated SH-SY5Y cells are susceptible to neurotoxin such as MMP+. SH-SY5Y cells were treated with various concentrations of MMP+ for 24 h. CCK-8 assay was used to

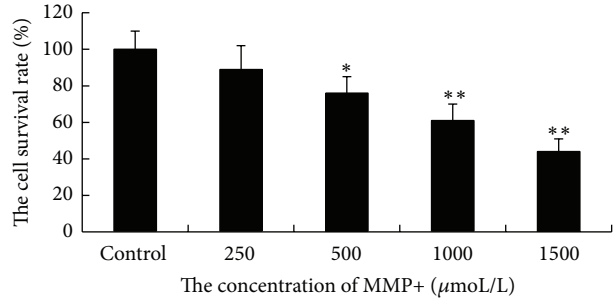


Figure 2: Cell viability of SH-SY5Y after being treated with various concentrations of MMP+ as assessed by CCK-8 assay. The results show that, with the increasing concentration of MMP+, SH-SY5Y cell viability decreased significantly at 500, 1000, and 1500 μM MMP+ compared to the control group ($^{*}P < 0.05$, $^{**}P < 0.01$).

assess the survival of the cells after treatment. A decreased cell viability was seen with the increase of the MMP+ concentration. Cell survival was about 90% at 250 μmol/L of MMP+ and dropped to 73.09%, 60.0%, and 50.0% at the concentrations of 500, 1000, and 1500 μM MPP+, respectively. Significant differences were found in the later three groups compared with the control group (Figure 2, $P < 0.05$). Treatment of MMP+ at a concentration of 1000 μM was chosen for PD model.

3.4. Protection on Damages of PD Model by CM-HGF and CM-MSC. SH-SY5Y cells treated with MMP+ at 1000 μM for 24 h were used as the established PD model. These cells were incubated with CM-HGF or CM-MSC, and CCK-8 assay was employed to study the viability of the cells. As shown in Figure 3(a), both CM-HGF and CM-MSC treatments were able to regenerate the damaged SH-SY5Y cells. At 48 hours after treatment, the proliferative effect from CM-HGF was more significant than that from CM-MSC ($P < 0.05$). Cell viability of PD model cells treated with CM-HGF was significantly higher compared to that of CM-MSC treated or untreated normal cells. Viability of PD model cells treated with CM-MSC as measured by O.D. 450 showed no significant difference compared to the normal control group at 48 h culture ($P > 0.05$), while the OD 450 value of the CM-HGF treated group was significantly higher than the normal control group ($P < 0.05$, Figure 3(a)). These results suggested that CM-HGF and CM-MSC could promote regeneration of SH-SY5Y. Under the microscope, while the control cells were showed in good condition with clear edge, most of the PD model cells (MMP+ treated for 72 h) appeared to have nuclear condensation and drastically shrunken cell bodies. However, these morphologies were significant improved after these cells were treated with CM-HGF and CM-MSC (Figure 3(b)).

3.5. Intracellular Ca^{2+} after Being Treated with CM-HGF and CM-MSC. Fluo-3-AM can be used as a calcium indicator due to its property of marked increased fluorescence intensity when it is bound with Ca^{2+}. It has low affinity for Ca^{2+} and can be readily dissociated. Therefore, it is ideal for

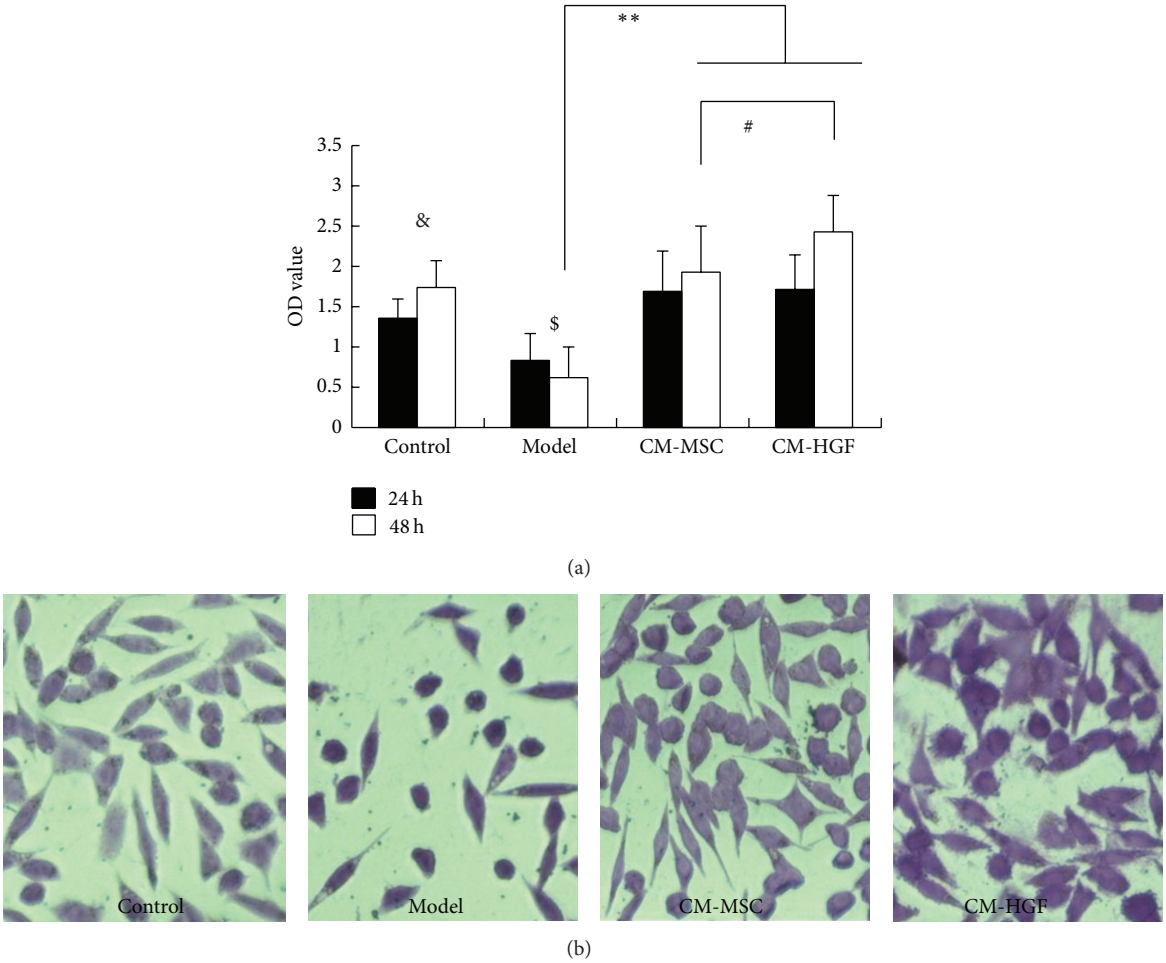


Figure 3: (a) Cell viability of SH-SY5Y underwent various treatments. Control: no-treatment; model: treated with MMP+ at 1000 μM; CM-MSC: treated with MMP+ at 1000 μM and then incubated with cultural supernatant of hUC-MSC; CM-HGF: treated with MMP+ at 1000 μM and then incubated with cultural supernatant of hUC-MSC infected with Ad-*HGF*. Viability in PD model cells was significantly decreased compared to control ($^{\$}P < 0.01$). After being treated with either CM-HGF or CM-MSC for 48 h, cell viability was significantly increased compared to the PD model ($^{**}P < 0.01$) and compared with the normal control group ($^{\&}P < 0.05$). Significant difference between CM-MSC and CM-HGF treatment was seen after 48 h treatment ($^{\#}P < 0.05$). (b) Microscopic images of cells with various treatments. The SH-SY5Y cells under each treatment were stained with crystal violet and observed under an inverted microscope.

measuring rapid and minimal changes of intracellular Ca^{2+}. LSCM was used to study the intracellular Ca^{2+} in these cells. Compared with the normal control cells, PD model cells showed enhanced intracellular fluorescence upon Fluo-3-AM staining, indicating an increase of intracellular free Ca^{2+}. Intracellular fluorescence intensity weakened after these cells were treated with CM-HGF or CM-MSC, indicating decreased intracellular free Ca^{2+} (Figure 4(a)). Cell fluorescence intensity was quantified with LaserSharp 2000 software by quantifying randomly picked 10 cells in each optical area (Figure 4(b)). The fluorescence intensity in cells of the PD model was significantly higher than that of the normal control cells ($^{*}P < 0.05$), whereas the fluorescence intensity in cells of the PD model treated with CM-MSC, though lower than that of the untreated PD model cells, was still significantly higher than that of the normal cells ($^{**}P < 0.01$). The fluorescence intensity of PD model treated with CM-HGF was higher than that of the normal control, but there is no significant difference between them ($P > 0.05$). Comparing the intracellular fluorescence intensity between PD model cells and PD model cells treated with either CM-HGF or CM-MSC, it was found that both treatments significantly decreased the intracellular free Ca^{2+} levels ($^{\#}P < 0.01$). In addition, CM-HGF treatment showed a better efficacy in reducing the intracellular free Ca^{2+} levels than CM-MSC treatment, as indicated by the lower fluorescence intensity in CM-HGF treated cells than in the CM-MSC treated cells ($^{\$}P < 0.05$, Figure 4(b)).

3.6. Expression of CaBP-D28k in CM-MSC and CM-HGF Treated Cells. Calbindin CaBP-D28k is a high affinity calcium-binding protein that plays an important role in calcium homeostasis. CaBP-D28k has been indicated to confer protection to SNC dopaminergic neurons against

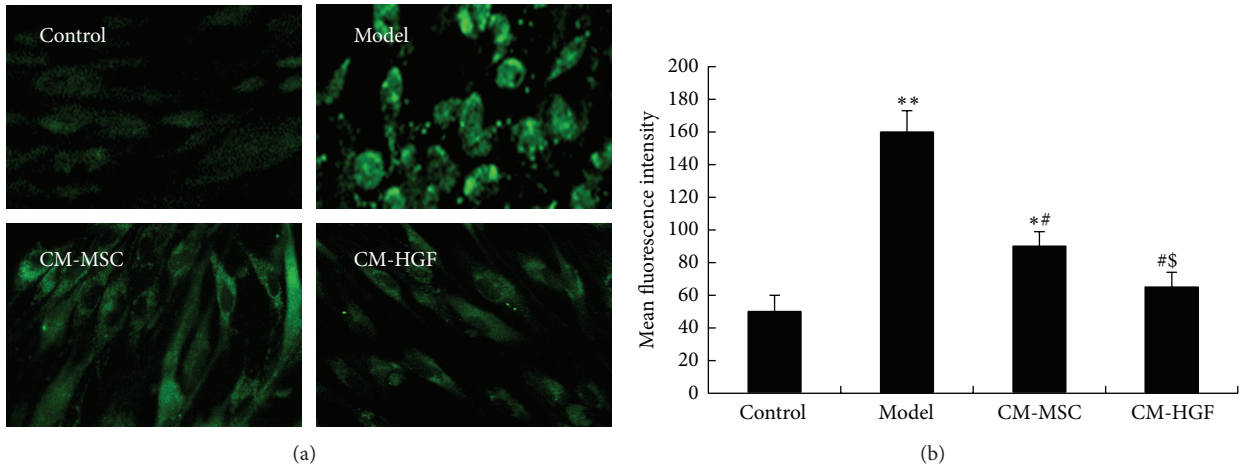


Figure 4: (a) Confocal microscope image of cells underwent various treatments. Fluo-3-AM was added to cells after incubation with CM-MSC or CM-HGF, and fluorescence was observed with LSCM. (b) Fluorescence intensity analysis with LaserSharp 2000 software. $^{*}P < 0.05$, between CM-MSC and control; $^{**}P < 0.01$, between PD model and control; $^{\#}P < 0.01$, between CM-HGF or CM-MSC and PD model; $^{\$}P < 0.05$, between CM-MSC and CM-HGF.

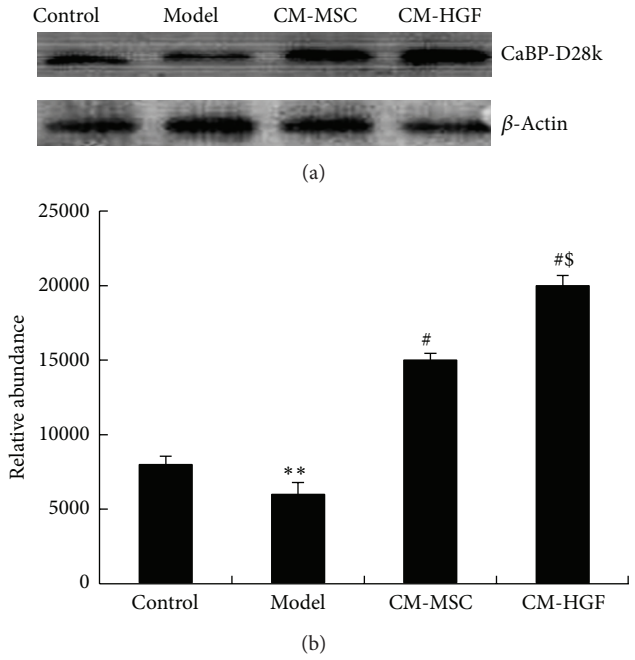


Figure 5: (a) Western blot results of total cell lysates for CaBP-D28k expression after various treatments. (b) Densitometry analysis with Quantity One software. β-actin was used as a loading quantity reference on which the expression levels of CaBP-D28k were normalized. $^{**}P < 0.01$, between model cells and control; $^{\#}P < 0.05$, between CM-HGF or CM-MSC and model cells; $^{\$}P < 0.05$, between CM-HGF and CM-MSC.

certain pathological process related to PD. The expression of CaBP-D28k was assessed by Western blot analysis in the PD cell model that underwent various treatments. As shown Figure 5, the expression of CaBP-D28k in SH-SY5Y cells treated with MPP+ (PD model cell) was significantly decreased compared to the normal cells. However, when these cells were treated with CM-MSC or CM-HGF, the expression of CaBP-D28k was elevated. Densitometry analysis showed that the differences between the normal control and PD model cells were statistically significant ($^{**}P < 0.01$). Both CM-HGF and CM-MSC treatments significantly upregulated CaBP-D28k expression ($^{\#}P < 0.05$); however, the efficacy of CM-HGF was shown to be significantly more potent than that of CM-MSC ($^{\$}P < 0.05$).

4. Discussion

MSCs are multipotent stem cells that can be easily obtained and expanded without getting involved with ethical issues. MSCs are of low immunogenicity, able to pass through the blood-brain barrier after intravenous transplantation, and have long survival time after transplantation. hUC-MSCs are particular attractive because they can be procured through noninvasive procedure and have abundant sources [5, 15]. MSCs have been shown to migrate to the site of brain injury, and safety of MSCs transplantation into brains has been demonstrated; thus, they are attractive therapeutic options for neurodegenerative disorders [16–19].

The major pathogenesis of PD is its loss of DA neurons, thus making it a good candidate for cell therapy. A lot of attentions have been given to use DA neurons differentiated from stem cells of various sources to replace the degenerated DA neurons [20]. However, studies have been shown that stem cells are also capable of protecting or stimulating the regeneration of damaged DA neurons in host [21, 22].

HGF is widely expressed in the nervous system, although its association with PD was still not clear. Salehi and Rajaei detected higher HGF concentrations in cerebrospinal fluid of PD patients than that of the normal population [12]. Lan et al. found that HGF can mediate proliferation and migration of primary dopaminergic nerve progenitor cells separated from the placenta [23]. A group in Japan injected plasmid carrying

HGF gene into PD rat model and found significant reduction of symptoms, suggesting that overexpression of *HGF* can prevent death of dopaminergic neurons in Parkinson rats [24].

Exogenous HGF protein has a very short half-life *in vivo*, and more critical, HGF is a macromolecular protein and cannot pass through the blood-brain barrier [8, 12]. Meanwhile, mesenchymal stem cells are appropriate cell carriers for exogenous genes [25]. Therefore, we decided to explore the idea of the combination of HGF and hUC-MSCs as a treatment of PD, by introducing *HGF* gene into hUC-MSCs to establish HGF producing MSCs. In our study, the supernatant from hUC-MSCs infected with Ad-*HGF* was more potent in inducing PD model cell regeneration than the supernatant from hUC-MSCs noninfected cells. Our data suggested that combination of HGF and hUC-MSCs had advantage over hUC-MSCs alone.

Damage to DA neurons can be triggered by toxin exposure, increased oxidative stress, and mitochondrial dysfunction, protein aggregation, and inflammation [26]. Calcium is important to normal cell function. Intracellular calcium regulation is closely related to mitochondrial function and oxidative stress, and there is increasing evidence that disruption of intracellular calcium homeostasis is crucial to the pathogenesis of PD [27]. CaBP-D28k is an intracellular calcium binding protein that is expressed in many neurons. It has been shown that neuron expressing CaBP-D28k is less susceptible to the damage [28]. CaBP-D28k is shown to activate Ca^{2+}/Mg^{2+}-ATPase, preventing excessive Ca^{2+} accumulation in the brain, thus, playing a role in Ca^{2+} transport and maintaining calcium homeostasis in the neurons [29, 30]. Recent studies also show that the levels of CaBP-D28k protein in neurons not only represent the Ca^{2+} change but are also closely related to the integrity of the structure and function [31]. Our present study demonstrated that intracellular free Ca^{2+} levels were increased in the PD model cells, and the reductions of the Ca^{2+} levels were correlated to the recovery of cell viability elucidated by the treatment of CM-HGF or CM-MSC. Moreover, CaBP-D28k expression levels in SH-SY5Y PD model cells were reversely correlated with the intracellular Ca^{2+} levels. Expression of CaBP-D28k was decreased in PB model cells but increased after these cells were treated with CM-HGF and CM-MSC. Similar to its effect on PD model cell regeneration, CM-HGF was more potent than CM-MSC in reducing intracellular Ca^{2+} levels and promoting CaBP-D28k expression. Our study clearly demonstrated that cultural supernatant of MSC overexpressing *HGF* was capable of eliciting regeneration of damaged PD model cells. This effect was probably achieved through the regulation of intracellular Ca^{2+} levels by modulating of CaBP-D28k expression. Further studies are needed to understand the active components in the cultural supernatant and the signaling cascade involved.

Conflict of Interests

There is no conflict of interests for any authors. All experiments were reviewed by the Ethics Committee of the 148th Hospital.

Authors' Contribution

Xin-Shan Liu and Jin-Feng Li contributed equally to this work.

References

[1] D. J. Pedrosa and L. Timmermann, "Review: management of Parkinson's disease," *Neuropsychiatric Disease and Treatment*, vol. 9, pp. 321–340, 2013.

[2] G. C. Pluck and R. G. Brown, "Apathy in Parkinson's disease," *Journal of Neurology Neurosurgery and Psychiatry*, vol. 73, no. 6, pp. 636–642, 2002.

[3] A. Bjorklund and J. H. Kordower, "Cell therapy for Parkinson's disease: what next?" *Movement Disorders*, vol. 28, no. 1, pp. 110–115, 2013.

[4] 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.

[5] A. Can and D. Balci, "Isolation, culture, and characterization of human umbilical cord stroma-derived mesenchymal stem cells," *Methods in Molecular Biology*, vol. 698, no. 1, pp. 51–62, 2011.

[6] D. P. Bottaro, J. S. Rubin, D. L. Faletto et al., "Identification of the hepatocyte growth factor receptor as the c-met proto-oncogene product," *Science*, vol. 251, no. 4995, pp. 802–804, 1991.

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

[8] G. K. Michalopoulos and R. Zarnegar, "Hepatocyte growth factor," *Hepatology*, vol. 15, no. 1, pp. 149–155, 1992.

[9] A. Ebens, K. Brose, E. D. Leonardo et al., "Hepatocyte growth factor/scatter factor is an axonal chemoattractant and a neurotrophic factor for spinal motor neurons," *Neuron*, vol. 17, no. 6, pp. 1157–1172, 1996.

[10] F. Maina and R. Klein, "Hepatocyte growth factor, a versatile signal for developing neurons," *Nature Neuroscience*, vol. 2, no. 3, pp. 213–217, 1999.

[11] Y. Tsuboi, K. Kakimoto, M. Nakajima et al., "Increased hepatocyte growth factor level in cerebrospinal fluid in Alzheimer's disease," *Acta Neurologica Scandinavica*, vol. 107, no. 2, pp. 81–86, 2003.

[12] Z. Salehi and F. Rajaei, "Expression of hepatocyte growth factor in the serum and cerebrospinal fluid of patients with Parkinson's disease," *Journal of Clinical Neuroscience*, vol. 17, no. 12, pp. 1553–1556, 2010.

[13] H. Ohmichi, U. Koshimizu, K. Matsumoto, and T. Nakamura, "Hepatocyte growth factor (HGF) acts as a mesenchyme-derived morphogenic factor during fetal lung development," *Development*, vol. 125, no. 7, pp. 1315–1324, 1998.

[14] J. Li, H. 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.

[15] C. Zhou, B. Yang, Y. Tian et al., "Immunomodulatory effect of human umbilical cord Wharton's jelly-derived mesenchymal stem cells on lymphocytes," *Cellular Immunology*, vol. 272, no. 1, pp. 33–38, 2011.

[16] R. A. Barker, J. Barrett, S. L. Mason, and A. Björklund, "Fetal dopaminergic transplantation trials and the future of neural grafting in Parkinson's disease," *The Lancet Neurology*, vol. 12, no. 1, pp. 84–91, 2013.

[17] M. A. Hellmann, H. Panet, Y. Barhum, E. Melamed, and D. Offen, "Increased survival and migration of engrafted mesenchymal bone marrow stem cells in 6-hydroxydopamine-lesioned rodents," *Neuroscience Letters*, vol. 395, no. 2, pp. 124–128, 2006.

[18] N. K. Venkataramana, S. K. V. Kumar, S. Balaraju et al., "Open-labeled study of unilateral autologous bone-marrow-derived mesenchymal stem cell transplantation in Parkinson's disease," *Translational Research*, vol. 155, no. 2, pp. 62–70, 2010.

[19] X. Zeng and L. A. Couture, "Pluripotent stem cells for Parkinson's disease: progress and challenges," *Stem Cell Research and Therapy*, vol. 4, no. 2, article 25, 2013.

[20] D. M. Gash, Z. Zhang, A. Ovadia et al., "Functional recovery in parkinsonian monkeys treated with GDNF," *Nature*, vol. 380, no. 6571, pp. 252–255, 1996.

[21] A. D. Ebert, A. J. Beres, A. E. Barber, and C. N. Svendsen, "Human neural progenitor cells over-expressing IGF-1 protect dopamine neurons and restore function in a rat model of Parkinson's disease," *Experimental Neurology*, vol. 209, no. 1, pp. 213–223, 2008.

[22] A. Glavaski-Joksimovic, T. Virag, Q. A. Chang et al., "Reversal of dopaminergic degeneration in a parkinsonian rat following micrografting of human bone marrow-derived neural progenitors," *Cell Transplantation*, vol. 18, no. 7, pp. 804–814, 2009.

[23] F. Lan, J. Xu, X. Zhang et al., "Hepatocyte growth factor promotes proliferation and migration in immortalized progenitor cells," *NeuroReport*, vol. 19, no. 7, pp. 765–769, 2008.

[24] H. Koike, A. Ishida, M. Shimamura et al., "Prevention of onset of Parkinson's disease by in vivo gene transfer of human hepatocyte growth factor in rodent model: a model of gene therapy for Parkinson's disease," *Gene Therapy*, vol. 13, no. 23, pp. 1639–1644, 2006.

[25] H. Duan, C. Wu, D. Wu et al., "Treatment of myocardial ischemia with bone marrow-derived mesenchymal stem cells overexpressing hepatocyte growth factor," *Molecular Therapy*, vol. 8, no. 3, pp. 467–474, 2003.

[26] S. R. Subramaniam and M. Chesselet, "Mitochondrial dysfunction and oxidative stress in Parkinson's disease," *Progress in Neurobiology*, vol. 106-107, pp. 17–32, 2013.

[27] D. J. Surmeier and P. T. Schumacker, "Calcium, bioenergetics, and neuronal vulnerability in Parkinson's disease," *The Journal of Biological Chemistry*, vol. 288, no. 15, pp. 10736–10741, 2013.

[28] M. K. Sanghera, J.-L. Zamora, and D. C. German, "Calbindin-D(28k)-containing neurons in the human hypothalamus: relationship to dopaminergic neurons," *Neurodegeneration*, vol. 4, no. 4, pp. 375–381, 1995.

[29] K. C. Luu, G. Y. Nie, A. Hampton, G. Fu, Y. Liu, and L. A. Salamonsen, "Endometrial expression of calbindin (CaBP)-d28k but not CaBP-d9k in primates implies evolutionary changes and functional redundancy of calbindins at implantation," *Reproduction*, vol. 128, no. 4, pp. 433–441, 2004.

[30] A. McMahon, B. S. Wong, A. M. Iacopino, M. C. Ng, S. Chi, and D. C. German, "Calbindin-D_{28k} buffers intracellular calcium and promotes resistance to degeneration in PC12 cells," *Molecular Brain Research*, vol. 54, no. 1, pp. 56–63, 1998.

[31] S. Sun, F. Li, X. Gao et al., "Calbindin-D28K inhibits apoptosis in dopaminergic neurons by activation of the PI3-kinase-Akt signaling pathway," *Neuroscience*, vol. 199, pp. 359–367, 2011.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 106252, 7 pages
http://dx.doi.org/10.1155/2014/106252

Research Article

Relationship between CCR and NT-proBNP in Chinese HF Patients, and Their Correlations with Severity of HF

Zhigang Lu,[1] Bo Wang,[2] Yunliang Wang,[3] Xueqing Qian,[4] Wei Zheng,[5] and Meng Wei[1]

[1] *Department of Cardiology, The 6th People's Hospital Affiliated to Shanghai Jiaotong University Medical College, No. 600 Yishan Road, Shanghai, China*
[2] *Department of Clinical Laboratory, Dalian Municipal Central Hospital, Dalian, China*
[3] *Department of Neurology, The 148th Hospital, Zibo, Shandong, China*
[4] *Medlogic Healthcare Technology Company Ltd., Shanghai, China*
[5] *Laboratory of Medicine, General Hospital of Shenyang Military Area Command, Shenyang, China*

Correspondence should be addressed to Bo Wang; wangbo@163.com and Meng Wei; 13801661299@163.com

Received 11 July 2014; Accepted 15 August 2014; Published 28 August 2014

Academic Editor: Hongwei Wang

Aim. To evaluate the relationship between creatinine clearance rate (CCR) and the level of N-terminal pro-B-type natriuretic peptide (NT-proBNP) in heart failure (HF) patients and their correlations with HF severity. *Methods and Results*. Two hundred and one Chinese patients were grouped according to the New York Heart Association (NYHA) classification as NYHA 1-2 and 3-4 groups and 135 cases out of heart failure patients as control group. The following variables were compared among these three groups: age, sex, body mass index (BMI), smoking status, hypertension, diabetes, NT-proBNP, creatinine (Cr), uric acid (UA), left ventricular end-diastolic diameter (LVEDD), and CCR. The biomarkers of NT-proBNP, Cr, UA, LVEDD, and CCR varied significantly in the three groups, and these variables were positively correlated with the NHYA classification. The levels of NT-proBNP and CCR were closely related to the occurrence of HF and were independent risk factors for HF. At the same time, there was a significant negative correlation between the levels of NT-proBNP and CCR. The area under the receiver operating characteristic curve suggested that the NT-proBNP and CCR have high accuracy for diagnosis of HF and have clinical diagnostic value. *Conclusion*. NT-proBNP and CCR may be important biomarkers in evaluating the severity of HF.

1. Introduction

Chronic heart failure (CHF) is a disorder associated with high mortality and prolonged hospitalization; it affects more than 10 million people in the countries represented by the European Society of Cardiology [1]. With the development of the society and the increase in the aging population, the prevalence of hypertension, diabetes, and myocardial infarction (MI) is significantly higher than before, thus increasing the incidence of CHF [2]. Over the last decades, despite advances in treatment, the number of CHF deaths has increased steadily. Approximately 20% of deaths are reported per year due to CHF [3]. Heart failure not only declines heart pump function, but also produces the change of complex molecules, neuroendocrine, and inflammation immune system and makes many biomarkers at different stages including neurohormonal markers, inflammatory markers, markers of oxidative stress and myocardial injury, and remodeling markers. The assessment of these biomarkers, alone or in combination, may be useful in early diagnosis, differential diagnosis, prognosis, guiding treatment, and risk stratification for the heart failure patients. In recent years, with the emergence of the B-type natriuretic peptide (BNP) and clinical research, there is more focus on the development of biomarkers in heart failure [4].

The natriuretic peptide family mainly includes A-type natriuretic peptide (ANP) or atrial natriuretic peptide, B-type natriuretic peptide (BNP) or brain natriuretic peptide, C-type natriuretic peptide (CNP), renal natriuretic peptide, and Dendroaspis natriuretic peptide (DNP). The ANP and BNP are mainly secreted by the atrium and ventricle, respectively, and have similar effects, which are natriuretic and can inhibit

renin angiotensin aldosterone system, and BNP is useful as a principal biomarker for CHF [5–9]. Since the level of BNP increases in heart failure, elevated plasma BNP concentration is used as a marker of heart failure. In recent years, like BNP, the NT-proBNP also was identified as a novel and important biomarker in heart failure to determine the severity of heart failure [10–12].

Heart failure and renal dysfunction are closely related. The ventricular dysfunction caused by CHF may lead to a series of adaptive responses, such as the activation of neuroendocrine system, peripheral vasoconstriction, and reduced renal perfusion pressure. All these changes can cause renal dysfunction, and the deterioration of renal function further increases the capacity of the load of the heart and leads to a vicious circle. The natriuretic peptide system can be activated in both heart failure and severe renal insufficiency patients. Renal dysfunction is often present in CHF patients with reported CCR lower than 60 mL/min in up to 50% of patients [13, 14].

Several studies have revealed that there is a relationship between NT-proBNP levels and clinical manifestations [4, 15, 16]. However, it remains unknown whether CCR and other biomarkers are correlated with the severity of heart failure. In this study, we aimed to determine whether the plasma levels of NT-proBNP, CCR, and other biomarkers altered with changes in the severity of heart failure and whether these markers are appropriate in immediately identifying symptomatic or asymptomatic heart failure in patients. While few similar studies have been conducted, this remains the first report in the Chinese population.

2. Materials and Methods

2.1. Ethics Statement. The investigation complied with the principles outlined in the Declaration of Helsinki [17]. The present cross-sectional study was performed in patients of outpatient setting. The study was approved by the Ethics Committee of the Hospital, Shanghai, Dalian, and Shenyang, China. Verbal informed consent was obtained from all patients. Each consent was recorded in the sample collection processing records and this consent procedure was approved by the Ethics Committee.

2.2. Patients. Three hundred thirty-six consecutive symptomatic or nonsymptomatic Chinese heart failure patients for suspected myocardial ischemia scheduled for coronary angiography were recruited between July 2011 and October 2012 at the 6th People's Hospital affiliated to Shanghai Jiaotong University Medical College, Dalian Municipal Central Hospital Affiliated of Dalian Medical University, and General Hospital of Shenyang Military Area Command, China. Severity of CHF was clinically evaluated according to the NYHA classification.

Two hundred and one Chinese patients were grouped according to the New York Heart Association (NYHA) classification as NYHA 1-2 and 3-4 groups and 135 cases out of heart failure patients as control group. Patients in NYHA class 1 showed cardiac disease but result in no limitation of physical activity. Ordinary physical activity does not cause undue fatigue, palpitation, dyspnea, or anginal pain. No cardiac disease, hypertension, or diabetes was diagnosed in control group.

A fasting venous blood sample was obtained for measurement of fasting glucose and HbA1c. Patients with HbA1c levels ≥6.5% were diagnosed as diabetic, even without previous history of diabetes. Body weight and height were measured to determine the body mass index (BMI; BMI, kg/m^2 = weight (kg)/[height (m)]^2) and blood pressure by standard methods.

2.3. Statistical Analysis. Data were expressed as mean ± standard deviation (SD) for continuous variables, or as percentages (%) for categorical variables. Statistical analysis was performed using SPSS for Windows (version 13.0). Variables such as age, sex, and smoking status were adjusted by covariance analysis. Repeated measures one-way ANOVA was used to determine the significance of trends within groups, and further comparison between the two groups was done using least significant difference procedure (LSD). Numerical data was compared using the Chi-square test. Spearman single-factor correlation analysis and Spearman coefficient of rank correlation were used to evaluate linear relationship between biomarkers and NYHA classification. Logistic regression was used to identify independent risk factors for heart failure. Association between variables in NT-proBNP and CCR was examined using Pearson's correlation coefficient and linear regression. Receiver operating characteristic (ROC) curves were used to obtain the biomarker cut-off points for predicting the prevalence of angiographic heart failure. The respective areas under the curve (AUC), sensitivity, and specificity were compared between biomarkers and NYHA classification. A value of P less than 0.05 was considered statistically significant.

2.4. Laboratory Assays. Fasting plasma glucose (FPG) was quantified by the glucose oxidase procedure and HbA1c was measured by ion-exchange high-performance liquid chromatography (HPLC, Bio-Rad, USA). Creatinine (Cr) and uric acid (UA) were measured by an enzymatic method with a chemical analyzer (Hitachi 7600-020, Tokyo, Japan); CCR was calculated using the Cockcroft-Gault formula. The chemiluminescence-based immunoanalytical system was used to determine plasma levels of NT-proBNP (VITROS 5600 integrated system, Johnson & Johnson Medical Company, USA).

2.5. Echocardiography. Two-dimensional and Doppler echocardiography scans were performed using the HP77020A echocardiograph (Hewlett Packard Company, USA) to assess the left ventricular end-diastolic diameter (LVEDD).

3. Results

3.1. Baseline Characteristics. The baseline characteristics of the patients are shown in Table 1. Of the 336 patients, 135 patients (40.18%, mean age = 65.84 ± 15.95 years, male = 47.40%) were in control group, 79 patients (23.51%, mean age = 69.91 ± 13.14 years, male = 54.43%) were in NYHA class 1-2, and 122 patients (36.31%, mean age = 68.27 ± 13.36 years, male = 63.93%) were in NYHA class 3-4.

Table 1: Characteristics of the control, NYHA classification 1-2, and 3-4 groups.

Groups	Control ($n = 135$)	1-2 ($n = 79$)	3-4 ($n = 122$)	F value	P value
Age, year (mean ± SD)	65.84 ± 15.95	69.91 ± 13.14	68.27 ± 13.36	2.14	**0.12**
Male, % (n)	47.40 (64/135)	54.43 (43/79)	63.93 (78/122)	3.59	**0.03**
BMI, kg/m^2 (mean ± SD)	23.48 ± 3.69	24.19 ± 4.31	24.05 ± 4.22	1.13	**0.33**
Current smokers, % (n)	27.40 (37/135)	29.11 (23/79)	17.21 (21/122)	0.55	**0.58**
Hypertension, % (n)	49.63 (67/135)	75.94 (60/79)	71.31 (87/122)	10.39	**0.00**
Diabetes, % (n)	16.30 (22/135)	30.38 (24/79)	38.52 (47/122)	8.43	**0.00**
NT-proBNP, pg/mL (mean ± SD)	78.83 ± 15.27	1611.54 ± 171.24	3162.19 ± 453.21	260.18	**0.00**
CCR, mL/min (mean ± SD)	89.94 ± 16.39	59.43 ± 19.57	53.57 ± 17.41	154.87	**0.00**
Cr, mg/dL (mean ± SD)	0.67 ± 0.15	1.07 ± 0.25	1.16 ± 0.23	195.55	**0.00**
UA, umol/L (mean ± SD)	299.22 ± 56.12	418.91 ± 49.21	471.54 ± 64.72	88.80	**0.00**
LVEDD, mm (mean ± SD)	45.21 ± 3.86	50.89 ± 6.65	52.85 ± 8.71	45.40	**0.00**

BMI: body mass index; CCR: creatinine clearance rate; Cr: creatinine; LVEDD: left ventricular end-diastolic diameter; NT-proBNP: N-terminal pro-B-type natriuretic peptide; NYHA: New York Heart Association; UA: uric acid.

There were no significant differences in age, BMI, and current smokers between any of heart failure groups and the control group ($P > 0.05$); the variables of male, hypertension, diabetes, NT-proBNP, CCR, Cr, UA, and LVEDD were significantly different among the three groups (all $P < 0.01$, excluded variable of male $P < 0.05$).

3.2. Comparison of NT-proBNP, CCR, Cr, UA, and LVEDD between the Heart Failure Groups and Control Group. The NT-proBNP, Cr, UA, and LVEDD levels were significantly higher in the NYHA class 1-2 and 3-4 groups than in the control group, and these variables in the NYHA class 3-4 group were significantly higher than that in the control and NYHA class 1-2 groups (all $P < 0.01$, excluded LVEDD level between NYHA class 1-2 and 3-4 groups, $P < 0.05$). The NT-proBNP level increased from control group (78.83 ± 15.27) to NYHA class 1-2 group (1611.54 ± 171.24) to class 3-4 group (3162.19 ± 453.21). The value for CCR significantly decreased from control group (89.94 ± 16.39) to NYHA class 1-2 group (59.43 ± 19.57) and class 3-4 group (53.57 ± 17.41) ($P < 0.01$). The patients with history of hypertension and diabetes were higher in the NYHA class 1-2 and 3-4 groups than in the control group ($P < 0.05$), and there was no significant difference between NYHA class 3-4 and 1-2 groups (Table 2).

3.3. Correlations Analysis of Individual Biomarkers with NYHA Classification in Heart Failure Groups. As shown in Table 3, the coefficient of rank correlation for NT-proBNP was 0.87, CCR was 0.74, Cr was 0.69, LVEDD was 0.44, and UA was 0.64, with $P = 0.00$. The variables of NT-proBNP, CCR, Cr, LVEDD, and UA showed positive correlation with the NHYA classification. With NHYA classification as dependent variable ($y = 1, n = 0$) and age, male, BMI, current smokers, hypertension, diabetes, NT-proBNP, CCR, Cr, UA, and LVEDD as independent variables, the results showed that NT-proBNP and CCR were independent risk factors for heart failure (Table 4).

The Pearson correlation analysis was carried out to determine the relationship between variables of NT-proBNP and CCR in control and heart failure groups. Table 5 shows that there was a significant negative correlation between the levels of NT-proBNP and CCR ($r = -0.62$, $P = 0.00$).

In univariate linear regression analysis, CCR showed a significant negative correlation with NT-proBNP in the control and heart failure groups ($r = -0.62$, $P = 0.00$, Figure 2). This indicates that with the elevated NT-proBNP levels, CCR gradually reduced in the control group to NYHA class 1-2 to class 3-4 group.

3.4. Diagnostic Power of NT-proBNP and CCR for Heart Failure. The ROC curves for NT-proBNP and CCR as indicators of heart failure are shown in Figure 1. The area under the ROC curve was higher for NT-proBNP (NHYA 1-2: 0.896; NHYA 3-4: 0.922) than for CCR (NHYA 1-2: 0.860; NHYA 3-4: 0.882). These results suggested that the NT-proBNP and CCR have high accuracy for diagnosis of heart failure and have clinical diagnostic value. The respective cut-off points for diagnosis of heart failure were estimated according to the ROC curves for NT-proBNP and CCR. With a cut-off value of 329.05 pg/mL, NT-proBNP had a sensitivity of 81.07% and a specificity of 75.62% for predicting NHYA 1-2, and a cut-off value of 324.40 pg/mL had a sensitivity of 82.34% and a specificity of 85.90% for predicting NHYA 3-4. Similarly, a cut-off value for CCR of 61.39 mL/min had a sensitivity of 74.71% and a specificity of 63.02% for predicting NHYA 1-2, and a cut-off value of 63.13 mL/min had a sensitivity of 82.42% and a specificity of 78.33% for predicting NHYA 3-4. Using these cut-off points, NT-proBNP showed higher sensitivity and specificity than CCR (Table 6).

4. Discussion

In this study, we observed that the variables, such as male, hypertension, diabetes, NT-pro BNP, CCR, Cr, UA, and LVEDD, were significantly different among all groups. Furthermore, we revealed that the biomarkers of NT-proBNP, Cr, UA, LVEDD, and CCR were positively correlated with the severity of heart failure.

Table 2: Paired comparison for biology markers between control group and heart failure group.

Variable	NYHA		SE	P value	95% CI	
					Lower	Upper
Male	0	2	0.07	0.32	−0.21	0.07
		3	0.06	0.01	−0.29	−0.04
	2	3	0.07	0.18	−2.24	0.05
Hypertension	0	2	0.07	0.00	−0.39	−0.13
		3	0.06	0.00	−0.33	−0.10
	2	3	0.07	0.49	−0.09	0.18
Diabetes	0	2	0.06	0.02	−0.26	−0.02
		3	0.05	0.00	−0.33	−0.11
	2	3	0.06	0.20	−0.21	0.04
NT-proBNP	0	2	153.31	0.00	−1834.29	−1231.12
		3	135.20	0.00	−3349.31	−2817.41
	2	3	156.30	0.00	−1858.11	−1243.20
CCR	0	2	2.49	0.00	25.62	35.40
		3	2.19	0.00	32.06	40.69
	2	3	2.53	0.02	0.88	10.85
Cr	0	2	0.03	0.00	−0.46	−0.35
		3	0.04	0.00	−0.54	−0.44
	2	3	0.03	0.01	−0.14	−0.02
UA	0	2	14.99	0.00	−149.18	−90.20
		3	13.22	0.00	−198.33	−146.31
	2	3	15.30	0.00	−82.70	−22.56
LVEDD	0	2	0.94	0.00	−7.52	−3.82
		3	0.83	0.00	−9.27	−6.01
	2	3	0.96	0.04	−3.85	−2.76

CCR: creatinine clearance rate; CI: confidence interval; Cr: creatinine; LVEDD: left ventricular end-diastolic diameter; NT-proBNP: N-terminal pro-B-type natriuretic peptide; NYHA: New York Heart Association; SE: standard error; UA: uric acid.

Table 3: Spearman correlation analysis of relations between variables and the NYHA classification (n = 336).

Variable	Correlation coefficient	Sig (2-tailed)
NT-proBNP	0.87	0.00
CCR	0.74	0.00
Cr	0.69	0.00
LVEDD	0.44	0.00
UA	0.64	0.00

CCR: creatinine clearance rate; Cr: creatinine; LVEDD: left ventricular end-diastolic diameter; NT-proBNP: N-terminal pro-B-type natriuretic peptide; UA: uric acid.

The NT-proBNP level significantly increased and the value for CCR significantly decreased from control group to NYHA class 1-2 to 3-4 group. The levels of NT-proBNP and CCR were closely related to heart failure and were independent risk factors for patients with heart failure. At the same time, there was a significant negative correlation between the level of NT-proBNP and CCR. The area under the ROC curve suggested that the NT-proBNP and CCR have high accuracy in the diagnosis of heart failure with clinical diagnostic value.

Our findings are similar to the results of several previous studies where NT-proBNP plasma levels were closely related to the severity of heart failure [18]. Furthermore, in our study, the mean levels of NT-proBNP in the NYHA class 1-2 and 3-4 groups were greater than the cut-off points for diagnosis of heart failure, indicating that severity of heart failure increased gradually from control group to class 1-2 and 3-4 groups.

Heart failure and renal dysfunction often coexist as the visceral damage in one organ will result in the other organ's pathological changes accordingly. As two most important organs in the body, heart and kidney influence each other in the physiological and pathological processes, and the renal blood flow accounts for 20–25% of the total output of heart and plays an important role in regulating blood volume, blood vessels tension, and blood pressure change. In patients with heart failure, moderately elevated serum creatinine, without a history of chronic renal insufficiency, is often noticed. Therefore, our research focused on heart failure with no history of chronic kidney disease patients to observe the correlation between kidney index and cardiac function.

Determination of endogenous CCR can effectively evaluate the glomerular filtration function. The CCR can determine the degree of renal impairment and whether glomerular filtration function was damaged. Previous studies have shown

Table 4: Logistic regression analysis of risk factors for heart failure.

Variable	Regression coefficient	wald	χ^2 value	P value	Correct class
NT-proBNP	0.03	4.52	265.55	0.03	87.50%
CCR	20.82	6.08	441.48	0.01	99.70%

CCR: creatinine clearance rate; NT-proBNP: N-terminal pro-B-type natriuretic peptide.

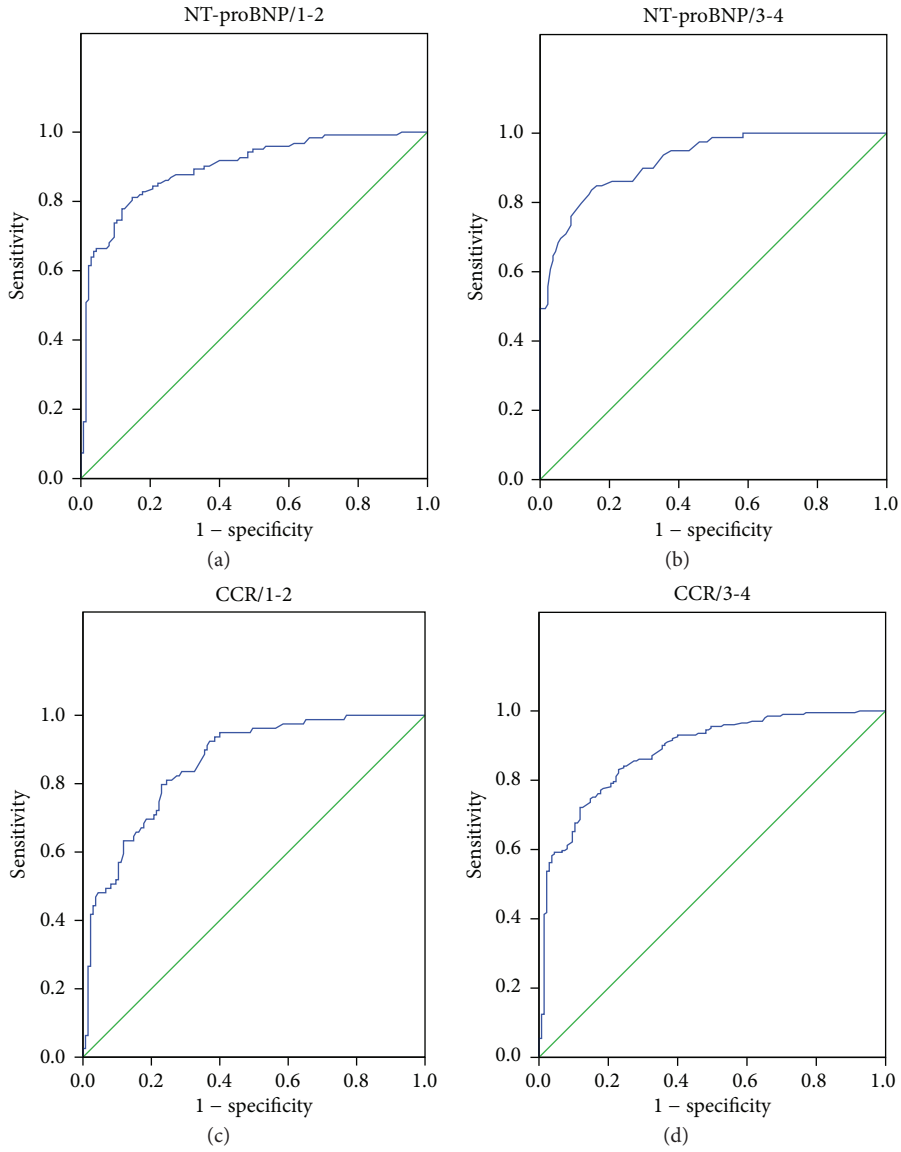


Figure 1: Receiver operating characteristic curve analysis of NT-proBNP and CCR for diagnosis for heart failure. ROC curve shows NT-proBNP for the prediction of heart failure patients with NHYA class: (a) 1-2 and (b) 3-4 groups; ROC curve shows CCR for the prediction of heart failure patients with NHYA class: (c) 1-2 and (d) 3-4 groups.

Table 5: Pearson correlation analysis for NT-proBNP and CCR (n = 336).

Variable	NT-proBNP	CCR
NT-proBNP		
Correlation coefficient	1.00	−0.62
Sig (2-tailed)		0.00
CCR		
Correlation coefficient	−0.62	1.00
Sig (2-tailed)	0.00	

CCR: creatinine clearance rate; NT-proBNP: N-terminal pro-B-type natriuretic peptide.

that renal insufficiency is the risk factor for prognosis of patients with myocardial infarction, cardiac insufficiency, and hypertension [19–21].

Uric acid, a product of purine metabolism whose elevated concentration in CHF is a sign of damaged oxygen metabolism, is associated with the severity of cardiac dysfunction [22]. There is a correlation between uric acid and the existing cardiovascular disease risk factors such as hypertension, diabetes, hyperlipidemia, and obesity [23]. Increased blood UA levels in patients with CHF may be because of excretion of UA occurring mainly through

Table 6: Cut-off points, sensitivity, specificity, and area under the curves for biomarkers and NYHA classification.

Marker	NYHA classification	Cut-off point	Sensitivity	Specificity	Area under ROC curve	SE	*P* value	95% CI	
								Lower	Upper
NT-proBNP	1-2	329.05 pg/mL	81.07%	75.62%	0.90	0.02	0.00	0.86	0.94
	3-4	324.40 pg/mL	82.34%	85.90%	0.92	0.02	0.00	0.89	0.96
CCR	1-2	61.39 mL/min	74.71%	63.02%	0.86	0.03	0.00	0.81	0.91
	3-4	63.13 mL/min	82.42%	78.33%	0.88	0.02	0.00	0.85	0.92

CCR: creatinine clearance rate; CI: confidence interval; NT-proBNP: N-terminal pro-B-type natriuretic peptide; NYHA: New York Heart Association; ROC: receiver operating characteristic; SE: standard error.

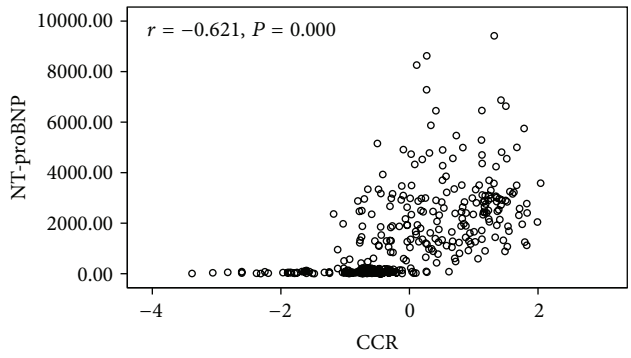


Figure 2: Linear regression analysis of the level of NT-proBNP with CCR ($r = -0.621$, $P = 0.00$, $n = 336$).

kidneys, hypoxemia, increased anaerobic metabolism, and activation of xanthine oxidase. With the severity of heart failure, rate of anaerobic metabolism and quantity of lactic acid increase and the excretion of UA and lactate compete for the anion channel in the proximal convoluted tubule. The excretion of UA is reduced and results in the increased blood UA concentration. The cardiac output quantity is significantly reduced in the patients with heart failure, which results in decreased renal blood flow, damaged kidney, reduced glomerular filtration rate, decreased excretion of UA, and the increased level of UA [24].

Left ventricular end-diastolic diameter is used to determine the abnormal changes of cardiac systolic function. Cardiac ischemia causes the interruption of coronary blood flow and decline of myocardial contraction ability. Cardiac contraction ability is closely related to the size of myocardial ischemic area. When ischemic area size exceeds 15%, LVEDD value increases. There is evidence to suggest that NT-proBNP levels may reflect increased left ventricular wall stress in the absence of cardiac ischemia; thus, in this study we observed that the LVEDD increased from control to NYHA class 1-2 to 3-4 group, associated with heart failure severity.

Conflict of Interests

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

Authors' Contribution

Zhigang Lu and Bo Wang contributed equally to this work.

Acknowledgments

This work was supported by Johnson & Johnson Medical. Data was analyzed at the 6th People's Hospital affiliated to Shanghai Jiaotong University Medical College in Shanghai, China.

References

[1] K. Dickstein, A. Cohen-Solal, G. Filippatos et al., "ESC Committee for Practice Guidelines (CPG). ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure 2008: the Task Force for the Diagnosis and Treatment of Acute and Chronic Heart Failure 2008 of the European Society of Cardiology. Developed in collaboration with the Heart Failure Association of the ESC (HFA) and endorsed by the European Society of Intensive Care Medicine (ESICM)," *European Heart Journal*, vol. 29, pp. 2388–2442, 2008.

[2] Shanghai Investigation Group of Heart Failure, "The evolving trends in the epidemiologic factors and treatment of patients with congestive heart failure in shanghai during years of 1980, 1990, 2000," *Journal of Chinese Cardiovascular Disease*, vol. 30, pp. 24–27, 1980.

[3] D. Lloyd-Jones, R. Adams, M. Carnethon et al., "American Heart Association Statistics Committee and Stroke Statistics Subcommittee. Heart disease and stroke statistics—2009 update: a report from the American Heart Association Statistics Committee and Stroke Statistics Subcommittee," *Circulation*, vol. 119, pp. e21–e181, 2009.

[4] J. P. Goetze, C. Christoffersen, M. Perko et al., "Increased cardiac BNP expression associated with myocardial ischemia," *The FASEB Journal*, vol. 17, no. 9, pp. 1105–1107, 2003.

[5] T. Tsutamoto, A. Wada, K. Maeda et al., "Attenuation of compensation of endogenous cardiac natriuretic peptide system in chronic heart failure: prognostic role of plasma brain natriuretic peptide concentration in patients with chronic symptomatic left ventricular dysfunction," *Circulation*, vol. 96, no. 2, pp. 509–516, 1997.

[6] G. A. Haldeman, J. B. Croft, W. H. Giles, and A. Rashidee, "Hospitalization of patients with heart failure: national hospital discharge survey, 1985 to 1995," *American Heart Journal*, vol. 137, no. 2, pp. 352–360, 1999.

[7] R. S. Gardner, F. Özalp, A. J. Murday, S. D. Robb, and T. A. McDonagh, "N-terminal pro-brain natriuretic peptide: a new gold standard in predicting mortality in patients with advanced heart failure," *European Heart Journal*, vol. 24, no. 19, pp. 1735–1743, 2003.

[8] S. Masson, R. Latini, I. S. Anand et al., "Direct comparison of B-type natriuretic peptide (BNP) and amino-terminal proBNP in a large population of patients with chronic and symptomatic heart failure: the valsartan heart failure (Val-HeFT) data," *Clinical Chemistry*, vol. 52, no. 8, pp. 1528–1538, 2006.

[9] Y. Tanino, J. Shite, O. L. Paredes et al., "Whole body bioimpedance monitoring for outpatient chronic heart failure follow up," *Circulation Journal*, vol. 73, no. 6, pp. 1074–1079, 2009.

[10] P. J. Hunt, A. M. Richards, M. G. Nicholls, T. G. Yandle, R. N. Doughty, and E. A. Espiner, "Immunoreactive amino-terminal pro-brain natriuretic peptide (NT-PROBNP): a new marker of cardiac impairment," *Clinical Endocrinology*, vol. 47, no. 3, pp. 287–296, 1997.

[11] C. Fisher, C. Berry, L. Blue, J. J. Morton, and J. McMurray, "NT-proBNP predicts prognosis in patients with chronic heart failure," *Heart*, vol. 89, pp. 879–881, 2003.

[12] T. Weber, J. Auer, and B. Eber, "The diagnostic and prognostic value of brain natriuretic peptide and aminoterminal (nt)-pro brain natriuretic peptide," *Current Pharmaceutical Design*, vol. 11, no. 4, pp. 511–525, 2005.

[13] K. D. Aaronson, J. S. Schwartz, T.-M. Chen, K.-L. Wong, J. E. Goin, and D. M. Mancini, "Development and prospective validation of a clinical index to predict survival in ambulatory patients referred for cardiac transplant evaluation," *Circulation*, vol. 95, no. 12, pp. 2660–2667, 1997.

[14] J. M. Flack, J. D. Neaton, B. Daniels, and P. Esunge, "Ethnicity and renal disease: lessons from the multiple risk factor intervention trial and the treatment of mild hypertension study," *American Journal of Kidney Diseases*, vol. 21, no. 4, pp. 31–40, 1993.

[15] R. J. Elin and W. E. Winter, "Laboratory and clinical aspects of B-type natriuretic peptides," *Archives of Pathology and Laboratory Medicine*, vol. 128, no. 6, pp. 697–699, 2004.

[16] S. G. Williams, L. L. Ng, R. J. O'Brien et al., "Complementary roles of simple variables, NYHA and N-BNP, in indicating aerobic capacity and severity of heart failure," *International Journal of Cardiology*, vol. 102, no. 2, pp. 279–286, 2005.

[17] P. P. Rickham, "Human experimentation. Code of ethics of the world medical association. Declaration of Helsinki," *British Medical Journal*, vol. 2, no. 5402, p. 177, 1964.

[18] T. A. McDonagh, S. Holmer, I. Raymond, A. Luchner, P. Hildebrant, and H. J. Dargie, "NT-proBNP and the diagnosis of heart failure: a pooled analysis of three European epidemiological studies," *European Journal of Heart Failure*, vol. 6, no. 3, pp. 269–273, 2004.

[19] N. S. Anavekar, J. J. V. McMurray, E. J. Velazquez et al., "Relation between renal dysfunction and cardiovascular outcomes after myocardial infarction," *The New England Journal of Medicine*, vol. 351, no. 13, pp. 1285–1295, 2004.

[20] H. L. Hillege, A. R. J. Girbes, P. J. De Kam et al., "Renal function, neurohormonal activation, and survival in patients with chronic heart failure," *Circulation*, vol. 102, no. 2, pp. 203–210, 2000.

[21] M. Rahman, S. Pressel, B. R. Davis et al., "Cardiovascular outcomes in high-risk hypertensive patients stratified by baseline glomerular filtration rate," *Annals of Internal Medicine*, vol. 144, no. 3, pp. 172–180, 2006.

[22] F. Leyva, S. Anker, J. W. Swan et al., "Serum uric acid as an index of impaired oxidative metabolism in chronic heart failure," *European Heart Journal*, vol. 18, no. 5, pp. 858–865, 1997.

[23] A. Dobson, "Is raised serum uric acid a cause of cardiovascular disease or death?" *The Lancet*, vol. 354, no. 9190, p. 1578, 1999.

[24] S. Spiekermann, U. Landmesser, S. Dikalov et al., "Electron spin resonance characterization of vascular xanthine and NAD(P)H oxidase activity in patients with coronary artery disease: relation to endothelium-dependent vasodilation," *Circulation*, vol. 107, no. 10, pp. 1383–1389, 2003.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 708364, 10 pages
http://dx.doi.org/10.1155/2014/708364

Research Article

Characterization of Putative *cis*-Regulatory Elements in Genes Preferentially Expressed in *Arabidopsis* Male Meiocytes

Junhua Li,[1] Jinhong Yuan,[1,2] and Mingjun Li[1]

[1] *College of Life Sciences, Henan Normal University, Xinxiang, Henan 453007, China*
[2] *Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing 100081, China*

Correspondence should be addressed to Junhua Li; lijh0909@gmail.com and Mingjun Li; 041013@htu.cn

Received 9 May 2014; Revised 19 July 2014; Accepted 20 July 2014; Published 27 August 2014

Academic Editor: Hongwei Wang

Meiosis is essential for plant reproduction because it is the process during which homologous chromosome pairing, synapsis, and meiotic recombination occur. The meiotic transcriptome is difficult to investigate because of the size of meiocytes and the confines of anther lobes. The recent development of isolation techniques has enabled the characterization of transcriptional profiles in male meiocytes of *Arabidopsis*. Gene expression in male meiocytes shows unique features. The direct interaction of transcription factors (TFs) with DNA regulatory sequences forms the basis for the specificity of transcriptional regulation. Here, we identified putative *cis*-regulatory elements (CREs) associated with male meiocyte-expressed genes using *in silico* tools. The upstream regions (1 kb) of the top 50 genes preferentially expressed in *Arabidopsis* meiocytes possessed conserved motifs. These motifs are putative binding sites of TFs, some of which share common functions, such as roles in cell division. In combination with cell-type-specific analysis, our findings could be a substantial aid for the identification and experimental verification of the protein-DNA interactions for the specific TFs that drive gene expression in meiocytes.

1. Introduction

Meiosis is a special type of cell division that, after two consecutive rounds of nuclear divisions, leads to the production of haploid gametes. The processes of homologous chromosome pairing, synapsis, and meiotic recombination all occur during meiosis. Meiotic recombination is essential for plant reproduction and breeding because it ensures equal segregation and genetic exchange between homologous chromosomes [1–4]. The male meiocytes of *Arabidopsis* occupy only a small fraction of the anther tissue and are surrounded by somatic anther lobes [5]. An effective meiocyte collection method was established only recently; this development has enabled investigations of the meiotic transcriptome [5, 6]. Genome-wide gene expression analysis revealed unique transcriptome landscapes during male meiosis [5, 6].

Gene expression in eukaryotic cells is regulated by transcription factors (TFs). There are around 2000 TFs in the *Arabidopsis* genome [7], and interactions of the DNA-binding domains of TFs with specific *cis*-regulatory elements (CREs) can activate the expression of several to many thousands of target genes. The transcriptional domains of regulatory genes are critically important in many developmental processes [8]. Meiosis operates in a highly specified cell cluster and thus requires precise spatial and temporal control [3]. In *Arabidopsis*, the expression of many meiotic genes such as *AtDMC1* [9, 10], *SDS* [11], *MMD1* [12], and *RCK* [13] is highly regulated. Studying the commonness and distribution of CREs in the promoters of coexpressed genes can help facilitate the identification of signaling networks in specific cell types (e.g., [14–17]). For example, CREs or promoter motifs have been investigated in sperm cells (mature pollen) of both rice and *Arabidopsis* [18, 19].

Transcriptome profiling experiments have shown that more than 1,000 genes were preferentially expressed in meiocytes [5]; a high proportion of the promoters of such preferentially expressed genes were sufficient to drive green fluorescent protein (GFP) reporter activity in meiocytes [20]. These preliminary studies laid a substantial foundation that has enabled the mining and the examination of the common structures of meiotically active promoters. In this study, the sequences of 50 meiotically active promoters were analyzed.

The putative CREs in these promoters were identified; these CREs may be responsible for the high activity of these promoters in male meiocytes.

2. Materials and Methods

We selected candidate genes from data generated in a previous mRNA deep-sequencing study of meiosis-specific genes in *Arabidopsis* [5]. These included the most highly expressed genes in male meiocytes. In a list with genes that had ≥4 times higher expression in meiocytes than in anthers, top 50 genes in the meiocytes to seedling comparison list were chosen with exclusion of transposable element genes. The difference in expression between meiocytes and anther, the difference in expression between meiocytes and seedlings, the annotated function, and the GO (gene ontology) functional categorization of the 50 top genes are presented in Supplemental File 1, available online at http://dx.doi.org/10.1155/2014/708364. As a negative control, 50 genes randomly selected from an Affymetrix ATH1 microarray experiment deposited in the NASC database were analyzed [21]; see Supplemental File 2 for descriptions of these control genes.

One Kb of upstream sequences relative to the transcription start sites were retrieved using Regulatory Sequence Analysis Tools (RSAT, http://rsat.ulb.ac.be/rsat/) [22]. Analysis of known CREs was initially performed using SIGNALSCAN program in plant *cis*-acting regulatory DNA elements (PLACE, http://www.dna.affrc.go.jp/PLACE/) [23, 24]. Analysis of statistically overrepresented elements was conducted by Pscan (http://159.149.160.51/pscan/) [25]. In the Pscan window, TAIR gene identifiers of the 50 genes were submitted, the source organism was specified as *Arabidopsis thaliana*, and the region to be analyzed was from −1000 to +0 with regard to the annotated transcription start site. For assessing the significance of the results, the P values were computed by Pscan with a z-test, a test that associated with each profile the probability of obtaining the same score on a random sequence set [25]. An element is considered to be significantly overrepresented if the P value is less than 0.01. Additional analysis for unknown novel motifs was conducted by Promzea (http://promzea.org) [26]. 1000 bp long promoter regions were analyzed and each predicted motif was provided with a mean normalized conditional probability (MNCP); a MNCP score greater than 1 indicates that the motif is more represented in the input data set compared to a random set of promoters/first introns [26]. Motifs predicted by Promzea were compared with experimentally defined motifs in the PLACE database using STAMP [27]. Strand bias analysis of putative CREs was performed using Athamap (http://www.athamap.de/) [28–32], −1000 to 0 regions relative to the transcription start site were analyzed, and the total strand distribution of CREs was the sum of the individual CRE numbers in each promoter in the "overview" search result.

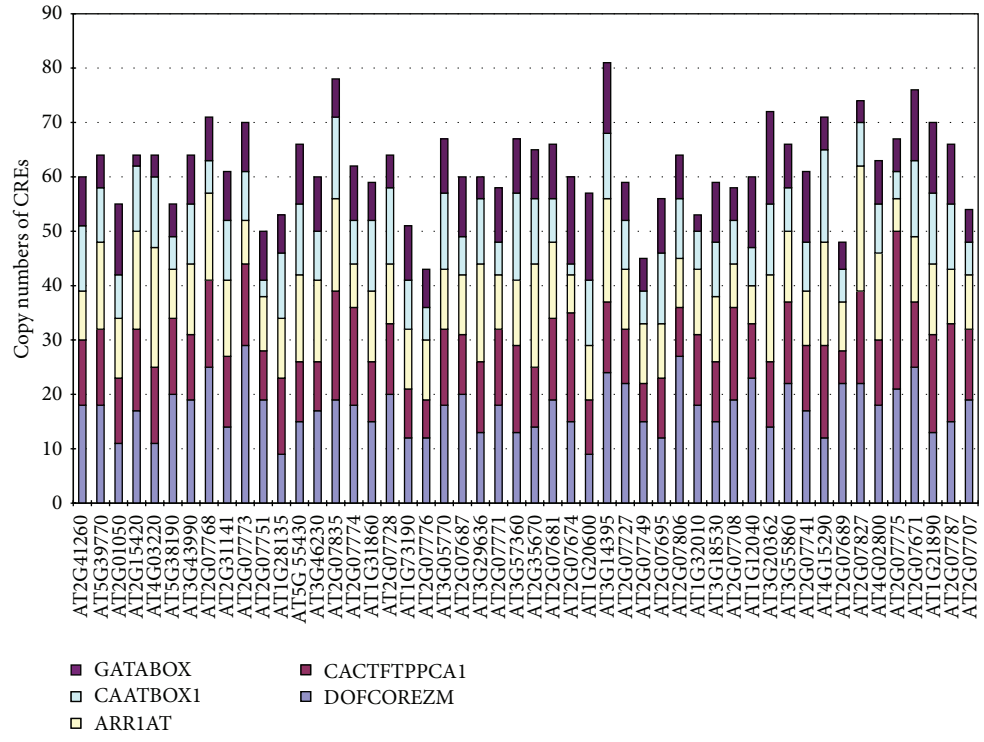
3. Results and Discussions

Putative 1000 bp promoter regions were selected and their CREs were analyzed by the use of the PLACE collection. Five CREs were found in all 50 promoters: DOFCOREZM (5′-AAAG-3′), CACTFTPPCA1 (5′-YACT-3′, Y=T/C), ARR1AT (5′-NGATT-3′, N=G/A/C/T), CAATBOX1 (5′-CAAT-3′), and GATABOX (5′-GATA-3′). The frequencies and distributions of these CREs in each promoter are shown in Figure 1(a).

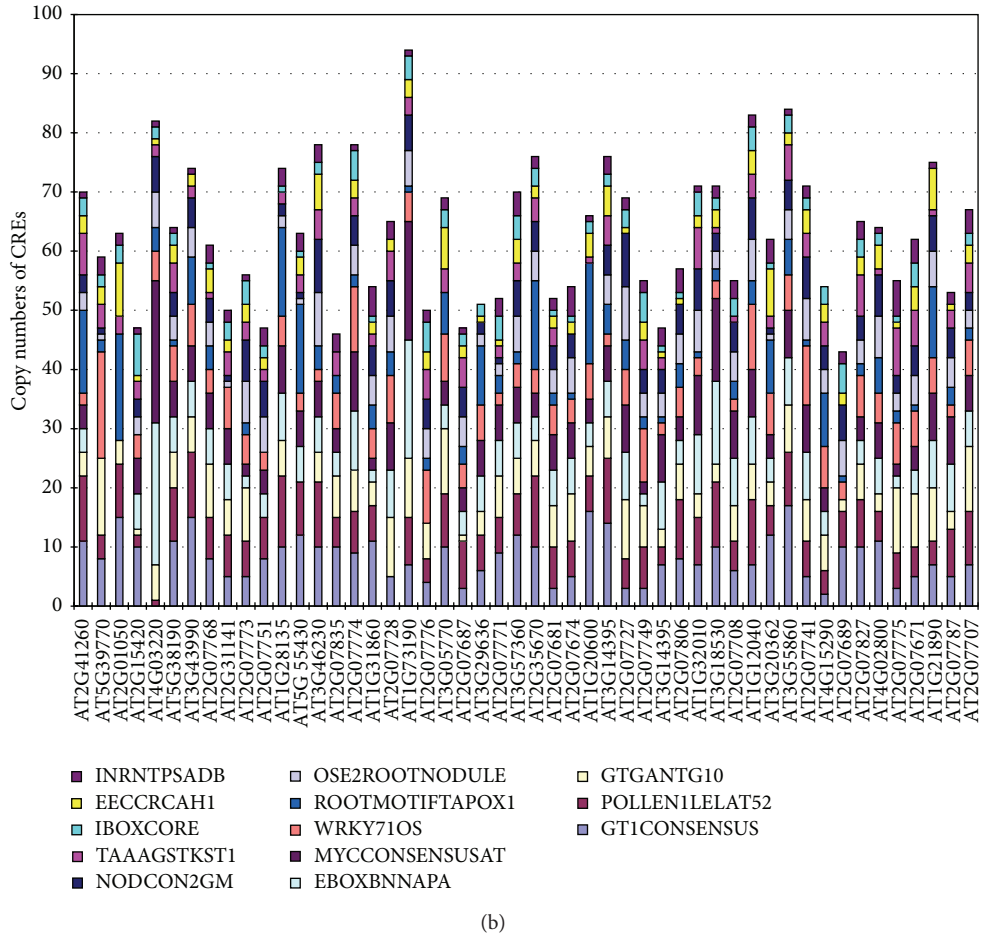
DOFCOREZM was the most abundant CRE in the 50 putative promoter sequences. It is a core site for the binding of Dof proteins in maize. The Dof proteins are a family of plant-specific TFs that includes Dof1, Dof2, Dof3, and PBF [33, 34]. Maize Dof1 was suggested to be a regulator of the expression of the C4 photosynthetic phosphoenolpyruvate carboxylase (*C4PEPC*) gene [35]. Dof1 also enhances transcription of the cytosolic orthophosphate dikinase (*cyPPDK*) genes and the nonphotosynthetic *PEPC* gene [33]. Maize Dof2 suppresses the promoter of *C4PEPC* [35]; PBF is an endosperm-specific Dof protein that binds to the prolamin box of a native B-hordein promoter in barley endosperm [36]. CACTFTPPCA1 is a key component of *Mem1* (*mesophyll expression module 1*) and is found in the distal promoter region of the C4 isoform of phosphoenolpyruvate carboxylase (*ppcA1*) in the C4 dicot *Flaveria trinervia*; it determines the mesophyll-specific expression of *ppcA1* [37]. ARR1AT is the binding element of ARR1 found in *Arabidopsis*. ARR1 is a response regulator [38]. CAATBOX1 is responsible for the tissue specific promoter activity of a pea legumin gene [39]. GATABOX is required for light-dependent and nitrate-dependent control of transcription in plants [40]. The GATA motif has been found in the promoter ofthe *Cab22* gene that encodes the *Petunia* chlorophyll a/b binding protein; this motif is the specific binding site of ASF-2 [41].

In addition to the five CREs that were found in all 50 promoters, there are 13 CREs that were found in at least 80% of the promoters (Figure 1(b)). These include GT1CONSENSUS (5′-GRWAAW-3′, R=A/G, W=A/T), POLLEN1LELAT52 (5′-AGAAA-3′), GTGANTG10 (5′-GTGA-3′), EBOXBNNAPA (5′-CANNTG-3′, N=G/A/C/T), MYCCONSENSUSAT (5′-CANNTG-3′, N=G/A/C/T), WRKY71OS (5′-TGAC-3′), ROOTMOTIFTAPOX1 (5′-ATATT-3′), OSE2ROOTNODULE (5′-CTCTT-3′), NODCON2GM (5′-CTCTT-3′), TAAAGSTKST1 (5′-TAAAG-3′), IBOXCORE (5′-GATAA-3′), EECCRCAH1 (5′-GANTTNC-3′, N=G/A/C/T), and INRNTPSADB (5′-YTCANTYY-3′, Y=T/C, N=G/A/C/T). Among these, seven elements are found in genes specifically expressed in particular organ. POLLEN1LELAT52 is one of two codependent regulatory elements responsible for pollen specific activation of tomato (*Lycopersicon esculentum*) *LAT52* gene [42]. GTGANTG10 is found in the promoter of the tobacco late pollen gene *g10* [43]. EBOXBNNAPA is a motif associated with storage proteins [44]. TAAAGSTKST1 is a target site in the control of guard cell-specific gene expression [45].

Six of the 13 CREs distributed in at least 80% of the examined promoters are annotated as being involved in plant responses to environmental factors, for example, GT1CONSENSUS for light and salicylic acid [46, 47], MYCCONSENSUSAT for cold [48–50], WRKY71OS for gibberellin and pathogenesis [51, 52], IBOXCORE and INRNTPSADB for light [53–55], and EECCRCAH1 for CO_2 [56, 57].



(a)



(b)

Figure 1: Distribution and occurrence of enriched PLACE motifs in the promoters of 50 genes preferentially expressed during meiosis. (a) Five common CREs found in all of the 50 promoters; (b) 13 CREs present in at least 80% of the promoters.

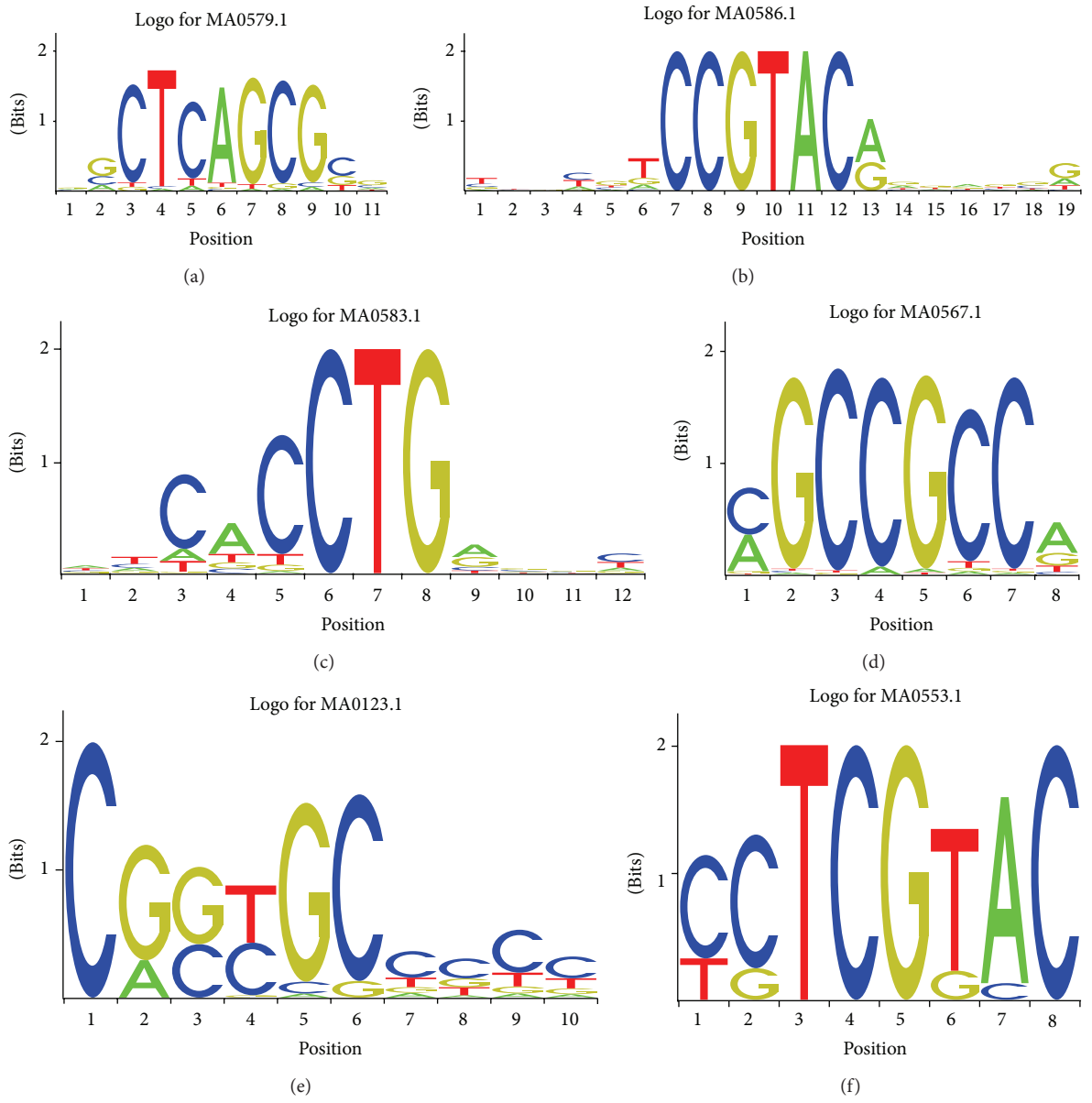


Figure 2: Sequence logos of overrepresented sequences in the promoters of genes preferentially expressed during meiosis, detected using Pscan. Letters in the logos abbreviate the nucleotides (A, C, G, and T) and are sized relative to their occurrence.

The involvement of these CREs in responses to environmental factors points to possible roles for these elements in combining signals from meiotic process and environmental factors, especially light and stress.

The aforementioned PLACE motifs represent the basic CREs required for a promoter but may not be statistically overrepresented as compared with the average level of CREs in the *Arabidopsis* genome. Among the PLACE motifs that were present in at least 80% of the promoters we examined (Figure 1), 17 of 18 are also present in rice sperm cell-specific genes [18]. The only exception to this striking similarity was the EECCRCAH1 CRE.

We further searched for motifs that were statistically overrepresented. That is, the frequency of an element in the 50 examined promoters is above the average level of the *Arabidopsis* genome. Six overrepresented putative TF binding site motifs were identified in our Pscan analysis (Figure 2 and Table 1). When we used 50 randomly selected genes (negative control) as input, only one such overrepresented motif was detected (Supplemental File 3), indicating that the meiotically active promoter sequences possess more conserved sequences.

The most significantly abundant motif detected by Pscan, CTCAGCG, is the binding sequence of *Arabidopsis* CELL DIVISION CYCLE 5 (AtCDC5), which is expressed extensively in shoot and root meristems and may function in cell cycle regulation [58, 59]. This result suggests that similar regulatory machinery functions in meiocytes and meristems

Table 1: Description of the most abundant motifs in promoters of genes preferentially expressed during meiosis, detected using Pscan.

Pscan ID	TF name	Class of the TF	Family of the TF	*P* value	References
MA0579.1	CDC5	Helix-turn-helix	Myb	$1.89231e-05$	Hirayama and Shinozaki [58]
MA0586.1	SPL14	Zinc-coordinating	SBP	$6.00715e-05$	Liang et al. [60]
MA0583.1	RAV1	EcoRII-fold	ABI3VP1	0.000477606	Kagaya et al. [61]
MA0567.1	ERF1	Beta-Hairpin-Ribbon	AP2 MBD-like	0.000997537	Godoy et al. [88]
MA0123.1	ABI4	Beta-Hairpin-Ribbon	AP2 MBD-like	0.00108575	Niu et al. [71]
MA0553.1	SMZ	AP2-ERF	AP2-ERF	0.00224012	Unpublished

and that such machinery leads to high mitotic or meiotic cell division activity. Another overrepresented motif contains the core binding motif GTAC that is recognized by the plant-specific SQUAMOSA promoter binding protein (SBP) domain transcription factor AtSPL14, which is involved in plant development and resistance to programmed cell death [60]. The binding motif of the RAV1 (RAV: for related to ABI3/VP1) DNA binding protein is overrepresented in the Pscan search results [61]; RAV1 is a regulator of plant development and is involved in plant responses to biotic and abiotic stress [62–65]. Another overrepresented motif is recognized by ERF1; a TF that belongs to the EREB/AP2 family and regulates plant responses to jasmonate, ethylene, and fungi [66–70]. The statistically overrepresented CE-1 like sequence CACCG is an ABA response sequence in a number of ABA-related genes, and it is the target of the maize abscisic acid insensitive 4 (ABI4) protein [71]. Another Pscan motif, the SCHLAFMÜTZE (SMZ) binding site, is the target of an AP2-like transcription factor that acts as a repressor of flowering [72].

As a complement to our Pscan analysis, we searched for novel promoter DNA motifs associated with upregulation in *Arabidopsis* meiocytes using the Promzea motif discovery tool [26]. Nine overrepresented motifs were detected by Promzea with MNCP scores >1 in the promoters of the 50 meiotically active genes; five were detected in the promoters of the 50 randomly selected control genes (Supplemental File 4). This result supports the result from the Pscan analysis that meiotically active promoters possess more conserved motifs than randomly selected promoters. The 14 motifs matched to different experimentally defined motifs in the literature (Figure 3 and Supplemental File 5).

Motif1 from the Promzea analysis was statistically close to the TATABOX1 element, an element that is critical for the initiation of tissue specific transcription (Figure 3) [73, 74]. Motif4 matched the phosphate response domain GMHDLGMVSPB [75]. Motif3 matched the experimentally defined motif PIIATGAPB, which is responsible for light-activated gene expression [75]. Motif2 matched the E2FAT motif that is the binding site of E2F. The E2F transcription factors control the cell cycle by regulating the transcription of genes required for cell cycle and DNA replication [76]; these processes are obviously important in meiosis. Motif8 was similar to the pathogen/elicitor-related element TL1ATSAR [77]. Of the nine motifs predicted by Promzea, four motifs (Motif5, Motif6, Motif7, and Motif9) were enriched with CG, a property found in regulatory elements that is related to DNA methylation. CpG methylation is known to suppress transcription [78]. The presence of CG-enriched motifs identified in our analysis suggests that like gene activation, gene repression is also important for meiotically active gene regulatory networks, for example, the suppression of meiosis-restricted processes in somatic tissues. In addition, motif comparison analysis using STAMP found that these motifs possess other properties: Motif9 matched to INTRONLOER that is involved in $3'$ intron-exon splice junctions in plants [79], Motif5 matched to REGION1OSOSEM that is involved in the control of transcription by ABA [80], Motif6 matched to the tissue specific expression element BS1EGCCR [81], and Motif7 matched to the ammonium response element AMMORESVDCRNIA1 [82].

In the promoters of the negative genes, CREs are almost equally distributed on both the sense and the antisense strands (CREs on sense strand/CREs on antisense strand = 2742/2706 = 1/0.987); however, comparatively large numbers of CREs are located on the antisense strand compared to the sense strand of the meiotically active promoters (CREs on sense strand/CREs on antisense strand = 2758/2941 = 1/1.066). Interestingly, a similar bias of CRE distribution on the antisense strand is observed in promoters of rice sperm cell-specific genes [18].

The information from this study can be used in efforts to characterize the interactions between regulatory elements and TFs in meiocytes. Cell-type-specific analysis of TF expression is one of the strategies for sorting true protein-DNA interaction from numerous potentially spurious candidates [83]. For example, one of the PLACE motifs identified in this study, the GATABOX (Figure 1(a)), is the binding motif of the conserved C2C2-GATA TFs that have two GATA zinc fingers [40]. There are 29 C2C2-GATA family members that have been identified in *Arabidopsis*. They are highly expressed in early flower domains, and a few are involved in flower development [84, 85]. In our analysis, we identified two members of this family of TFs that are highly expressed in male meiocytes (*AT5G47140* and *AT1G08000*, Table 2). Therefore, AT5G47140 and AT1G08000 are better candidates than other C2C2-GATA family members for being proteins that can bind to GATABOX CREs in meiocytes. E2F transcription factors are essential for the regulation of the cell cycle and DNA replication. Three classical E2F proteins (E2Fa–c) and three atypical E2F proteins (E2Fd–f) have been characterized in *Arabidopsis* [86, 87]. Among these, *E2Fa* (*AT2G36010*) and *E2Fe* (*AT3G48160*) are highly expressed in meiocytes (Table 2); they may therefore be better candidates than other E2Fs for being proteins that can bind to E2FAT-like CREs in meiocytes, and this may link the E2Fs to the

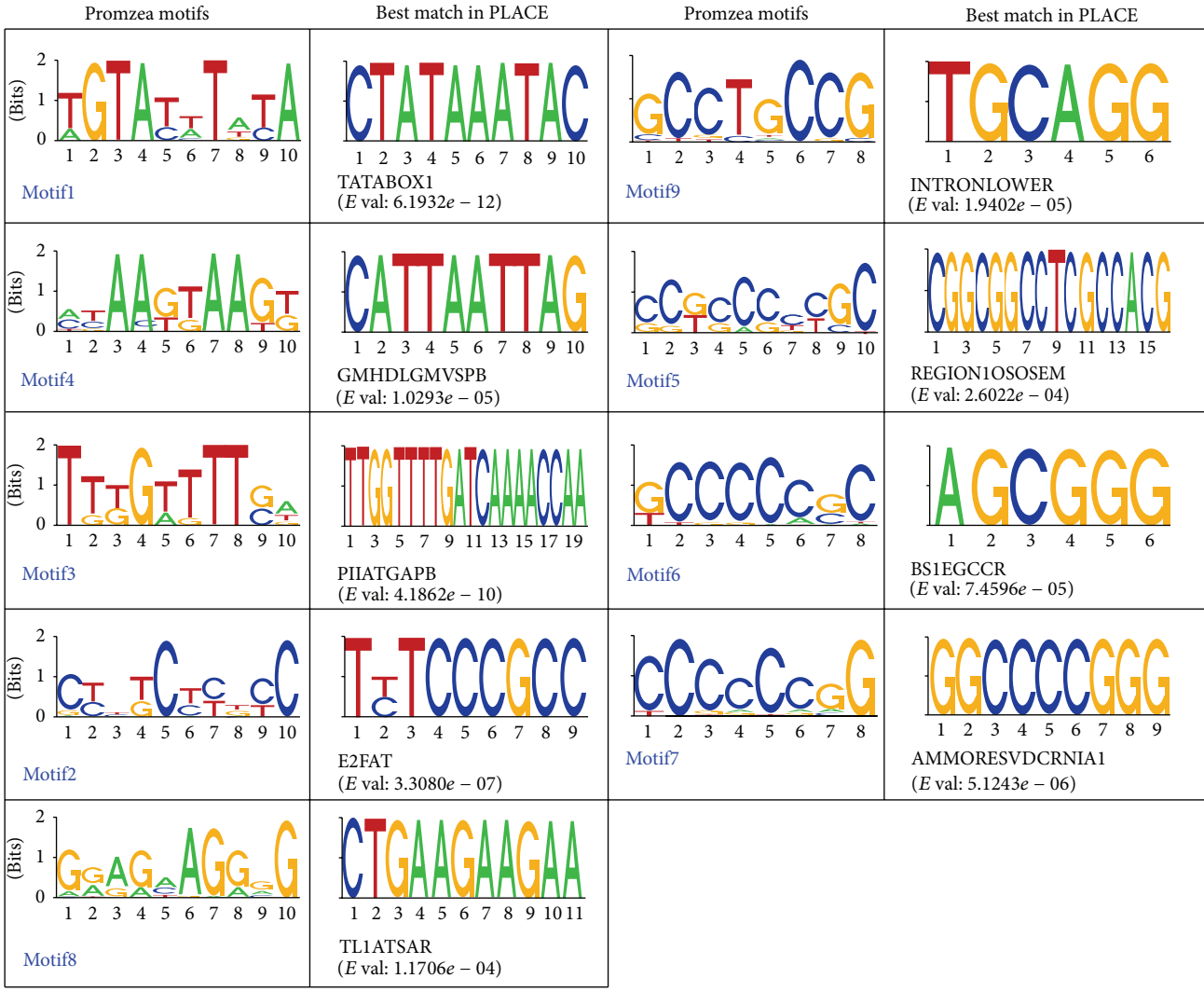


Figure 3: Sequence logos of novel motifs in the promoters of genes preferentially expressed during meiosis, detected using Promzea. The best match of each motif in the PLACE database is indicated in the panels to the right. Letters in the logos abbreviate the nucleotides (A, C, G, and T) and are sized relative to their occurrence. The *e*-value for STAMP is indicated by the false discovery ratio (FDR).

Table 2: Transcriptional factor genes preferentially expressed during meiosis with putative target binding sites highly enriched in meiocytes. Numbers in the boxes are ratios of read counts that indicate the difference in expression in bidirectional comparisons between each of the tissue pairs. M: meiocytes; A: anther; S: seedling.

Gene ID (name)	M/A	M/S	A/S
AT5G47140	2.30	4.09	1.78
AT1G08000	1.02	2.43	2.37
AT2G36010 (*E2Fa*)	0.73	2.19	3.01
AT3G48160 (*E2Fe*)	0.76	2.19	2.88

control of meiotic processes such as the meiotic cell cycle and DNA replication.

More than half of the overrepresented CREs identified in this study are binding sites of TFs that function in plant responses to environmental factors. We therefore infer that, during meiosis, exogenous signals are perceived largely through particular CRE and that this is especially likely for light and stress signals [89–92].

4. Conclusions

In this study, which aimed to identify CREs associated with genes preferentially expressed during meiosis, we analyzed 1 kb upstream regions of the 50 genes that were highly expressed in *Arabidopsis* meiocytes. Although the CREs in 10 promoters of meiotically active genes were analyzed in our previous study [20], here we performed a more comprehensive *in silico* study with a larger number of genes. The CREs that we identified in the promoters of these 50 genes may be responsible for the high activity of corresponding promoters in male meiocytes. The information obtained from this study can be used to identify TFs that regulate meiotically active gene expression and, more attractively, the synthesis

of artificial promoters that could drive high gene expression in meiocytes. As meiosis is evolutionarily conserved, the information on transcriptional domains obtained from the model system *Arabidopsis* has value not only in assessing the conservation of functional pathways in meiosis of other eukaryotes but also in applications seeking to improve crop plants.

Conflict of Interests

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

Authors' Contribution

Junhua Li and Jinhong Yuan contributed equally to the work.

Acknowledgments

The authors thank Dr. Changbin Chen (Department of Horticultural Science, University of Minnesota) for providing the transcriptome information on the *Arabidopsis* male meiocytes. This work was supported by the Scientific Research Starting Foundation of Henan Normal University (qd12127) and the Young Scientist Fund of Henan Normal University (2013QK11).

References

[1] J. Yanowitz, "Meiosis: making a break for it," *Current Opinion in Cell Biology*, vol. 22, no. 6, pp. 744–751, 2010.

[2] H. Ma, "A molecular portrait of Arabidopsis meiosis," *The Arabidopsis Book*, vol. 4, Article ID e0095, 2006.

[3] H. Ma, "Molecular genetic analyses of microsporogenesis and microgametogenesis in flowering plants," *Annual Review of Plant Biology*, vol. 56, pp. 393–434, 2005.

[4] A. Ronceret, M. Sheehan, and W. Pawlowski, "Chromosome dynamics in meiosis," in *Cell Division Control in Plants*, pp. 103–124, 2008.

[5] C. Chen, A. D. Farmer, R. J. Langley et al., "Meiosis-specific gene discovery in plants: RNA-Seq applied to isolated Arabidopsis male meiocytes," *BMC Plant Biology*, vol. 10, article 280, 2010.

[6] H. Yang, P. Lu, Y. Wang, and H. Ma, "The transcriptome landscape of Arabidopsis male meiocytes from high-throughput sequencing: the complexity and evolution of the meiotic process," *Plant Journal*, vol. 65, no. 4, pp. 503–516, 2011.

[7] A. Guo, K. He, D. Liu et al., "DATF: a database of *Arabidopsis* transcription factors," *Bioinformatics*, vol. 21, no. 10, pp. 2568–2569, 2005.

[8] K. Watanabe and K. Okada, "Two discrete *cis* elements control the abaxial side-specific expression of the *FILAMENTOUS FLOWER* gene in *Arabidopsis*," *Plant Cell*, vol. 15, no. 11, pp. 2592–2602, 2003.

[9] V. I. Klimyuk and J. D. G. Jones, "*AtDMC1*, the *Arabidopsis* homologue of the yeast *DMC1* gene: characterization, transposon-induced allelic variation and meiosis-associated expression," *Plant Journal*, vol. 11, no. 1, pp. 1–14, 1997.

[10] F. Couteau, F. Belzile, C. Horlow, O. Grandjean, D. Vezon, and M. Doutriaux, "Random chromosome segregation without meiotic arrest in both male and female meiocytes of a *dmc1* mutant of *Arabidopsis*," *Plant Cell*, vol. 11, no. 9, pp. 1623–1634, 1999.

[11] Y. Azumi, D. Liu, D. Zhao et al., "Homolog interaction during meiotic prophase I in *Arabidopsis* requires the *SOLO DANCERS* gene encoding a novel cyclin-like protein," *EMBO Journal*, vol. 21, no. 12, pp. 3081–3095, 2002.

[12] X. Yang, C. A. Makaroff, and H. Ma, "The Arabidopsis *MALE MEIOCYTE DEATH1* gene encodes a PHD-finger protein that is required for male meiosis," *Plant Cell*, vol. 15, no. 6, pp. 1281–1295, 2003.

[13] C. Chen, W. Zhang, L. Timofejeva, Y. Gerardin, and H. Ma, "The Arabidopsis *ROCK-N-ROLLERS* gene encodes a homolog of the yeast ATP-dependent DNA helicase MER3 and is required for normal meiotic crossover formation," *Plant Journal*, vol. 43, no. 3, pp. 321–334, 2005.

[14] J. W. Tullai, M. E. Schaffer, S. Mullenbrock, S. Kasif, and G. M. Cooper, "Identification of transcription factor binding sites upstream of human genes regulated by the phosphatidylinositol 3-kinase and MEK/ERK signaling pathways," *The Journal of Biological Chemistry*, vol. 279, no. 19, pp. 20167–20177, 2004.

[15] J. L. DeRisi, V. R. Iyer, and P. O. Brown, "Exploring the metabolic and genetic control of gene expression on a genomic scale," *Science*, vol. 278, no. 5338, pp. 680–686, 1997.

[16] S. L. Harmer, J. B. Hogenesch, M. Straume et al., "Orchestrated transcription of key pathways in Arabidopsis by the circadian clock," *Science*, vol. 290, no. 5499, pp. 2110–2113, 2000.

[17] J. L. Nemhauser, T. C. Mockler, and J. Chory, "Interdependency of brassinosteroid and auxin signaling in Arabidopsis," *PLoS Biology*, vol. 2, no. 9, article e258, 2004.

[18] N. Sharma, S. D. Russell, P. L. Bhalla, and M. B. Singh, "Putative cis-regulatory elements in genes highly expressed in rice sperm cells," *BMC Research Notes*, vol. 4, article 319, 2011.

[19] M. L. Engel, R. Holmes-Davis, and S. McCormick, "Green sperm. Identification of male gamete promoters in *Arabidopsis*," *Plant Physiology*, vol. 138, no. 4, pp. 2124–2133, 2005.

[20] J. Li, A. D. Farmer, I. E. Lindquist et al., "Characterization of a set of novel meiotically-active promoters in *Arabidopsis*," *BMC Plant Biology*, vol. 12, article 104, 2012.

[21] D. J. Craigon, N. James, J. Okyere, J. Higgins, J. Jotham, and S. May, "NASCArrays: a repository for microarray data generated by NASC's transcriptomics service," *Nucleic Acids Research*, vol. 32, pp. D575–D577, 2004.

[22] M. Thomas-Chollier, M. Defrance, A. Medina-Rivera et al., "RSAT 2011: regulatory sequence analysis tools," *Nucleic Acids Research*, vol. 39, no. 2, pp. W86–W91, 2011.

[23] F. Gubler, D. Raventos, M. Keys, R. Watts, J. Mundy, and J. V. Jacobsen, "Target genes and regulatory domains of the GAMYB transcriptional activator in cereal aleurone," *Plant Journal*, vol. 17, no. 1, pp. 1–9, 1999.

[24] K. Higo, Y. Ugawa, M. Iwamoto, and T. Korenaga, "Plant *cis*-acting regulatory DNA elements (PLACE) database: 1999," *Nucleic Acids Research*, vol. 27, no. 1, pp. 297–300, 1999.

[25] F. Zambelli, G. Pesole, and G. Pavesi, "Pscan: Finding over-represented transcription factor binding site motifs in sequences from co-regulated or co-expressed genes," *Nucleic Acids Research*, vol. 37, no. 2, pp. W247–W252, 2009.

[26] C. Liseron-Monfils, T. Lewis, D. Ashlock et al., "Promzea: a pipeline for discovery of co-regulatory motifs in maize and other plant species and its application to the anthocyanin and phlobaphene biosynthetic pathways and the Maize Development Atlas," *BMC Plant Biology*, vol. 13, no. 1, article 42, 2013.

[27] S. Mahony and P. V. Benos, "STAMP: a web tool for exploring DNA-binding motif similarities," *Nucleic Acids Research*, vol. 35, no. 2, pp. W253–W258, 2007.

[28] L. Bülow, Y. Brill, and R. Hehl, "AthaMap-assisted transcription factor target gene identification in *Arabidopsis thaliana*," *Database*, vol. 2010, Article ID baq034, 2010.

[29] L. Bülow, S. Engelmann, M. Schindler, and R. Hehl, "AthaMap, integrating transcriptional and post-transcriptional data," *Nucleic Acids Research*, vol. 37, no. 1, pp. D983–D986, 2009.

[30] C. Galuschka, M. Schindler, L. Bülow, and R. Hehl, "AthaMap web tools for the analysis and identification of co-regulated genes," *Nucleic Acids Research*, vol. 35, no. 1, pp. D857–D862, 2007.

[31] N. Ole Steffens, C. Galuschka, M. Schindler, L. Bülow, and R. Hehl, "AthaMap: an online resource for *in silico* transcription factor binding sites in the *Arabidopsis thaliana* genome," *Nucleic Acids Research*, vol. 32, pp. D368–D372, 2004.

[32] N. O. Steffens, C. Galuschka, M. Schindler, L. Bülow, and R. Hehl, "AthaMap web tools for database-assisted identification of combinatorial cis-regulatory elements and the display of highly conserved transcription factor binding sites in *Arabidopsis thaliana*," *Nucleic Acids Research*, vol. 33, no. 2, pp. W397–W402, 2005.

[33] S. Yanagisawa, "Dof1 and Dof2 transcription factors are associated with expression of multiple genes involved in carbon metabolism in maize," *Plant Journal*, vol. 21, no. 3, pp. 281–288, 2000.

[34] S. Yanagisawa and R. J. Schmidt, "Diversity and similarity among recognition sequences of Dof transcription factors," *Plant Journal*, vol. 17, no. 2, pp. 209–214, 1999.

[35] S. Yanagisawa and J. Sheen, "Involvement of maize Dof zinc finger proteins in tissue-specific and light-regulated gene expression," *Plant Cell*, vol. 10, no. 1, pp. 75–89, 1998.

[36] M. Mena, J. Vicente-Carbajosa, R. J. Schmidt, and P. Carbonero, "An endosperm-specific DOF protein from barley, highly conserved in wheat, binds to and activates transcription from the prolamin-box of a native B-hordein promoter in barley endosperm," *Plant Journal*, vol. 16, no. 1, pp. 53–62, 1998.

[37] U. Gowik, J. Burscheidt, M. Akyildiz et al., "*Cis*-regulatory elements for mesophyll-specific gene expression in the C_4 plant *Flaveria trinervia*, the promoter of the C_4 phosphoenolpyruvate carboxylase gene," *Plant Cell*, vol. 16, no. 5, pp. 1077–1090, 2004.

[38] H. Sakai, T. Aoyama, and A. Oka, "*Arabidopsis* ARR1 and ARR2 response regulators operate as transcriptional activators," *Plant Journal*, vol. 24, no. 6, pp. 703–711, 2000.

[39] A. Shirsat, N. Wilford, R. Croy, and D. Boulter, "Sequences responsible for the tissue specific promoter activity of a pea legumin gene in tobacco," *MGG Molecular & General Genetics*, vol. 215, no. 2, pp. 326–331, 1989.

[40] J. C. Reyes, M. I. Muro-Pastor, and F. J. Florencio, "The GATA family of transcription factors in arabidopsis and rice," *Plant Physiology*, vol. 134, no. 4, pp. 1718–1732, 2004.

[41] E. Lam and N. H. Chua, "ASF-2: a factor that binds to the cauliflower mosaic virus 35S promoter and a conserved GATA motif in *Cab* promoters.," *The Plant cell*, vol. 1, no. 12, pp. 1147–1156, 1989.

[42] S. A. Filichkin, J. M. Leonard, A. Monteros, P. Liu, and H. Nonogaki, "A novel endo-β-mannanase gene in tomato LeMAN5 is associated with anther and pollen development," *Plant Physiology*, vol. 134, no. 3, pp. 1080–1087, 2004.

[43] H. J. Rogers, N. Bate, J. Combe et al., "Functional analysis of *cis*-regulatory elements within the promoter of the tobacco late pollen gene *g*10," *Plant Molecular Biology*, vol. 45, no. 5, pp. 577–585, 2001.

[44] K. Stålberg, M. Ellerström, I. Ezcurra, S. Ablov, and L. Rask, "Disruption of an overlapping E-box/ABRE motif abolished high transcription of the *napA* storage-protein promoter in transgenic *Brassica napus* seeds," *Planta*, vol. 199, no. 4, pp. 515–519, 1996.

[45] G. Plesch, T. Ehrhardt, and B. Mueller-Roeber, "Involvement of TAAAG elements suggests a role for Dof transcription factors in guard cell-specific gene expression," *Plant Journal*, vol. 28, no. 4, pp. 455–464, 2001.

[46] P. Villain, R. Mache, and D. Zhou, "The mechanism of GT element-mediated cell type-specific transcriptional control," *The Journal of Biological Chemistry*, vol. 271, no. 51, pp. 32593–32598, 1996.

[47] A. S. Buchel, F. T. Brederode, J. F. Bol, and H. J. M. Linthorst, "Mutation of GT-1 binding sites in the *Pr*-1*A* promoter influences the level of inducible gene expression *in vivo*," *Plant Molecular Biology*, vol. 40, no. 3, pp. 387–396, 1999.

[48] H. Abe, T. Urao, T. Ito, M. Seki, K. Shinozaki, and K. Yamaguchi-Shinozaki, "Arabidopsis AtMYC2 (bHLH) and AtMYB2 (MYB) function as transcriptional activators in abscisic acid signaling," *Plant Cell*, vol. 15, no. 1, pp. 63–78, 2003.

[49] V. Chinnusamy, M. Ohta, S. Kanrar et al., "ICE1: a regulator of cold-induced transcriptome and freezing tolerance in *Arabidopsis*," *Genes and Development*, vol. 17, no. 8, pp. 1043–1054, 2003.

[50] U. Hartmann, M. Sagasser, F. Mehrtens, R. Stracke, and B. Weisshaar, "Differential combinatorial interactions of *cis*-acting elements recognized by R2R3-MYB, BZIP, and BHLH factors control light-responsive and tissue-specific activation of phenylpropanoid biosynthesis genes," *Plant Molecular Biology*, vol. 57, no. 2, pp. 155–171, 2005.

[51] T. Eulgem, P. J. Rushton, E. Schmelzer, K. Hahlbrock, and I. E. Somssich, "Early nuclear events in plant defence signalling: Rapid gene activation by WRKY transcription factors," *The EMBO Journal*, vol. 18, no. 17, pp. 4689–4699, 1999.

[52] Z. L. Zhang, Z. Xie, X. Zou, J. Casaretto, T. D. Ho, and Q. J. Shen, "A rice *WRKY* gene encodes a transcriptional repressor of the gibberellin signaling pathway in aleurone cells," *Plant Physiology*, vol. 134, no. 4, pp. 1500–1513, 2004.

[53] S. D. Simpson, K. Nakashima, Y. Narusaka, M. Seki, K. Shinozaki, and K. Yamaguchi-Shinozaki, "Two different novel *cis*-acting elements of *erd1*, a *clpA* homologous *Arabidopsis* gene function in induction by dehydration stress and dark-induced senescence," *Plant Journal*, vol. 33, no. 2, pp. 259–270, 2003.

[54] A. Bovy, C. Van Den Berg, G. De Vrieze, W. F. Thompson, P. Weisbeek, and S. Smeekens, "Light-regulated expression of the *Arabidopsis thaliana* ferredoxin gene requires sequences upstream and downstream of the transcription initiation site," *Plant Molecular Biology*, vol. 27, no. 1, pp. 27–39, 1995.

[55] M. Nakamura, T. Tsunoda, and J. Obokata, "Photosynthesis nuclear genes generally lack TATA-boxes: a tobacco photosystem I gene responds to light through an initiator," *Plant Journal*, vol. 29, no. 1, pp. 1–10, 2002.

[56] K. Kucho, S. Yoshioka, F. Taniguchi, K. Ohyama, and H. Fukuzawa, "*Cis*-acting elements and DNA-binding proteins involved in CO_2-responsive transcriptional activation of *Cah1* encoding a periplasmic carbonic anhydrase in *Chlamydomonas reinhardtii*," *Plant Physiology*, vol. 133, pp. 783–793, 2003.

[57] S. Yoshioka, F. Taniguchi, K. Miura, T. Inoue, T. Yamano, and H. Fukuzawa, "The novel Myb transcription factor LCR1 regulates the CO_2-responsive gene *Cah1*, encoding a periplasmic carbonic anhydrase in *Chlamydomonas reinhardtii*," *Plant Cell*, vol. 16, no. 6, pp. 1466–1477, 2004.

[58] T. Hirayama and K. Shinozaki, "A cdc5+ homolog of a higher plant, *Arabidopsis thaliana*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 93, no. 23, pp. 13371–13376, 1996.

[59] Z. Lin, K. Yin, D. Zhu, Z. Chen, H. Gu, and L. Qu, "AtCDC5 regulates the G2 to M transition of the cell cycle and is critical for the function of Arabidopsis shoot apical meristem," *Cell Research*, vol. 17, no. 9, pp. 815–828, 2007.

[60] X. Liang, T. J. Nazarenus, and J. M. Stone, "Identification of a consensus DNA-binding site for the *Arabidopsis thaliana* SBP domain transcription factor, AtSPL14, and binding kinetics by surface plasmon resonance," *Biochemistry*, vol. 47, no. 12, pp. 3645–3653, 2008.

[61] Y. Kagaya, K. Ohmiya, and T. Hattori, "RAV1, a novel DNA-binding protein, binds to bipartite recognition sequence through two distinct DNA-binding domains uniquely found in higher plants," *Nucleic Acids Research*, vol. 27, no. 2, pp. 470–478, 1999.

[62] Y. X. Hu, Y. H. Wang, X. F. Liu, and J. Y. Li, "Arabidopsis RAV1 is down-regulated by brassinosteroid and may act as a negative regulator during plant development," *Cell Research*, vol. 14, no. 1, pp. 8–15, 2004.

[63] K. H. Sohn, S. C. Lee, H. W. Jung, J. K. Hong, and B. K. Hwang, "Expression and functional roles of the pepper pathogen-induced transcription factor RAV1 in bacterial disease resistance, and drought and salt stress tolerance," *Plant Molecular Biology*, vol. 61, no. 6, pp. 897–915, 2006.

[64] K. Yamasaki, T. Kigawa, M. Inoue et al., "Solution structure of the B3 DNA binding domain of the Arabidopsis cold-responsive transcription factor RAV1," *Plant Cell*, vol. 16, no. 12, pp. 3448–3459, 2004.

[65] H. R. Woo, J. H. Kim, U. Lee et al., "The RAV1 transcription factor positively regulates leaf senescence in *Arabidopsis*," *Journal of Experimental Botany*, vol. 61, no. 14, pp. 3947–3957, 2010.

[66] M. Berrocal-Lobo and A. Molina, "Ethylene response factor 1 mediates *Arabidopsis* resistance to the soilborne fungus *Fusarium oxysporum*," *Molecular Plant-Microbe Interactions*, vol. 17, no. 7, pp. 763–770, 2004.

[67] M. Berrocal-Lobo, A. Molina, and R. Solano, "Constitutive expression of *ETHYLENE − RESPONSE − FACTOR1* in *Arabidopsis* confers resistance to several necrotrophic fungi," *Plant Journal*, vol. 29, no. 1, pp. 23–32, 2002.

[68] D. Hao, M. Ohme-Takagi, and A. Sarai, "Unique mode of GCC box recognition by the DNA-binding domain of ethylene-responsive element-binding factor (ERF domain) in plant," *Journal of Biological Chemistry*, vol. 273, no. 41, pp. 26857–26861, 1998.

[69] M. Ohme-Takagi and H. Shinshi, "Ethylene-inducible DNA binding proteins that interact with an ethylene-responsive element," *Plant Cell*, vol. 7, no. 2, pp. 173–182, 1995.

[70] R. Solano, A. Stepanova, Q. Chao, and J. R. Ecker, "Nuclear events in ethylene signaling: A transcriptional cascade mediated by ETHYLENE-INSENSITIVE3 and ETHYLENE-RESPONSE-FACTOR1," *Genes and Development*, vol. 12, no. 23, pp. 3703–3714, 1998.

[71] X. Niu, T. Helentjaris, and N. J. Bate, "Maize ABI4 binds coupling element1 in abscisic acid and sugar response genes," *Plant Cell*, vol. 14, no. 10, pp. 2565–2575, 2002.

[72] J. Mathieu, L. J. Yant, F. Mürdter, F. Küttner, and M. Schmid, "Repression of flowering by the miR172 target SMZ," *PLoS Biology*, vol. 7, no. 7, Article ID e1000148, 2009.

[73] J. Kharazmi and C. Moshfegh, "Investigation of *dmyc* promoter and regulatory regions," *Gene Regulation and Systems Biology*, vol. 2013, no. 7, pp. 85–102, 2013.

[74] M. L. Grace, M. B. Chandrasekharan, T. C. Hall, and A. J. Crowe, "Sequence and spacing of TATA box elements are critical for accurate initiation from the β-phaseolin promoter," *The Journal of Biological Chemistry*, vol. 279, no. 9, pp. 8102–8110, 2004.

[75] N. M. Creux, M. Ranik, D. K. Berger, and A. A. Myburg, "Comparative analysis of orthologous cellulose synthase promoters from *Arabidopsis*, *Populus* and *Eucalyptus*: evidence of conserved regulatory elements in angiosperms," *New Phytologist*, vol. 179, no. 3, pp. 722–737, 2008.

[76] K. Helin, "Regulation of cell proliferation by the E2F transcription factors," *Current Opinion in Genetics and Development*, vol. 8, no. 1, pp. 28–35, 1998.

[77] F. Yu, Y. Huaxia, W. Lu, C. Wu, X. Cao, and X. Guo, "*GhWRKY15*, a member of the WRKY transcription factor family identified from cotton (*Gossypium hirsutum* L.), is involved in disease resistance and plant development," *BMC Plant Biology*, vol. 12, article 144, 2012.

[78] C. Hsieh, "Dependence of transcriptional repression on CpG methylation density," *Molecular and Cellular Biology*, vol. 14, no. 8, pp. 5487–5494, 1994.

[79] Y. López, A. Patil, and K. Nakai, "Identification of novel motif patterns to decipher the promoter architecture of co-expressed genes in *Arabidopsis thaliana*," *BMC Systems Biology*, vol. 7, article S10, 2013.

[80] T. Hattori, T. Terada, and S. Hamasuna, "Regulation of the *Osem* gene by abscisic acid and the transcriptional activator VP1: analysis of *cis*-acting promoter elements required for regulation by abscisic acid and VP1," *Plant Journal*, vol. 7, no. 6, pp. 913–925, 1995.

[81] E. Lacombe, J. van Doorsselaere, W. Boerjan, A. M. Boudet, and J. Grima-Pettenati, "Characterization of *cis*-elements required for vascular expression of the *Cinnamoyl CoA Reductase* gene and for protein—DNA complex formation," *The Plant Journal*, vol. 23, no. 5, pp. 663–676, 2000.

[82] R. Loppes and M. Radoux, "Identification of short promoter regions involved in the transcriptional expression of the nitrate reductase gene in *Chlamydomonas reinhardtii*," *Plant Molecular Biology*, vol. 45, no. 2, pp. 215–227, 2001.

[83] Y. Jiao and E. M. Meyerowitz, "Cell-type specific analysis of translating RNAs in developing flowers reveals new levels of control," *Molecular Systems Biology*, vol. 6, article 419, 2010.

[84] Y. Zhao, L. Medrano, K. Ohashi et al., "Hanaba taranu is a GATA transcription factor that regulates shoot apical meristem and flower development in *Arabidopsis*," *Plant Cell*, vol. 16, no. 10, pp. 2586–2600, 2004.

[85] C. D. Mara and V. F. Irish, "Two GATA transcription factors are downstream effectors of floral homeotic gene action in *Arabidopsis*," *Plant Physiology*, vol. 147, no. 2, pp. 707–718, 2008.

[86] L. de Veylder, T. Beeckman, and D. Inzé, "The ins and outs of the plant cell cycle," *Nature Reviews Molecular Cell Biology*, vol. 8, no. 8, pp. 655–665, 2007.

[87] T. Lammens, J. Li, G. Leone, and L. De Veylder, "Atypical E2Fs: new players in the E2F transcription factor family," *Trends in Cell Biology*, vol. 19, no. 3, pp. 111–118, 2009.

[88] M. Godoy, J. M. Franco-Zorrilla, J. Pérez-Pérez, J. C. Oliveros, Ó. Lorenzo, and R. Solano, "Improved protein-binding microarrays for the identification of DNA-binding specificities of transcription factors," *Plant Journal*, vol. 66, no. 4, pp. 700–711, 2011.

[89] M. Jain, A. K. Tyagi, and J. P. Khurana, "Constitutive expression of a meiotic recombination protein gene homolog, *OsTOP6A1*, from rice confers abiotic stress tolerance in transgenic *Arabidopsis* plants," *Plant Cell Reports*, vol. 27, no. 4, pp. 767–778, 2008.

[90] B. C. Lu, "Genetic recombination in *Coprinus*. IV. A kinetic study of the temperature effect on recombination frequency," *Genetics*, vol. 78, no. 2, pp. 661–677, 1974.

[91] B. C. Lu, "The control of meiosis progression in the fungus *Coprinus cinereus* by light/dark cycles," *Fungal Genetics and Biology*, vol. 31, no. 1, pp. 33–41, 2000.

[92] X. Niu, L. Renshaw-Gegg, L. Miller, and M. J. Guiltinan, "Bipartite determinants of DNA-binding specificity of plant basic leucine zipper proteins," *Plant Molecular Biology*, vol. 41, no. 1, pp. 1–13, 1999.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 815672, 9 pages
http://dx.doi.org/10.1155/2014/815672

Research Article

A Genome-Wide Identification of Genes Undergoing Recombination and Positive Selection in *Neisseria*

Dong Yu, Yuan Jin, Zhiqiu Yin, Hongguang Ren, Wei Zhou, Long Liang, and Junjie Yue

Beijing Institute of Biotechnology, Beijing 100071, China

Correspondence should be addressed to Long Liang; ll@bmi.ac.cn and Junjie Yue; yue_junjie@126.com

Received 2 June 2014; Revised 18 July 2014; Accepted 18 July 2014; Published 10 August 2014

Academic Editor: Shiwei Duan

Currently, there is particular interest in the molecular mechanisms of adaptive evolution in bacteria. *Neisseria* is a genus of gram negative bacteria, and there has recently been considerable focus on its two human pathogenic species *N. meningitidis* and *N. gonorrhoeae*. Until now, no genome-wide studies have attempted to scan for the genes related to adaptive evolution. For this reason, we selected 18 *Neisseria* genomes (14 *N. meningitidis*, 3 *N. gonorrhoeae* and 1 commensal *N. lactamics*) to conduct a comparative genome analysis to obtain a comprehensive understanding of the roles of natural selection and homologous recombination throughout the history of adaptive evolution. Among the 1012 core orthologous genes, we identified 635 genes with recombination signals and 10 genes that showed significant evidence of positive selection. Further functional analyses revealed that no functional bias was found in the recombined genes. Positively selected genes are prone to DNA processing and iron uptake, which are essential for the fundamental life cycle. Overall, the results indicate that both recombination and positive selection play crucial roles in the adaptive evolution of *Neisseria* genomes. The positively selected genes and the corresponding amino acid sites provide us with valuable targets for further research into the detailed mechanisms of adaptive evolution in *Neisseria*.

1. Introduction

Homologous recombination and positive selection are two indispensable sources of genetic variation and play central roles in the adaptive evolution of many bacteria species [1, 2]. Of the two mechanisms, homologous recombination occurs frequently in some bacteria, such as *Streptomyces* [3], *Helicobacter pylori* [4], and *Neisseria* [5], and could possibly speed adaptation by reducing competition between beneficial mutations [6]. There is also evidence for positive selection in specific genes in certain pathogens, such as *Listeria monocytogenes* [7], *Salmonella* [8], *Streptococcus* [9], *Campylobacter* [10], and *Actinobacilus pleuropneumoniae* [11]. These positively selected genes are usually involved in the dynamic interaction between host and pathogen [12, 13].

At present, there are well-developed methods for detecting genes undergoing recombination and selection. Phi [14] and GENECONV [15] are two common methods used to detect recombination based on different statistical tests. The d_N/d_S-based method is typically used to estimate the ratio of the rate of nonsynonymous nucleotide substitutions to that of synonymous substitutions [16, 17]. This ratio indicates whether a gene has been under positive selection ($\omega > 1$), neutral selection ($\omega = 1$), or purifying selection ($\omega < 1$). Combined with the codon models developed by Nielsen and Yang [16, 18], which allow variation in ω among sites, this method can identify positive selection signals when there are only few positive sites. All these methods will be employed in this study to detect the genes with the history of recombination or positive selection.

Neisseria is a genus of bacteria that colonizes the mucosal surfaces of many animals. Of the known 14 species, only 2 species, *Neisseria meningitides* and *Neisseria gonorrhoeae*, are human pathogens; and the remainders are all commensal or nonpathogenic. Until now, there have been many comparative genomic studies on the genomic evolution of these two pathogenic species [5, 19–27]. Homologous recombination has been found to play a key role in the adaptive evolution of *Neisseria*; however, few studies have characterised the effect of positive selection on the *Neisseria* genome. Only two genes, porB [28] and pilE [29], have received attention, and both have undergone strong positive selection pressure. In this

study, we used the genome sequences available for the strains of *N. meningitidis, N. gonorrhoeae*, and nonpathogenic *N. lactamica* to investigate the contributions of recombination and positive selection to the evolution of *Neisseria* genomes. Considering the high sequence diversity and open pan-genome, we focused on the core genome genes during our scan for recombined genes and positively selected genes. Statistical tests and a literature review were conducted to determine the association between genes and the properties of this genus.

2. Materials and Methods

2.1. Data Preparation. Eighteen genome sequences of *Neisseria*, including complete proteomes and the corresponding coding genes, were retrieved from the NCBI Genome database (http://www.ncbi.nlm.nih.gov/genome/bacte- ria). Detailed information, such as Genbank ID and genome size, is listed in Table 1. The COGs (clusters of orthologous groups of proteins) functional classification for each proteome was conducted with ID mapping from the Uniprot database [30]. Then, using *Neisseria gonorrhoeae* FA 1090 as the reference genome, stand-alone BLAST was performed against the proteomes of the remaining 17 strains for homologs (sequence identity > 80% and alignment coverage > 80%) of each of the FA_1090 proteins. For each of the core genes from FA_1090, BLAST was performed against all 18 genomes (including the reference genome) with the same thresholds, and multiple copies in any genome were reported and removed from further analysis. The remaining core proteins were defined as the core orthologs of *Neisseria*.

2.2. Alignment and Calculation of Nucleotide Diversity, Informative Sites, Codon Bias, d_N, and d_S. The orthologous protein sequences were aligned using the method implemented in muscle [31]. Then, multiple codon alignments of genes corresponding to protein sequence alignments were obtained using PAL2NAL [32]. Using the resulting gene alignments, the gene-by-gene number of informative sites and the nucleotide diversity were obtained from the output of the PhiPack program [14].

In this study, the effective number of codons (N_c) was used to measure the codon bias. The N_c value ranges from 20 for the strongest bias to 61 for no bias [33], and the program CodonW (http://sourceforge.net/projects/codonw/) was used to calculate the values of N_c for each gene. The number of synonymous nucleotide substitutions per synonymous site (d_S) and the number of nonsynonymous nucleotide substitutions per nonsynonymous site (d_N) were estimated from the gene alignments using the program SNAP [34].

2.3. Detection of Recombination. Four statistical procedures GENECONV [15], pairwise homoplasy index (Phi) [14], maximum χ^2 [35], and neighbor similarity score (NSS) [36] were run on the aligned genes to discover the homologous recombination signals. For the analyses of GENECONV, the parameter g-scale was set to 1, which allows mismatches within a recombining fragment. The P values were calculated from 10000 random permutations of the data. The remaining three programs were implemented in the PhiPack package and were run with default parameters.

2.4. Detection of Selection. FastTree [37] was used to construct maximum likelihood phylogenetic trees with a general time-reversible (GTR) model of nucleotide substitution for each gene alignment. The resulting topologies of ML trees were applied to subsequent selection analysis.

The codeml program from PAML [38] was used to detect the genes under positive selection. Two site-specific models were applied: the null model M1a (nearly neutral) and the alternative model M2a (positive selection); the two models differ by the statistical distribution assumed for the ω ratio. The latter model allows sites with $\omega > 1$, whereas the former only allows sites with ω varying between 0 and 1. To ensure convergence to the best likelihood, all calculations were performed three times. A likelihood ratio test (LRT) was then carried out to infer the occurrence of sites under positive selection pressure through comparing M1a against M2a. P values were determined from the LRT scores calculated by the module χ^2 of the PAML package.

2.5. Statistical Analysis. Correction for multiple testing was performed using the method presented by Benjamini and Hochberg [39]. For all genes tested for recombination and positive selection, q-values were calculated for each P value using the R package [40, 41] (q-value with the proportion of true null hypothesis set to 1). According to the conservation of tests, false discovery rates of 10% and 20% were used for the recombination analyses and positive selection detection, respectively.

The significance level for differences among the properties, including nucleotide diversity, codon bias, d_S, and d_N, between a COG and other COGs was determined using the nonparametric Mann-Whitney U-test. Correlation between each COG and evolutionary forces (homologous recombination and positive selection) was estimated using a binomial test. Then, Bonferroni corrections for multiple comparisons were performed according to the number of one-sided tests. The significance level was set to 5%. All statistical tests were carried out using Python scripts and R.

3. Results and Discussion

3.1. Characterization of the Orthologous Genes in 18 Neisseria Genomes. Previous studies [42–45] showed that both intraspecies and interspecies recombination could act as the important genetic mechanism in generating new clones and alleles in *Neisseria*. The genus *Neisseria* consists of two important pathogenic species and a dozen species that are never or rarely pathogenic. At present, there are only 18 completely sequenced genomes of genus *Neisseria* available, including 14 *N. meningitidis*, 3 *N. gonorrhoeae*, and 1 *N. lactamics* genomes. Thus, we selected all 18 genomes to conduct a genome-wide scan for the identification of genes exhibiting recombination or positively selected signals.

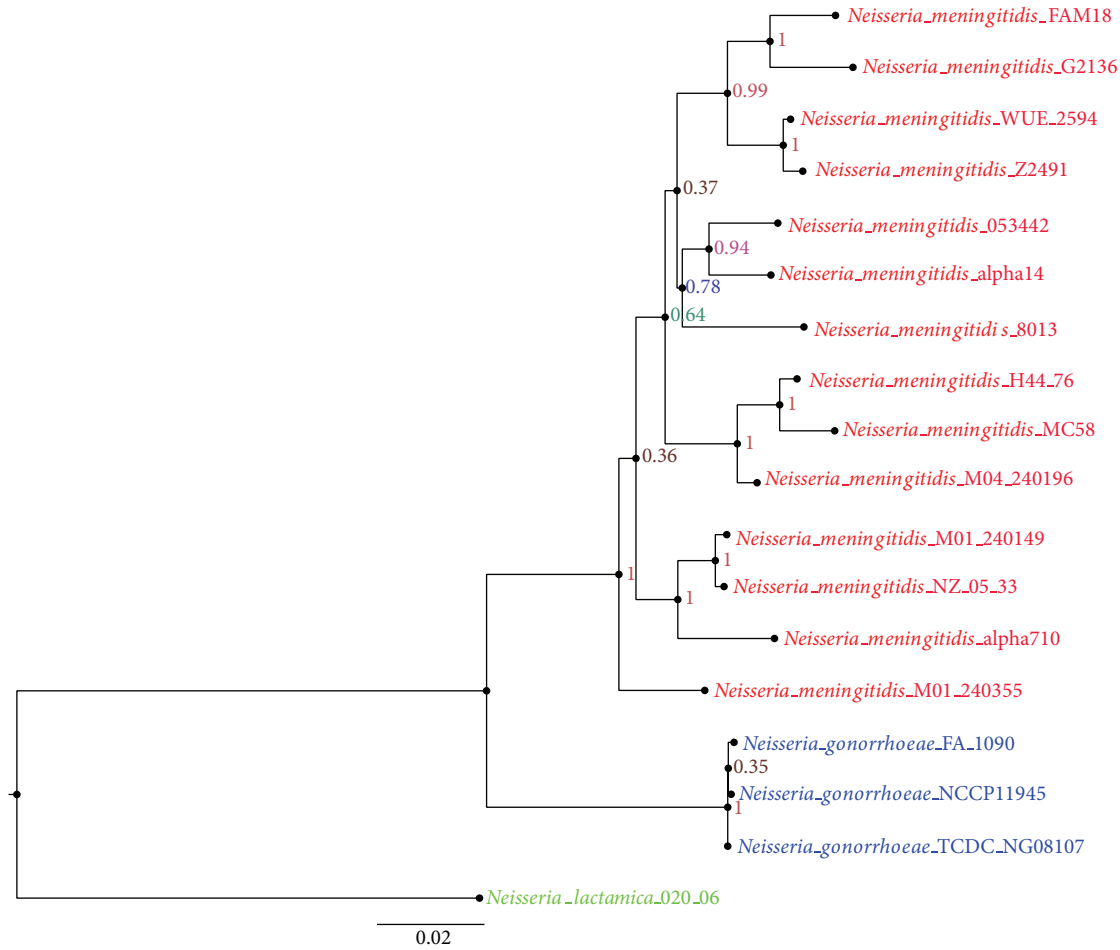


Figure 1: Phylogram of concatenated sequences of 7 housekeeping genes (abcZ, adk, aroE, fumC, gdh, pdhC, and pgm) for the 18 *Neisseria* genomes analyzed. The genomes in different species are marked with different colors: red for *Neisseria meningitides*, blue for *Neisseria gonorrhoeae*, and green for *Neisseria lactamics*. The numbers labeled on each internal node are the boostrap values.

The phylogenetic relationships of the 18 strains were first established based on the 7 housekeeping genes frequently used for multilocus sequence typing (MLST) analysis of *Neisseria*: abcZ, adk, aroE, fumC, gdh, pdhC, and pgm [46]. The 7 genes were concatenated to construct a maximum likelihood tree with high bootstrap values as shown in Figure 1. In the tree, the three species were divided into three clades and formed a monophyly, respectively.

In the next step, *N. gonorrhoeae* FA 1090 was used as the reference genome to perform a BLAST search against the other 17 *Neisseria* genomes for orthologs. Finally, 1034 genes were identified as present in all 18 genomes, containing the initial definition of the core genome for these *Neisseria* strains and accounting for 38.73% to 55.45% of the coding genes in each genome. This proportion is similar to that in previous analysis of *Neisseria meningitides* genomes [5, 27]. Of the 1034 core genes, 22 genes occurred as two or more copies in some genomes and were excluded from further analysis. The remaining 1012 genes with a single copy per genome were then defined as the core orthologous genes for subsequent analysis of homologous recombination and natural selection.

Among these genes, genes in COGs "Replication, recombination, and repair" were found to show higher nucleotide diversity than genes in other COGs (Table 2). For the association between the number of informative sites and COGs, the same result was obtained, which means genes in category "Replication, recombination, and repair" also had more informative sites than genes in other COGs (Table 2).

The effective number of codons, abbreviated as N_c, was used to measure the codon bias for each orthologous gene. Genes categorised into the COG "Translation, ribosomal structure and biogenesis" were evident to have a significant higher codon bias compared with genes in other COG categories (Table 2). It is well known that genes with a lower N_c can have a strong bias and are more likely to be highly expressed [47–49]. So, the genes in the two COGs might present housekeeping features in the fundamental life cycle and essential physiological activities of *Neisseria*.

In the same way, an association between COGs and d_N or d_S was also observed. There were 4 COGs in which genes were found to have higher rates of synonymous nucleotide substitutions in comparison with other categories. On the other

Table 1: Genome sequences used in this study.

Strain name	GenBank accession no.	Genome size (Mbp)	No. of CDS	CC ID
*Neisseria_meningitidis*_FAM18	NC_008767	2.19	1917	CC11
*Neisseria_meningitidis*_G2136	NC_017513	2.18	1928	CC8
*Neisseria_meningitidis*_WUE_2594	NC_017512	2.23	1941	CC5
*Neisseria_meningitidis*_Z2491	NC_003116	2.18	1909	CC4
*Neisseria_meningitidis*_8013	NC_017501	2.28	1913	CC18
*Neisseria_meningitidis*_053442	NC_010120	2.15	2020	CC4821
*Neisseria_meningitidis*_alpha14	NC_013016	2.14	1872	CC53
*Neisseria_meningitidis*_M04_240196	NC_017515	2.25	1947	CC269
*Neisseria_meningitidis*_H44_76	NC_017516	2.24	1961	CC32
*Neisseria_meningitidis*_MC58	NC_003112	2.27	2063	CC32
*Neisseria_meningitidis*_alpha710	NC_017505	2.24	2017	CC41/44
*Neisseria_meningitidis*_M01_240149	NC_017514	2.22	1936	CC41/44
*Neisseria_meningitidis*_NZ_05_33	NC_017518	2.24	1948	CC41/44
*Neisseria_meningitidis*_M01_240355	NC_017517	2.29	1971	CC213
*Neisseria_gonorrhoeae*_TCDC_NG08107	NC_017511	2.15	2196	
*Neisseria_gonorrhoeae*_FA_1090	NC_002946	2.15	2002	
*Neisseria_gonorrhoeae*_NCCP11945	NC_011035	2.23	2680	
*Neisseria_lactamica*_020_06	NC_014752	2.22	1972	

Table 2: Association between COGs and descriptive variables.

Functional category	Number of genes analyzed	Bonferroni-corrected P value for one-sided U−test for association between genes in a given COG and[(1)]					
		>nt diversity	>Number of Informative sites	>Codon bias[(2)]	<Codon bias[(2)]	$>d_S$	$>d_N$
Energy production and conversion	85						**<0.001**
Nucleotide metabolism and transport	37					**0.03**	
Translation, ribosomal structure, and biogenesis	110			**<0.001**		**0.03**	
Replication, recombination, and repair	70	**<0.001**	**<0.001**				**0.03**
Cell wall/membrane/envelope biogenesis	75					**0.002**	
Function unknown	93				**0.023**		**0.03**
Intracellular trafficking, secretion and vesicular transport	29				**0.020**	**0.039**	
Not in COGs	47				**0.040**		**<0.001**

[(1)]" >" or "<" indicates the direction of the one-sided tests (i.e. ">Codon bias" shows Bonferroni-corrected P-values for associations between genes in a given COG and higher codon bias as compared to the genes in other COGs, and "<Codon bias" represents a contrast tendency).
[(2)]Tests for codon bias were performed using N_c values (a lower N_c means increased codon bias).

hand, genes in the other 4 COGs also showed a tendency to have higher rates of nonsynonymous substitutions in comparison with genes in other COGs (Table 2). It is worth noting that all the genes in the core genome in *Neisseria* had higher d_S and d_N rates than the genes in other bacteria, for example, *E. coli* [50] and *A. pleuropneumoniae* [11], indicating that strong natural selection might act on *Neisseria*.

3.2. A Considerable Number of Genes Showing Evidence of Recombination. Until now, there were several different strategies for identifying the homologous recombination regions in sequences. In this study, four common statistical test methods, including NSS, Max-χ^2, Phi, and GENECONV, were employed to detect the recombination signals among the 1012 orthologous genes. As a result, a total of 996 genes (98.4% of all 1012 core genome genes) were found to show significant evidence (FDR < 10%) of recombination by at least one of the four tests. Overall, 951, 968, 842, and 727 genes were identified to show significant evidence of recombination by NSS, Max-χ^2, Phi, and GENECONV, respectively. Additionally, a total of 635 genes (62.7% of 1012 core genome genes) were showed recombination signals in all four tests. The proportion of genes undergoing recombination ranged from 62.7% to 98.4%, which is higher than those typically observed in other

Table 3: Genes under positive selection.

Gene	Cluster ID	COG	Function	$2\Delta L$	q-value	ω	Positively selected sites
dnaE	N35	L	DNA polymerase III alpha subunit	48.459	0.016	13.082	413, 968, 971, 972
	N139	P	Ammonium transporter	42.474	0.096	63.631	12, 14, 18, 19, 20, 21, 67
recB	N245	L	DNA helicase	67.811	0.000	4.287	4, 251, 865, 869, 882, 1036, 1137, 1184
hup	N352	P	TonB-dependent receptor	129.195	0.000	7.064	263, 265, 282, 287, 288, 290, 291, 293, 304, 378, 380, 535, 538, 553, 557, 561, 646, 810, 884, 889, 891
	N380	M	Hypothetical protein	46.418	0.029	152.674	18, 19, 20, 23, 24, 25, 26, 28, 30
dnaX	N436	L	DNA polymerase III gamma and tau subunit	57.997	0.001	6.032	228, 294, 329, 512, 559
uraA	N514	F	Uracil permease	55.222	0.002	16.737	2, 9, 10, 17, 24, 25, 29, 31, 395, 455
	N832	S	Hypothetical protein	51.544	0.006	6.580	190, 207, 212, 228, 232, 276, 314, 368, 401, 514, 729
frpB	N966	P	Iron-regulated outer membrane protein	125.098	0.000	4.333	341, 342, 343, 348, 394, 409, 415, 451, 459, 466, 467, 471, 472, 473, 476, 483, 674, 688, 718, 730, 739
polA	N973	L	DNA polymerase I	54.283	0.003	6.489	212, 866, 867, 881, 882, 898

bacteria, such as *E. coli*. The result suggests that homologous recombination plays an important role in the evolution of *Neisseria* genomes.

In a previous work [5], Joseph et al. identified 459 ortholog genes with signs of recombination in *Neisseria meningitidis* genomes, which accounts for 39.6% of all core genome genes. In this work, only *Neisseria meningitidis* genomes were for recombination test, the abovementioned 459 orthologous genes with signs of recombination could be considered intraspecies recombinations. In our present work, in addition to the *N. meningitidis* genomes, the genomes of *Neisseria gonorrhoeae*, and *Neisseria lactamica* were also selected for the recombination analyses and several interspecies recombination genes were identified. The interspecies recombination events in the genus *Neisseria* have been reported many times [44–46]. It is not surprising that the proportion of genes with recombination signals in the present work is markedly higher than the value observed by Joseph et al. It can be deduced that both intraspecies and interspecies recombination could act as important genetic mechanisms for generating new clones and alleles [47] in *Neisseria*.

To test whether the high percentage of core genome genes with a recombination signal is caused by the choice of genomes, we carried out the same analysis on the 14 *N. meningitidis* genome sequences with the same parameters. We first obtained 1211 orthologous genes with a single copy per genome. Among these orthologous genes, 634 (52.4%) genes were identified to show significant evidence of recombination by all the four tests. In this case, a lower percentage of genes with recombination signals were identified, confirming that the choice of genomes really has an impact on the percentage of recombined genes in the core genome. It also indicated that interspecies recombination indeed has a role in the evolution of *Neisseria* genomes. Additionally, a higher proportion of genes with recombination signals were observed in these 14 *N. meningitidis* genomes compared with the results in Joseph's work. The reason could lie in the differences in the specific genomes in both analyses, suggesting that intraspecies recombination plays an unexpected role in the evolution of the *N. meningitidis* genome. In a word, recombination acts as an important and irreplaceable genetic mechanism in shaping the genomes of genus *Neisseria*.

Moreover, it is worth noting that the core genes identified as recombinants have high rates of d_S and d_N, nucleotide diversity and the number of information sites ($P < 0.001$, $P < 0.001$, $P < 0.001$ and $P < 0.001$, respectively, one-sided U-test). The association between COG categories and the number of recombined genes was also estimated (Figure 2). Only two COGs "general function prediction only" and "function unknown" were significantly overrepresented with recombined genes. However, after Bonferroni correction, all the genes exhibiting evidence of recombination were distributed with no significance in all COGs. This unbiasedness of recombined genes in function further confirmed the role of recombination in shaping genomes during the evolution of *Neisseria*.

3.3. 10 Genes Showing Evidence of Positive Selection. The detection of positive selection for the 1012 orthologs was conducted in PAML, and models M1a and M2a of variable selective pressure across codon sites were used to estimate selective pressure and test for positive selection. Based on LRT statistics for comparing the null model and alternative model with χ^2 distribution and correction for multiple testing (FDR < 20%), a total of 10 genes were identified to be under strong selected pressure. Of the 10 genes, 4 belonged to the COG "Replication, recombination, and repair", and 3 were in the COG "Inorganic ion transport and metabolism." The remaining three genes were classified into the "cell wall/membrane/envelope biogenesis," "nucleotide transport and metabolism," and "function unknown", respectively (Table 3).

In the same way, two obvious discrepancies were observed, respectively, for values of d_S and the number of informative sites between genes under positive selection and

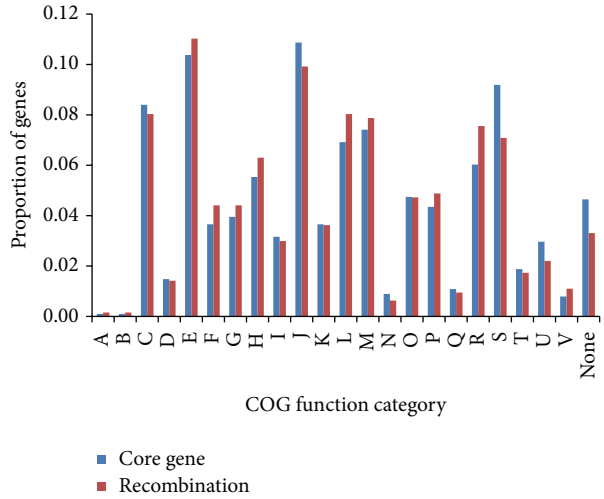


Figure 2: Genes with recombination signals are distributed with no significance in all COGs. The x axis represents different COG categories. The y axis represents the proportion of genes in each COG category. The proportion of genes with evidence for recombination and core genes for each COG are represented by red and blue bars, respectively. The COG categories are coded as follows: A, RNA processing and modification; B, chromatin structure and dynamics; C, energy production and conversion; D, cell cycle control, cell division, and chromosome partitioning; E, amino acid transport and metabolism; F, nucleotide transport and metabolism; G, carbohydrate transport and metabolism; H, coenzyme transport and metabolism; I, lipid transport and metabolism; J, translation, ribosomal structure and biogenesis; K, transcription; L, replication, recombination, and repair; M, cell wall/membrane/envelope biogenesis; N, cell motility; O, posttranslational modification, protein turnover, and chaperone; P, inorganic ion transport and metabolism; Q, secondary metabolites biosynthesis, transport, and catabolism; R, general function prediction only; S, function unknown; T, signal transduction mechanisms; U, intracellular trafficking, secretion, and vesicular transport; V, defense mechanisms; None, not in COGs.

the remaining genes ($P = 0.024$ and $P = 0.005$, one-sided U-test). Furthermore, all 10 positively selected genes were found to show significant evidence of recombination detected by at least one recombination test. Only one gene was not in the genes identified by all four tests. The probable reason for this is that recombination could form phylogenetic incongruence [51, 52].

Compared to the high proportion of recombined genes, few positively selected genes (10) were identified, accounting for approximately 1% of the core genome. Similar proportion was also obtained in *E. coli* [12], but is smaller than those of other pathogenic bacteria, such as *A. pleuropneumoniae* [11].

Among the protein products encoded by the 10 positively selected genes, only 8 proteins were annotated with definite functions. We found that these proteins were either involved in DNA processing or inorganic transport and metabolism.

Of the 10 genes, recB, encoding the DNA helicase, is an integral part of recBCD homologous recombined enzyme. Mutations in recB are required for double-strand break repair [53] and can also reduce the frequency of many types of recombination events [54]. dnaE, dnaX and polA are all DNA polymerase genes. The first two encode the polymerase iii subunits, and the last encodes polymerase I. All three play fundamental roles in DNA metabolism, including DNA replication, recombination, and repair. In a word, positive selection on the four genes might ensure the strain to adapt to frequent recombination in the genomes.

AmtB encodes an ammonium transporter and is involved in ammonium transmembrane transporter activity. uraA encodes a uracil permease involved in transmembrane transport as well and acts as a membrane-bound facilitator for the transport of uracil across the cell membrane into the cytoplasm [55]; it is therefore necessary for uracil uptake, especially at low exogenous uracil concentrations and even under conditions with high UPRTase activity.

Hup encodes a TonB-dependent receptor that utilizes heme as an iron source [56]. It has been reported that mutations in the hemoglobin receptor gene have profound effects on the survival of *N. meningitidis* in an infant rat, indicating that this gene is important for the virulence of *Neisseria* [57].

FrpB is clearly a virulence gene, encoding an iron-regulated outer membrane protein. It is a member of the TonB-dependent transporter family and is responsible for iron uptake into the periplasm. FrpB is subject to a high degree of antigenic variation, principally through a region of hypervariable sequence exposed on the cell surface [58, 59].

In a word, the four genes play important roles in the uptake of nutrition. So the adaptive changes in these proteins might be beneficial for *Neisseria* to survive in the host.

4. Conclusion

Our analysis reported here indicates that both homologous recombination and positive selection play important roles in the evolution of the core genome in *Neisseria*. Additionally, homologous recombination has a greater contribution to the genetic variation of a large number of genes with recombination signals. Only 10 genes were identified to be under positive selection, which also showed significant evidence of recombination. However, the positively selected genes were found to be involved in DNA processing or located on the cell membrane. The former reduce the frequency of

recombination and enables a stable genetic environment, while the latter maintain a dynamic interaction with the external environment, as well as with the host. Overall, the changes in these positively selected genes result in an improvement in bacterial fitness in response to a variety of environmental signals. These genes can be regarded as a screened gene set for further analysis of the mechanisms of adaptive evolution in *Neisseria*.

Conflict of Interests

The authors declare that they have no conflict of interests.

Authors' Contribution

Junjie Yue and Long Liang formulated the study. Dong Yu performed the research. Yuan Jin and Zhiqiu Yin analysed the data. Hongguang Ren and Wei Zhou participated in analysis and discussion. Dong Yu wrote the paper. All authors read and approved the final paper.

Acknowledgments

This work was supported by the National Key Program for Infectious Diseases of China (2011ZX10004-001), the National Basic Research Program of China (2013CB910804), and the Innovation Foundation of AMMS (No. 2012CXJJ023).

References

[1] G. Bell, *Selection: The Mechanism of Evolution*, Chapman & Hall, 1997.

[2] B. Alberts, A. Johnson, J. Lewis et al., "DNA replication, repair, and recombination," in *Molecular Biology of the Cell*, p. 845, Garland Science, 2002.

[3] J. R. Doroghazi and D. H. Buckley, "Widespread homologous recombination within and between Streptomyces species," *ISME Journal*, vol. 4, no. 9, pp. 1136–1143, 2010.

[4] S. Suerbaum, J. Maynard Smith, K. Bapumia et al., "Free recombination within Helicobacter pylori," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 21, pp. 12619–12624, 1998.

[5] B. Joseph, R. F. Schwarz, B. Linke et al., "Virulence evolution of the human pathogen neisseria meningitidis by recombination in the core and accessory genome," *PLoS ONE*, vol. 6, no. 4, Article ID e18441, 2011.

[6] T. F. Cooper, "Recombination speeds adaptation by reducing competition between beneficial mutations in populations of Escherichia coli," *PLoS Biology*, vol. 5, article e225, no. 9, 2007.

[7] Y. L. Tsai, S. B. Maron, P. McGann, K. K. Nightingale, M. Wiedmann, and R. H. Orsi, "Recombination and positive selection contributed to the evolution of *Listeria monocytogenes* lineages III and IV, two distinct and well supported uncommon *L. monocytogenes* lineages," *Infection, Genetics and Evolution*, vol. 11, no. 8, pp. 1881–1890, 2011.

[8] Y. Soyer, R. H. Orsi, L. D. Rodriguez-Rivera, Q. Sun, and M. Wiedmann, "Genome wide evolutionary analyses reveal serotype specific patterns of positive selection in selected *Salmonella* serotypes," *BMC Evolutionary Biology*, vol. 9, no. 1, article 264, 2009.

[9] T. Lefébure and M. J. Stanhope, "Evolution of the core and pangenome of Streptococcus: positive selection, recombination, and genome composition," *Genome Biology*, vol. 8, no. 5, article R71, 2007.

[10] T. Lefébure and M. J. Stanhope, "Pervasive, genome-wide positive selection leading to functional divergence in the bacterial genus Campylobacter," *Genome Research*, vol. 19, no. 7, pp. 1224–1232, 2009.

[11] Z. Xu, H. Chen, and R. Zhou, "Genome-wide evidence for positive selection and recombination in Actinobacillus pleuropneumoniae," *BMC Evolutionary Biology*, vol. 11, no. 1, article 203, 2011.

[12] L. Petersen, J. P. Bollback, M. Dimmic, M. Hubisz, and R. Nielsen, "Genes under positive selection in Escherichia coli," *Genome Research*, vol. 17, no. 9, pp. 1336–1343, 2007.

[13] R. C. Brunham, F. A. Plummer, and R. S. Stephens, "Bacterial antigenic variation, host immune response, and pathogen-host coevolution," *Infection and Immunity*, vol. 61, no. 6, pp. 2273–2276, 1993.

[14] T. C. Bruen, H. Philippe, and D. Bryant, "A simple and robust statistical test for detecting the presence of recombination," *Genetics*, vol. 172, no. 4, pp. 2665–2681, 2006.

[15] S. Sawyer, "Statistical tests for detecting gene conversion," *Molecular Biology and Evolution*, vol. 6, no. 5, pp. 526–538, 1989.

[16] Z. Yang, R. Nielsen, N. Goldman, and A. K. Pedersen, "Codon-substitution models for heterogeneous selection pressure at amino acid sites," *Genetics*, vol. 155, no. 1, pp. 431–449, 2000.

[17] Z. Yang and J. R. Bielawski, "Statistical methods for detecting molecular adaptation," *Trends in Ecology and Evolution*, vol. 15, no. 12, pp. 496–503, 2000.

[18] R. Nielsen and Z. Yang, "Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene," *Genetics*, vol. 148, no. 3, pp. 929–936, 1998.

[19] J. C. Dunning Hotopp, R. Grifantini, N. Kumar et al., "Comparative genomics of Neisseria meningitidis: core genome, islands of horizontal transfer and pathogen-specific genes," *Microbiology*, vol. 152, part 12, pp. 3733–3749, 2006.

[20] B. Joseph, S. Schneiker-Bekel, A. Schramm-Glück et al., "Comparative genome biology of a serogroup B carriage and disease strain supports a polygenic nature of meningococcal virulence," *Journal of Bacteriology*, vol. 192, no. 20, pp. 5363–5377, 2010.

[21] M. Unemo and W. M. Shafer, "Antibiotic resistance in Neisseria gonorrhoeae: origin, evolution, and lessons learned for the future," *Annals of the New York Academy of Sciences*, vol. 1230, pp. E19–E28, 2011.

[22] D. A. Caugant, "Genetics and evolution of *Neisseria meningitidis*: importance for the epidemiology of meningococcal disease," *Infection, Genetics and Evolution*, vol. 8, no. 5, pp. 558–565, 2008.

[23] J. S. Bennett, S. D. Bentley, G. S. Vernikos et al., "Independent evolution of the core and accessory gene sets in the genus *Neisseria*: insights gained from the genome of *Neisseria lactamica* isolate 020-06," *BMC Genomics*, vol. 11, no. 1, article 652, 2010.

[24] C. O. Buckee, K. A. Jolley, M. Recker et al., "Role of selection in the emergence of lineages and the evolution of virulence in Neisseria meningitidis," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 39, pp. 15082–15087, 2008.

[25] J. S. Bennett, K. A. Jolley, P. F. Sparling et al., "Species status of Neisseria gonorrhoeae: evolutionary and epidemiological inferences from multilocus sequence typing," *BMC Biology*, vol. 5, article 35, 2007.

[26] K. A. Jolley, D. J. Wilson, P. Kriz, G. McVean, and M. C. J. Maiden, "The influence of mutation, recombination, population history, and selection on patterns of genetic diversity in Neisseria meningitidis," *Molecular Biology and Evolution*, vol. 22, no. 3, pp. 562–569, 2005.

[27] C. Schoen, J. Blom, H. Claus et al., "Whole-genome comparison of disease and carriage strains provides insights into virulence evolution in Neisseria meningitidis," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 9, pp. 3473–3478, 2008.

[28] N. H. Smith, J. M. Smith, and B. G. Spratt, "Sequence evolution of the porB gene of Neisseria gonorrhoeae and Neisseria meningitidis: evidence of positive Darwinian selection," *Molecular Biology and Evolution*, vol. 12, no. 3, pp. 363–370, 1995.

[29] T. D. Andrews and T. Gojobori, "Strong positive selection and recombination drive the antigenic variation of the PilE protein of the human pathogen Neisseria meningitidis," *Genetics*, vol. 166, no. 1, pp. 25–32, 2004.

[30] "Activities at the Universal Protein Resource (UniProt)," *Nucleic Acids Research*, vol. 42, pp. D191–D198, 2014.

[31] R. C. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Research*, vol. 32, no. 5, pp. 1792–1797, 2004.

[32] M. Suyama, D. Torrents, and P. Bork, "PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments," *Nucleic Acids Research*, vol. 34, pp. W609–W612, 2006.

[33] F. Wright, "The "effective number of codons" used in a gene," *Gene*, vol. 87, no. 1, pp. 23–29, 1990.

[34] T. Ota and M. Nei, "Variance and covariances of the numbers of synonymous and nonsynonymous substitutions per site," *Molecular Biology and Evolution*, vol. 11, no. 4, pp. 613–619, 1994.

[35] J. M. Smith, "Analyzing the mosaic structure of genes," *Journal of Molecular Evolution*, vol. 34, no. 2, pp. 126–129, 1992.

[36] I. B. Jakobsen and S. Easteal, "A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences," *Computer Applications in the Biosciences*, vol. 12, no. 4, pp. 291–295, 1996.

[37] M. N. Price, P. S. Dehal, and A. P. Arkin, "Fasttree: computing large minimum evolution trees with profiles instead of a distance matrix," *Molecular Biology and Evolution*, vol. 26, no. 7, pp. 1641–1650, 2009.

[38] Z. Yang, "PAML 4: phylogenetic analysis by maximum likelihood," *Molecular Biology and Evolution*, vol. 24, no. 8, pp. 1586–1591, 2007.

[39] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society B*, vol. 57, no. 1, pp. 289–300, 1995.

[40] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, 2014.

[41] J. D. Storey and R. Tibshirani, "Statistical significance for genomewide studies," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 16, pp. 9440–9445, 2003.

[42] L. D. Bowler, Q. Y. Zhang, J. Y. Riou, and B. G. Spratt, "Interspecies recombination between the penA genes of Neisseria meningitidis and commensal Neisseria species during the emergence of penicillin resistance in N. meningitidis: natural events and laboratory simulation," *Journal of Bacteriology*, vol. 176, no. 2, pp. 333–337, 1994.

[43] J. Zhou, L. D. Bowler, and B. G. Spratt, "Interspecies recombination, and phylogenetic distortions, within the glutamine synthetase and shikimate dehydrogenase genes of Neisseria meningitidis and commensal Neisseria species," *Molecular Microbiology*, vol. 23, no. 4, pp. 799–812, 1997.

[44] E. Feil, J. Zhou, J. M. Smith, and B. G. Spratt, "A comparison of the nucleotide sequences of the adk and recA genes of pathogenic and commensal Neisseria species: evidence for extensive interspecies recombination within adk," *Journal of Molecular Evolution*, vol. 43, no. 6, pp. 631–640, 1996.

[45] E. C. Holmes, R. Urwin, and M. C. J. Maiden, "The influence of recombination on the population structure and evolution of the human pathogen Neisseria meningitidis," *Molecular Biology and Evolution*, vol. 16, no. 6, pp. 741–749, 1999.

[46] M. C. J. Maiden, J. A. Bygraves, E. Feil et al., "Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 6, pp. 3140–3145, 1998.

[47] M. Gouy and C. Gautier, "Codon usage in bacteria: correlation with gene expressivity," *Nucleic Acids Research*, vol. 10, no. 22, pp. 7055–7074, 1982.

[48] A. Carbone, F. Képès, and A. Zinovyev, "Codon bias signatures, organization of microorganisms in codon space, and lifestyle," *Molecular Biology and Evolution*, vol. 22, no. 3, pp. 547–561, 2005.

[49] H. Willenbrock and D. W. Ussery, "Prediction of highly expressed genes in microbes based on chromatin accessibility," *BMC Molecular Biology*, vol. 8, article 11, 2007.

[50] I. K. Jordan, I. B. Rogozin, Y. I. Wolf, and E. V. Koonin, "Essential genes are more evolutionarily conserved than are nonessential genes in bacteria," *Genome Research*, vol. 12, no. 6, pp. 962–968, 2002.

[51] M. Anisimova, R. Nielsen, and Z. Yang, "Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites," *Genetics*, vol. 164, no. 3, pp. 1229–1236, 2003.

[52] R. H. Orsi, Q. Sun, and M. Wiedmann, "Genome-wide analyses reveal lineage specific contributions of positive selection and recombination to the evolution of Listeria monocytogenes," *BMC Evolutionary Biology*, vol. 8, no. 1, article 233, 2008.

[53] C. J. Saveson and S. T. Lovett, "Tandem repeat recombination induced by replication fork defects in Escherichia coli requires a novel factor, RadC," *Genetics*, vol. 152, no. 1, pp. 5–13, 1999.

[54] S. T. Lovett, C. Luisi-DeLuca, and R. D. Kolodner, "The genetic dependence of recombination in recD mutants of *Escherichia coli*," *Genetics*, vol. 120, no. 1, pp. 37–45, 1988.

[55] P. S. Andersen, D. Frees, R. Fast, and B. Mygind, "Uracil uptake in Escherichia coli K-12: isolation of uraA mutants and cloning of the gene," *Journal of Bacteriology*, vol. 177, no. 8, pp. 2008–2013, 1995.

[56] D. Perkins-Balding, M. Ratliff-Griffin, and I. Stojiljkovic, "Iron transport systems in Neisseria meningitidis," *Microbiology and Molecular Biology Reviews*, vol. 68, no. 1, pp. 154–171, 2004.

[57] I. Stojiljkovic, V. Hwa, L. de Saint Martin et al., "The Neisseria meningitidis haemoglobin receptor: Its role in iron utilization

and virulence," *Molecular Microbiology*, vol. 15, no. 3, pp. 531–541, 1995.

[58] P. van der Ley, J. van der Biezen, R. Sutmuller, P. Hoogerhout, and J. T. Poolman, "Sequence variability of FrpB, a major iron-regulated outer-membrane protein in the pathogenic neisseriae," *Microbiology*, vol. 142, no. 11, part 1, pp. 3269–3274, 1996.

[59] M. Beucher and P. F. Sparling, "Cloning, sequencing, and characterization of the gene encoding FrpB, a major iron-regulated, outer membrane protein of Neisseria gonorrhoeae," *Journal of Bacteriology*, vol. 177, no. 8, pp. 2041–2049, 1995.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 439840, 7 pages
http://dx.doi.org/10.1155/2014/439840

Research Article

Novel Approach for Coexpression Analysis of E2F1–3 and MYC Target Genes in Chronic Myelogenous Leukemia

Fengfeng Wang,[1] Lawrence W. C. Chan,[1] William C. S. Cho,[2] Petrus Tang,[3] Jun Yu,[4] Chi-Ren Shyu,[5] Nancy B. Y. Tsui,[1] S. C. Cesar Wong,[1] Parco M. Siu,[1] S. P. Yip,[1] and Benjamin Y. M. Yung[1]

[1] *Department of Health Technology and Informatics, Hong Kong Polytechnic University, Lee Shau Kee Building, Hung Hom, Kowloon, Hong Kong*
[2] *Department of Clinical Oncology, Queen Elizabeth Hospital, 30 Gascoigne Road, Kowloon, Hong Kong*
[3] *Bioinformatics Center, Chang Gung University, Taoyuan 333, Taiwan*
[4] *Beijing Institute of Genomics, Chinese Academy of Sciences, Chaoyang District, Beijing 100029, China*
[5] *Department of Computer Science, Informatics Institute, University of Missouri, Columbia, MO 65211, USA*

Correspondence should be addressed to Lawrence W. C. Chan; wing.chi.chan@polyu.edu.hk

Received 26 June 2014; Accepted 23 July 2014; Published 10 August 2014

Academic Editor: Ryuji Hamamoto

Background. Chronic myelogenous leukemia (CML) is characterized by tremendous amount of immature myeloid cells in the blood circulation. E2F1–3 and MYC are important transcription factors that form positive feedback loops by reciprocal regulation in their own transcription processes. Since genes regulated by E2F1–3 or MYC are related to cell proliferation and apoptosis, we wonder if there exists difference in the coexpression patterns of genes regulated concurrently by E2F1–3 and MYC between the normal and the CML states. *Results*. We proposed a method to explore the difference in the coexpression patterns of those candidate target genes between the normal and the CML groups. A disease-specific cutoff point for coexpression levels that classified the coexpressed gene pairs into strong and weak coexpression classes was identified. Our developed method effectively identified the coexpression pattern differences from the overall structure. Moreover, we found that genes related to the cell adhesion and angiogenesis properties were more likely to be coexpressed in the normal group when compared to the CML group. *Conclusion*. Our findings may be helpful in exploring the underlying mechanisms of CML and provide useful information in cancer treatment.

1. Introduction

Chronic myelogenous leukemia (CML) is a clonal myeloproliferative disorder that is characterized by the premature circulation of many immature myeloid cells in the blood stream [1]. The incidence rate of CML is about 1-2 per 100,000 per year. CML accounts for 20% of all leukemias affecting adults with a median age of 45 to 55 years [2]. The characteristics of CML at the cellular level include increased proliferation, increased resistance to apoptosis, and alterations in adhesion properties of leukemic progenitors [1]. Recently, there are many more studies on the analysis of microarray gene expression profiles in CML. Most of them investigate the function of differentially expressed genes such as the study to explore the relationship between pathways and differentially expressed genes from untreated CML patients in the chronic phase [3]. However, few studies are available on the coexpression analysis.

Transcription factor (TF), a kind of transacting factor, plays the most vital role in the regulation of gene expression and process of signal transduction [4]. E2F family of transcription factors is important to control cellular proliferation by regulating transcription of various genes involved in DNA replication, DNA repair, mitosis, and cell cycle progression [5]. According to structure-function studies and amino acid sequence analysis, members of the E2F family can be classified into two main subclasses: activators E2F1–3 and repressors E2F4–8 [5]. The transcription activators E2F1,

2, and 3 are vital for cell cycle progression, especially in the G1/S transition process [6]. The protooncogene c-myc encodes a transcription factor (MYC) that can induce both cell proliferation and apoptosis [7]. As a transcription factor, MYC can both activate and repress transcription of target genes. High-throughput techniques have shown that MYC-activated genes are involved in growth, protein synthesis, and mitochondrial function. Most of MYC-repressed genes participate in the interaction and communication between cells and their external environment, and several genes are found to have antiproliferative or antimetastatic properties [8]. In addition, E2F1–3 and MYC are reciprocally regulated in the transcription process to form positive feedback loops among them [9].

Target genes regulated by the same TF tend to be coexpressed, and the coexpression degree is increased if genes share more TFs [10]. Moreover, coexpression analysis has been used to study functionally related genes since the coexpressed genes are more likely to participate in the similar cellular processes and pathways [11]. Furthermore, coexpressed genes are different in different states and cell types [12]. As a result, coexpression pattern analysis is a powerful strategy for grouping genes and further analyzing the underlying mechanisms of diseases. The different coexpression pattern can be regarded as the signature of a disease.

Since target genes regulated by E2F1–3 or MYC are related to cell proliferation and apoptosis, we wonder if there exists difference in the coexpression patterns of genes regulated concurrently by E2F1–3 and MYC between the normal and the CML states. In order to answer this research question, we proposed a method to explore the difference in the coexpression patterns by identifying a disease-specific cutoff point for coexpression levels that classified the coexpressed gene pairs into strong and weak coexpression classes so that the class was best coherent with the disease phenotype. Traditional methods on the coexpression analysis identify significantly coexpressed gene pairs by calculating a P value of correlation coefficient for each gene pair individually, which cannot reflect the overall difference between two different groups. Our method calculated all the correlation coefficients in each group to form two different cumulative distributions including all the gene pairs, which can identify the difference between two different groups from the overall structure. Also, the different coexpression pattern reflected the biological alterations in CML compared to the normal state. Annotation of the candidate target genes and mapping the coexpressed gene pairs to the annotated gene pairs from enriched process networks provided important information to understand the underlying mechanisms of the CML and the normal states.

2. Methods

2.1. Microarray Expression Data. Microarray technology is used to monitor the expression levels of thousands of genes in cells simultaneously [13]. Gene expression analysis across different conditions, the normal and the disease states, may contribute much to the exploration of disease mechanisms. In this study, we analyzed the microarray dataset GSE5550, normalized by variance stabilizing transformations (VSN) method, which is publicly available on the *Gene Expression Omnibus* (*GEO*) repository [3]. The data were obtained from gene expression measurements of 8,537 unique mRNAs. CD34+ hematopoietic stem and progenitor cells were collected from the bone marrow of patients with untreated CML in the chronic phase and health controls [3]. The subjects recruited for this dataset are Caucasians in Germany. The CML group consisted of nine patient samples, and the control group included eight normal samples. In this dataset, a gene may be interrogated by more than one probe. In this case, we took the average of all the probes for the same mRNA [14, 15].

2.2. Identification of Candidate Target Genes Regulated Concurrently by E2F1–3 and MYC. The interactions between TFs (E2F1, E2F2, E2F3, and MYC) and target genes (TGs) were obtained from *prediction of transcriptional regulatory modules* (*PReMod*) database [16]. TF binding sites are often clustered together, called cis-regulatory modules (CRMs). *PReMod* database predicts relationships between TFs and their TGs based on the binding affinity and conservation of CRM. It consists of more than 100,000 computationally predicted modules within the human genome [16]. These modules give a description of 229 potential transcription factor families and are the first genome-wide collection of predicted regulatory modules for the human genome [17]. In this study, we called the set of TF binding predictions (TF-TG pairs) from *PReMod* a molecular interaction set. This set was regarded as the reference data. After obtaining the TGs of each TF (E2F1, E2F2, E2F3, and MYC) individually, we identified the common TGs of these four TFs, which were regarded as the candidate target genes for further analysis. The flowchart is shown in Figure S1 in Supplementary Material available online at http://dx.doi.org/10.1155/2014/439840.

2.3. Coexpression Measure. We chose Pearson correlation coefficient as the similarity measure. It is represented by the direction cosine between two vectors normalized by the subtraction of their own means, and its value accounts for the angle between two feature vectors instead of the vector lengths. Moreover, Pearson correlation coefficient numerically indicates the biological relationship of two genes but does not vary with the magnitudes of their expression profiles [11, 18]. In general, similarity measure is a kernel function between two feature vectors. In this study, each feature vector consisted of the expression intensity of a gene across all the samples in the normal group or the CML group, respectively. The correlation coefficient of any two genes among the candidate target genes was calculated. We took the absolute value of correlation coefficient ($|r|$) since the coexpression measure output a scalar in the range from 0 to 1 where a high output indicated a strong biological relationship in either positive or negative direction, and a low output indicated a weak biological relationship. The coexpression level was denoted by $C_d(i, j)$ if two expression profiles were extracted from samples of the disease (CML) group and

$C_n(i, j)$ for the normal group, shown in Formulas (1) as follows:

$$C_d(i,j) = \left|\mathrm{cor}\left(x_{di}, x_{dj}\right)\right|,$$
$$C_n(i,j) = \left|\mathrm{cor}\left(x_{ni}, x_{nj}\right)\right|, \quad (1)$$

where $C_d(i, j)$ and $C_n(i, j)$ are defined as the absolute values of correlation coefficients between the expression profiles of genes i and j in the CML group and the normal group, respectively [18]; x_{di} and x_{dj} are the expression profiles of the ith and jth genes in the CML group; x_{ni} and x_{nj} are the expression profiles of the ith and jth genes in the normal group; $\mathrm{cor}(x_{di}, x_{dj})$ and $\mathrm{cor}(x_{ni}, x_{nj})$ are the Pearson correlation coefficients between them in the CML group and the normal group, respectively.

2.4. Classification of Coexpressed Gene Pairs. There was a set of correlation coefficients in either the normal group or the CML group. The two sets of correlation coefficients formed two cumulative distributions. We applied two-sample Kolmogorov-Smirnov (KS) test to identify the difference in the overall distributions of these two conditions (C_d and C_n), including all the gene pairs. The maximum deviation between two cumulative distributions of C_d and C_n was identified (Formulas (2)), at which a threshold was found to classify the coexpressed gene pairs into strong and weak coexpression classes, called the disease-specific cutoff point (C). The cutoff point represented a coexpression level, at which F_d and F_n were extremely deviated. Gene pairs were further classified into four coexpression classes: (i) strongly coexpressed gene pairs in the normal group: pairs with coexpression levels ($|r|$ values) bigger than or equal to C in the normal group; (ii) strongly coexpressed gene pairs in the CML group: pairs with coexpression levels ($|r|$ values) bigger than or equal to C in the CML group; (iii) weakly coexpressed gene pairs in the normal group: pairs with coexpression levels ($|r|$ values) smaller than C in the normal group; and (iv) weakly coexpressed gene pairs in the CML group: pairs with coexpression levels ($|r|$ values) smaller than C in the CML group. Chi-square test was used to determine if the proportions of strongly and weakly coexpressed gene pairs significantly differed between the normal and the CML groups

$$D = \max_C \left|F_d(C) - F_n(C)\right|,$$
$$F_d(C) = \mathrm{Prob}\left(C_d \geq C\right), \quad (2)$$
$$F_n(C) = \mathrm{Prob}\left(C_n \geq C\right),$$

where F_d and F_n are the cumulative distribution functions (CDFs) of C_d and C_n, respectively; D is the maximum deviation; C is the cutoff point.

We further identified the specifically coexpressed gene pairs in different groups. Each type of gene pair represented a particular biological meaning. The normal-specific strongly coexpressed gene pairs were the gene pairs strongly coexpressed only in the normal group, which were regarded as the potential molecular interactions maintaining physiological balance in healthy individuals, and the impairment of these connections may lead to diseases. Obviously, these pairs were the CML-specific weakly coexpressed gene pairs, which were weakly coexpressed only in the CML group. The CML-specific strongly coexpressed gene pairs were the gene pairs strongly coexpressed only in the CML group, which represented the characteristics of the disease and may be the pathogenic alternatives when the corresponding normal-specific gene pairs cannot be coexpressed for responding to stress. Similarly, these pairs were regarded as the normal-specific weakly coexpressed gene pairs.

2.5. Functional Annotation for Candidate Target Genes. We applied *MetaCore* from GeneGo Inc. to annotate the candidate target genes. Specifically, when we uploaded the candidate target genes from Section 2.2 into this database, it mapped these genes to a set of cellular and molecular process networks, which are defined and annotated by Thomson Reuters scientists. In *MetaCore*, each process is defined as a preset network describing the protein interactions among them. In each process network, the annotated target genes were those genes included in both Section 2.2 and this process network. Enrichment analysis for a process network in *MetaCore* is performed based on the P value of hypergeometric intersection between the uploaded candidate target genes and the process-related genes in this database. The lower the P value is obtained, the higher the relevance of this process network to the candidate target genes and the rating of this process network are indicated. Only the top 10 statistically enriched process networks are shown according to the sorted P values in *MetaCore*.

2.6. Mapping Coexpressed Gene Pairs to Annotated Gene Pairs. The annotated target genes in each process network were paired with all the possible combinations to form the annotated gene pairs. The annotated gene pairs from each process network were mapped to the coexpressed gene pairs identified in Section 2.4: the normal-specific strongly coexpressed, the normal-specific weakly coexpressed, the CML-specific strongly coexpressed, and the CML-specific weakly coexpressed gene pairs. We applied Fisher's exact test to identify if there were more mapped normal-specific strongly coexpressed gene pairs than mapped CML-specific strongly coexpressed gene pairs in each process network. In other words, we planned to identify if these genes were more likely to be coexpressed in the normal group compared to the CML group. As a result, one-sided P value was chosen. False discovery rates (FDRs) are usually used to control the expected proportion of false positives for the multiple hypotheses. In this study, the FDRs were calculated based on the P values obtained from Fisher's exact test [19]. A process network was significantly mapped, if its FDR value was smaller than 0.05 [20]. The FDR values were estimated via the *Matlab* function, *mafdr* [21].

3. Results

3.1. Identification of Structural Coexpression Difference. In total, we identified 217 common TGs of E2F1–3 and MYC

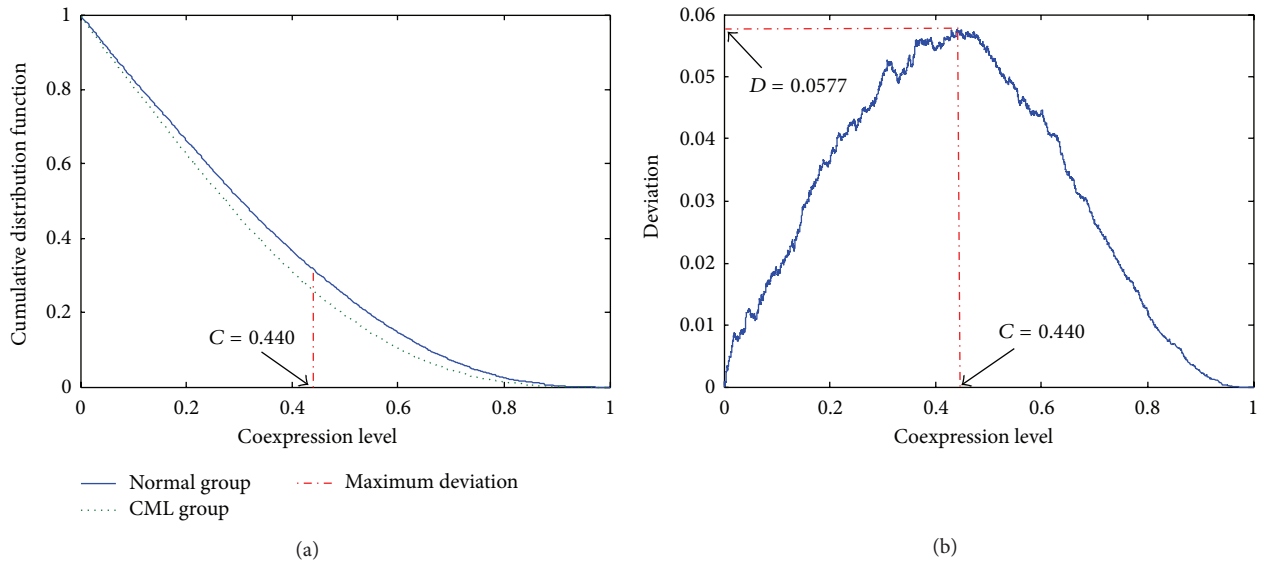


Figure 1: Plots of distributions for coexpression analysis. (a) Cumulative distribution functions of coexpression levels in the normal and the CML groups. (b) Deviation distribution against different coexpression cutoff points.

Table 1: The coexpressed gene pairs identified by the disease-specific cutoff point.

Group	Number of strongly coexpressed pairs	Number of weakly coexpressed pairs
Normal	7436	16000
CML	6083	17353

that can be found in the microarray dataset GSE5550 (Table S1). We further extracted the available expression profiles of these TGs and calculated the correlation coefficients in both the normal and the CML groups. In each group, there was a set of correlation coefficients of 23,436 gene pairs. We plotted the cumulative distributions of these two sets of data. The distributions between the normal and the CML groups were significantly different (P value = 2.00×10^{-34} for D = 0.0577). The disease-specific cutoff point that classified the coexpressed gene pairs into strong and weak coexpression classes was $C = 0.440$ (Figure 1(a)). Figure 1(b) illustrates that the deviation was small at the two extremes, and the peak ($D = 0.0577$) was found at the disease-specific cutoff point. Two coexpression patterns were so distinct that the normal group had more strongly coexpressed (level above ~0.440) and less weakly coexpressed (level below ~0.440) gene pairs than those in the CML group (Figure 1(a)). The cutoff point classified the gene pairs into four coexpression classes, shown in Table 1. The number of strongly coexpressed gene pairs in the normal group (7436) was larger than that in the CML group (6083). Chi-square test indicated that the proportions of strongly and weakly coexpressed gene pairs significantly differed between the normal and the CML groups (P value = 2.74×10^{-43} for $\chi^2 = 190$).

3.2. MetaCore Analysis for Enriched Process Networks. The top 10 statistically enriched process networks for functional annotation of the 217 candidate target genes are shown in Table S2. All the P values for hypergeometric intersection test were smaller than 0.05. We got the annotated target genes involved in each process network and mapped the annotated gene pairs to the coexpressed gene pairs. Fisher's exact test was used to identify if there were more mapped normal-specific strongly coexpressed gene pairs than mapped CML-specific strongly coexpressed gene pairs in each process network. The results showed that 8 out of 10 process networks had more mapped normal-specific strongly coexpressed gene pairs (Table 2). Fisher's exact test demonstrated that "*Cell adhesion_Attractive and repulsive receptors*" and "*Development_Regulation of angiogenesis*" process networks were significantly mapped (P values = 0.001 and 0.012, <0.05, and FDR values were 0.004 and 0.026, <0.05).

We further plotted the coexpression networks for the mapped normal-specific strongly coexpressed gene pairs (a = 6 and 8) (Figure 2). Both "*Cell adhesion_Attractive and repulsive receptors*" and "*Development_Regulation of angiogenesis*" process networks had ephrin-B2 (EFNB2), ephrin-A5 (EFNA5), and EPH receptor A4 (EPHA4) (Figure S2). From *National Center for Biotechnology Information* (*NCBI*) database, we obtained the basic information for these genes/proteins. EFNB2 and EFNA5 are the members of the ephrin gene family. EPHA4 protein product is an ephrin receptor. The ephrins (EPH) and EPH-related receptors belong to the largest subfamily of receptor protein-tyrosine kinases, which play a vital role in mediating developmental events. Figure 2 shows that the connection from EFNA5 to EPHA4 was identified as a strongly coexpressed gene pair for these two process networks in the normal group. In addition, protein products from neuropilin 2 (NRP2), transforming growth factor, beta receptor II (TGFBR2), and somatostatin receptor 2 (SSTR2) also belong to receptors, which are very important in signal transduction process. The encoded protein from integrin, alpha 2 (ITGA2), plays a vital

Table 2: Mapping coexpressed gene pairs to annotated gene pairs from each process network.

Process networks	Fisher's exact test					FDR
	a	*b*	*c*	*d*	*P* value	
Development_Neurogenesis in general	14	11	11	14	0.286	0.251
Development_Hedgehog signaling	22	15	15	22	0.081	0.118
Signal transduction_WNT signaling	10	7	7	10	0.247	0.270
Signal transduction_TGF-beta, GDF, and activin signaling	6	5	5	6	0.500	0.365
Cell adhesion_Attractive and repulsive receptors	**6**	**0**	**0**	**6**	***0.001***	***0.004***
Development_Regulation of angiogenesis	**8**	**2**	**2**	**8**	***0.012***	***0.026***
Cardiac development_BMP_TGF_beta_signaling	2	1	1	2	0.500	0.313
Neurophysiological process_Melatonin signaling	3	2	2	3	0.500	0.274

a: mapped normal-specific strongly coexpressed gene pairs; *b*: mapped normal-specific weakly coexpressed gene pairs; *c*: mapped CML-specific strongly coexpressed gene pairs; *d*: mapped CML-specific weakly coexpressed gene pairs.

role in leukocyte intercellular adhesion process. There were three enzymes identified in the coexpression networks: (i) the protein encoded by protein kinase, cAMP-dependent, catalytic, beta (PRKACB) is a protein kinase; (ii) the protein product from prolyl endopeptidase (PREP) is a protease; and (iii) the protein encoded by HIV-1 Tat interactive protein 2 (HTATIP2) is an oxidoreductase required for tumor suppression. From the results, we can infer that these genes/proteins were well connected with each other to transduce signals and maintain physiological balance in healthy individuals. However, in the CML group, these connections were impaired.

4. Discussion and Conclusion

In this study, our developed method successfully identified the difference in the coexpression patterns of those candidate target genes regulated concurrently by E2F1–3 and MYC between the normal and the CML groups from the overall structure (Figure 1). We further found that genes involved in the cell adhesion and angiogenesis properties were more likely to be coexpressed in the normal group compared to the CML group (Table 2 and Figure 2). The alteration in adhesion properties of leukemic progenitors is one CML characteristic at the cellular level [1]. In addition, Bhatia et al. hypothesized that decreased integrin-mediated adhesion of CML progenitors to stroma can lead to continuous cell proliferation [22]. They treated the cells with interferon-α (IFN-α). The results showed that the treatment restored the CML progenitor adhesion to stroma and also the regulation of CML progenitor proliferation [22]. Angiogenesis is the process forming new blood from the preexisting vasculature, including degradation of extracellular matrix proteins, as well as activation, proliferation, and migration of endothelial cells [23]. In leukemia, hematopoietic cells are supported from the normal vascular bed in bone marrow [23]. Increased vascularity was found in acute myeloid leukemia (AML) patients [24]. Importantly, in CML, the number of blood vessels and vascular areas were found to be increased when compared to control bone marrows [23]. Our results showed that the connection from EFNA5 to EPHA4 was identified as a strongly coexpressed gene pair in the normal group ($|r|$ values were 0.720 and 0.013 in the normal group and the CML group, resp.) (Figure 2). Ephrin-A receptors belong to the largest subfamily of receptor tyrosine kinases that regulate cell shape, mobility, and attachment [25]. Interactions between Ephrin-A receptors and ligands are important in cell-cell communication, initiating unique bidirectional signaling cascades to transduce the information [26]. There may be some relationships between adhesion property and angiogenesis. These two process networks were found to be well controlled in the normal group compared to the CML group. Dysregulation of adhesion and angiogenesis properties is a possible reason leading to CML.

The advantage of our study is the application of coexpression analysis to target genes regulated concurrently by more than one transcription factor under different conditions. We identified different coexpression patterns between the normal and the CML groups. A limitation for differential expression analysis is that it only reflects the upregulation or downregulation of existing components in the well-known pathways under the normal or the disease condition, which cannot identify the functionally associated linkages among genes during signal transduction. In addition, differential expression analysis does not take account of the level of correlations that may exist between gene expression patterns [12]. Coexpression analysis is useful for analyzing the underlying mechanisms of diseases. Moreover, the different coexpression pattern can be regarded as the signature of a disease.

Several methods have been proposed to analyze coexpressed genes. The two-stage screening procedure was applied to select statistically and biologically significant gene pairs in Zhu et al.'s study [27]. Gupta et al. proposed a method for determining the correlation threshold using the clustering coefficient. R^2 metric was used as a measure of similarity between two genes [28]. Previous studies cannot reflect the overall difference between two different groups. Our method calculated all the correlation coefficients in each group (the normal group and the CML group) to form two distributions, which can find the difference between two different groups from the overall structure.

In summary, we have presented a detailed method to identify a disease-specific cutoff point for coexpression levels

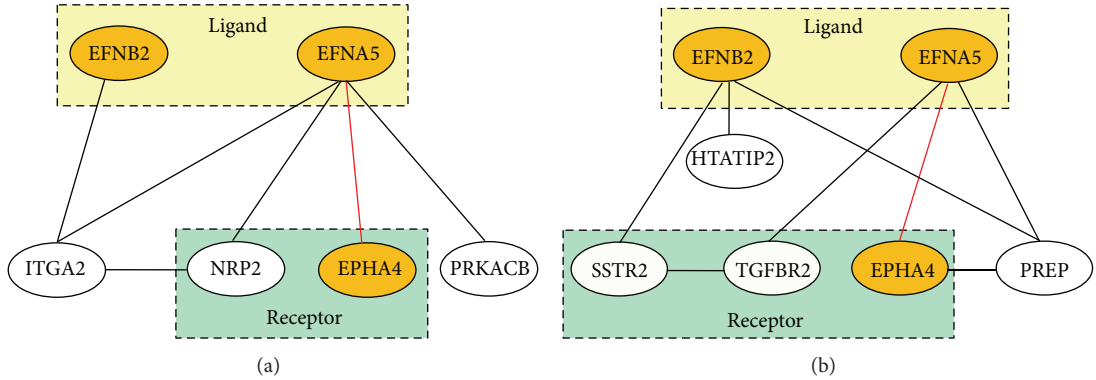


FIGURE 2: Coexpression networks for the mapped normal-specific strongly coexpressed gene pairs. The yellow ellipses are those genes found in both process networks. (a) Mapped normal-specific strongly coexpressed gene pairs in the "*Cell adhesion_Attractive and repulsive receptors*" process network. (b) Mapped normal-specific strongly coexpressed gene pairs in the "*Development_Regulation of angiogenesis*" process network.

that classified the coexpressed gene pairs into strong and weak coexpression classes so that the class was best coherent with the disease phenotype. We applied this method to explore the difference in the coexpression patterns of target genes regulated concurrently by E2F1–3 and MYC between the normal and the CML groups. Our method effectively identified the statistical differences between the normal and the CML groups from the overall structure. We further found the potentially altered cell adhesion and angiogenesis properties in the CML state when compared to the normal group. The different coexpression pattern can reflect the biological alterations in CML. Our significant findings will be helpful in exploring the underlying mechanisms of CML and provide useful information in cancer treatment.

Conflict of Interests

The authors declare that there is no conflict of interests.

Acknowledgments

This work was supported by Internal Grants of Hong Kong Polytechnic University (G.55.09.YL61) and (1-ZE17). The authors thank Dr. Thomas Lui for his advice on the analyses using *Matlab*. The *MetaCore* license was provided by the Bioinformatics Center from Chang Gung University.

References

[1] S. Salesse and C. M. Verfaillie, "Mechanisms underlying abnormal trafficking and expansion of malignant progenitors in CML: BCR/ABL-induced defects in integrin function in CML," *Oncogene*, vol. 21, no. 56, pp. 8605–8611, 2002.

[2] R. Frazer, A. E. Irvine, and M. F. McMullin, "Chronic myeloid leukaemia in the 21st century," *Ulster Medical Journal*, vol. 76, no. 1, pp. 8–17, 2007.

[3] E. Diaz-Blanco, I. Bruns, F. Neumann et al., "Molecular signature of $CD34^+$ hematopoietic stem and progenitor cells of patients with CML in chronic phase," *Leukemia*, vol. 21, no. 3, pp. 494–504, 2007.

[4] Q. Cui, Z. Yu, Y. Pan, E. O. Purisima, and E. Wang, "MicroRNAs preferentially target the genes with high transcriptional regulation complexity," *Biochemical and Biophysical Research Communications*, vol. 352, no. 3, pp. 733–738, 2007.

[5] C. Timmers, N. Sharma, R. Opavsky et al., "E2f1, E2f2, and E2f3 control E2F target expression and cellular proliferation via a p53-dependent negative feedback loop," *Molecular and Cellular Biology*, vol. 27, no. 1, pp. 65–78, 2007.

[6] L. Wu, C. Timmers, B. Malti et al., "The E2F1-3 transcription factors are essential for cellular proliferation," *Nature*, vol. 414, no. 6862, pp. 457–462, 2001.

[7] S. Pelengaris, B. Rudolph, and T. Littlewood, "Action of Myc in vivo—Proliferation and apoptosis," *Current Opinion in Genetics & Development*, vol. 10, no. 1, pp. 100–105, 2000.

[8] S. Pelengaris and M. Khan, "The many faces of c-MYC," *Archives of Biochemistry and Biophysics*, vol. 416, no. 2, pp. 129–136, 2003.

[9] H. A. Coller, J. J. Forman, and A. Legesse-Miller, "'Myc'ed messages': myc induces transcription of E2F1 while inhibiting its translation via a microRNA polycistron," *PLoS Genetics*, vol. 3, no. 8, p. e146, 2007.

[10] H. Yu, N. M. Luscombe, J. Qian, and M. Gerstein, "Genomic analysis of gene expression relationships in transcriptional regulatory networks," *Trends in Genetics*, vol. 19, no. 8, pp. 422–427, 2003.

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

[12] A. Torkamani, B. Dean, N. J. Schork, and E. A. Thomas, "Coexpression network analysis of neural tissue reveals perturbations in developmental processes in schizophrenia," *Genome Research*, vol. 20, no. 4, pp. 403–412, 2010.

[13] S. Dudoit, Y. H. Yang, M. J. Callow, and T. P. Speed, "Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments," *Statistica Sinica*, vol. 12, no. 1, pp. 111–139, 2002.

[14] T. Breslin, M. Krogh, C. Peterson, and C. Troein, "Signal transduction pathway profiling of individual tumor samples," *BMC Bioinformatics*, vol. 6, article 163, 2005.

[15] A. V. Kapp, S. S. Jeffrey, A. Langerød et al., "Discovery and validation of breast cancer subtypes," *BMC Genomics*, vol. 7, article 231, 2006.

[16] V. Ferretti, C. Poitras, D. Bergeron, B. Coulombe, F. Robert, and M. Blanchette, "PReMod: a database of genome-wide mammalian cis-regulatory module predictions," *Nucleic Acids Research*, vol. 35, no. 1, pp. D122–D126, 2007.

[17] M. Blanchette, A. R. Bataille, X. Chen et al., "Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression," *Genome Research*, vol. 16, no. 5, pp. 656–668, 2006.

[18] S. Horvath and J. Dong, "Geometric interpretation of gene coexpression network analysis," *PLoS Computational Biology*, vol. 4, no. 8, Article ID e1000117, 2008.

[19] J. D. Storey, "A direct approach to false discovery rates," *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, vol. 64, no. 3, pp. 479–498, 2002.

[20] J. Fu, W. Tang, P. Du et al., "Identifying MicroRNA-mRNA regulatory network in colorectal cancer by a combination of expression profile and bioinformatics analysis," *BMC Systems Biology*, vol. 6, article 68, 2012.

[21] W. E. Haskins, K. Petritis, and J. Zhang, "MRCQuant-an accurate LC-MS relative isotopic quantification algorithm on TOF instruments," *BMC Bioinformatics*, vol. 12, no. 74, 2011.

[22] R. Bhatia, J. B. McCarthy, and C. M. Verfaillie, "Interferon-α restores normal β1 integrin-mediated inhibition of hematopoietic progenitor proliferation by the marrow microenvironment in chronic myelogenous leukemia," *Blood*, vol. 87, no. 9, pp. 3883–3891, 1996.

[23] A. Aguayo, H. Kantarjian, T. Manshouri et al., "Angiogenesis in acute and chronic leukemias and myelodysplastic syndromes," *Blood*, vol. 96, no. 6, pp. 2240–2245, 2000.

[24] J. W. Hussong, G. M. Rodgers, and P. J. Shami, "Evidence of increased angiogenesis in patients with acute myeloid leukemia," *Blood*, vol. 95, no. 1, pp. 309–313, 2000.

[25] J. P. Himanen, N. Saha, and D. B. Nikolov, "Cell-cell signaling via Eph receptors and ephrins," *Current Opinion in Cell Biology*, vol. 19, no. 5, pp. 534–542, 2007.

[26] J. Himanen and D. B. Nikolov, "Eph receptors and ephrins," *International Journal of Biochemistry and Cell Biology*, vol. 35, no. 2, pp. 130–134, 2003.

[27] D. Zhu, A. O. Hero, H. Cheng, R. Khanna, and A. Swaroop, "Network constrained clustering for gene microarray data," *Bioinformatics*, vol. 21, no. 21, pp. 4014–4020, 2005.

[28] A. Gupta, C. D. Maranas, and R. Albert, "Elucidation of directionality for co-expressed genes: Predicting intra-operon termination sites," *Bioinformatics*, vol. 22, no. 2, pp. 209–214, 2006.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 765320, 10 pages
http://dx.doi.org/10.1155/2014/765320

Research Article

The Effects of the Context-Dependent Codon Usage Bias on the Structure of the nsp1α of Porcine Reproductive and Respiratory Syndrome Virus

Yao-zhong Ding,[1,2] Ya-nan You,[1] Dong-jie Sun,[1] Hao-tai Chen,[1,2] Yong-lu Wang,[1,2] Hui-yun Chang,[1,2] Li Pan,[1,2] Yu-zhen Fang,[1,2] Zhong-wang Zhang,[1,2] Peng Zhou,[1,2] Jian-liang Lv,[1,2] Xin-sheng Liu,[1,2] Jun-jun Shao,[1,2] Fu-rong Zhao,[1,2] Tong Lin,[1,2] Laszlo Stipkovits,[3] Zygmunt Pejsak,[4] Yong-guang Zhang,[1,2] and Jie Zhang[1,2]

[1] *State Key Laboratory of Veterinary Etiological Biology, National Foot-and-Mouth Disease Reference Laboratory, Lanzhou Veterinary Research Institute, Chinese Academy of Agricultural Sciences, Lanzhou, Gansu 730046, China*
[2] *Jiangsu Co-Innovation Center for Prevention and Control of Important Animal Infectious Diseases and Zoonoses, Yangzhou, Jiangsu 225009, China*
[3] *RT-Europe Research Center Ltd., Budapest, Hungary*
[4] *Department of Swine Diseases, National Veterinary Research Institute, 57 Partyzantow, 24-100 Puławy, Poland*

Correspondence should be addressed to Yong-guang Zhang; zhangyongguang@caas.cn and Jie Zhang; zhangjie03@caas.cn

Received 18 March 2014; Revised 5 June 2014; Accepted 19 June 2014; Published 3 August 2014

Academic Editor: Hongwei Wang

The information about the crystal structure of porcine reproductive and respiratory syndrome virus (PRRSV) leader protease nsp1α is available to analyze the roles of tRNA abundance of pigs and codon usage of the *nsp1α* gene in the formation of this protease. The effects of tRNA abundance of the pigs and the synonymous codon usage and the context-dependent codon bias (CDCB) of the *nsp1α* on shaping the specific folding units (α-helix, β-strand, and the coil) in the nsp1α were analyzed based on the structural information about this protease from protein data bank (PDB: 3IFU) and the *nsp1α* of the 191 PRRSV strains. By mapping the overall tRNA abundance along the *nsp1α*, we found that there is no link between the fluctuation of the overall tRNA abundance and the specific folding units in the nsp1α, and the low translation speed of ribosome caused by the tRNA abundance exists in the *nsp1α*. The strong correlation between some synonymous codon usage and the specific folding units in the nsp1α was found, and the phenomenon of CDCB exists in the specific folding units of the nsp1α. These findings provide an insight into the roles of the synonymous codon usage and CDCB in the formation of PRRSV nsp1α structure.

1. Introduction

Porcine reproductive and respiratory syndrome virus (PRRSV) is an economically important pathogen of swine. The PRRSV belongs to the order Nidovirales, family Arteriviridae, genus *Arterivirus* [1]. The PRRSV genome contains at least 9 open reading frames, including ORF1a encoding papain-like cysteine protease, ORF1b encoding RNA dependent RNA polymerase, ORFs 2–6 encoding envelop proteins, and ORF7 encoding the nucleocapsid protein [2, 3]. PRRSV strains can be divided into two distinct serotypes, namely, the North American isolate (US) and the European isolate (EU) [4–8].

The replicative enzymes of the PRRSV are encoded in ORF1a and ORF1b, which locate in the 5′ proximal three quarters of the viral genome. The two polyproteins encoded by ORF1a and ORF1b are cleaved extensively by the nonstructural protein 4 (nsp4) deriving from ORF1a, yielding a series of nonstructural proteins [9]. In particular, the nsp1 and the nsp2 proteases release themselves from the ORF1a polyprotein firstly, and the nsp1 can be further processed into two multifunctional proteases, namely, the nsp1α and

the nsp1β [10, 11]. The arterivirus nsp1 region contains a tandem of papain-like autoprotease domains (PCPα and PCPβ), and the arterivirus PCPα and PCPβ domains were found to be active in the reticulocyte lysates and the *E. coli* systems [12, 13]. This biological feature might indicate that the active functions of PCPα and PCPβ are free from the different types of the expression systems and depend on the correct folding by themselves. As for the nsp1α, it plays an important role in regulating the accumulation of both genome- and subgenome-length minus-strand RNA and thereby fine-tuning the relative abundance of each of viral mRNAs in the infected cells [10, 14, 15]. The correct secondary structure of the nsp1α is required for the biological functions of the protease. Based on the crystal structure of the nsp1α, it was found that this nonstructural protein has three domains, namely, the N-terminal zinc finger (ZF) domain, the papain-like cysteine protease domain, and the carboxyl-terminal extension [16]. Recently, the role of the nsp1α in impairing the host immune response has been reported [17]; however, little information about the relationship between synonymous codon usage and the secondary structure of the PRRSV nsp1α is available to date.

The synonymous codon usage and translational speed of gene play important roles in many biological functions, like translation efficiency, genetic diversity, amino acid conservation, transfer RNA abundance, coevolution of the virus and its hosts, and context-dependent codon bias (CDCB), and so forth [18–22]. The nucleotide composition of a coding sequence (CDS) is nonrandom, and the CDS nonrandomness is influenced by the preferences in the selection of synonymous codons pairing to the same amino acid (termed as the synonymous codon usage bias SCUB). The link between SCUB and specific folding unit of protein gives us a new insight into the correct formation of the secondary structure of proteins [23–26]. It is noted that mRNA sequences generally have an additional potential to carry correct structural information in the forms of SCUB, which can be involved in a single codon or a nucleotide context of the target coding sequence [27, 28]. As for SCUB, neighboring nucleotides flanking a codon regulate the usage of the specific codon from the synonymous family, termed as context-dependent codon bias (CDCB) [20, 29–31]. It has been reported that the most important nucleotide determining CDCB is the first nucleotide after a codon, termed as the N_1 context [32]. Although several evidences indicate the link between SCUB and the formation of the specific folding unit of viral protein, little information about the role of CDCB in the formation of the specific folding unit is reported up to date. In this study, we employed the structural information about the nsp1α of PRRSV and several simple formulas to analyze the relationship between the CDCB of the PRRSV nsp1α gene and the protease.

2. Materials and Methods

2.1. Information of PRRSV Gene and Structure of the nsp1α. The 191 coding sequences of PRRSV containing the *nsp1α* gene were downloaded from the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/Genbank/) and the accession numbers of the sequences were listed in Table S1 available online at http://dx.doi.org/10.1155/2014/765320. To investigate SCUB of the *nsp1α*, the related genes were obtained from these 191 coding sequences by the multiple sequence alignments performed with the Clustal W (1.7) computer programs [33]. The information about the secondary structure of the PRRSV nsp1α was obtained from protein data bank (PDB: 3IFU).

2.2. Analysis of the Overall tRNA Abundance of Each Codon Position along the nsp1α Gene. To identify the translation selection caused by the various tRNA copy numbers (reflecting tRNA abundance) of the pigs (http://gtrnadb.ucsc.edu/) at each codon position in the PRRSV *nsp1α*, we devised an index (C value) representing the overall tRNA abundance for a particular codon position in a target gene. Consider

$$C = \sqrt[n]{\prod_{1}^{n}\left(\frac{Wij}{Wj}\right)}, \tag{1}$$

where C value indicates the overall tRNA abundance for a particular codon position in the target gene, W_{ij} represents the tRNA copy numbers of a synonymous codon (i) for the corresponding amino acid (j), W_j represents the optimal tRNA copy numbers of a synonymous codon for the same amino acid, and n means the number of the interesting gene. The C value ranges from 0 to 1.0. The C value less than 0.3 for a codon position represents low tRNA abundance, and the C value more than 0.7 for a codon position represents high tRNA abundance.

2.3. Estimation of the Relationship between the Synonymous Codon Usage Bias and the Secondary Structure of the nsp1α. Based on the alignment between the amino acid sequences of the PRRSV (PDB: 3IFU) and the 191 nsp1α genes involved in this study, we can locate the different folding units in the target protein. We devised the formula for the P value based on the previous research which analyzed the relationship between the codon usage bias and the structure of the target protein [25]. Consider

$$\begin{aligned} P &= \ln\frac{f_{\text{obs}}}{f_{\text{exp}}}, \\ f_{\text{obs}} &= \frac{N_{(i,\text{sec-}k)}}{N_{(k)}}, \\ f_{\text{exp}} &= \frac{\sum N_{(i,\text{sec-}j)}}{N_{\text{total}}}, \end{aligned} \tag{2}$$

where $N_{(i,\text{sec-}k)}$ represents the amount of a specific synonymous codon for the corresponding amino acid in a specific folding unit (the α-helix, the β-strand, or the coil) of protein; sec-k represents the corresponding amino acid in a specific secondary unit; $N_{(k)}$ represents the amount of the amino acid in the corresponding folding unit. In addition, $\sum N_{(i,\text{sec-}j)}$ represents the total number of amino acids in a specific

folding unit; sec-j contains the three kinds of folding unit, namely, α-helix, β-strand, and the coil; N_{total} represents the total number of codons in the target genes. When the P value is more than zero, the corresponding synonymous codon (i) owns a potential to be selected in a specific folding unit. When the P value is less than zero, the synonymous codon (i) has no tendency to be chosen in a specific folding unit. Furthermore, we defined that when the P value is more than 0.1, the synonymous codon has a strong ability to exist in the specific folding unit; on the contrary, when the P value is less than −0.1, the synonymous codon has a strong tendency to avoid the specific folding unit.

2.4. Calculation of the Relative Abundance of Codons with Context. With the purpose to estimate the synonymous codons playing an important role in the formation of the specific folding units, codons having a significant tendency to exist in the specific folding unit of the PRRSV *nsp1α* were analyzed by the formula for the relative abundance of codons with context. Berg and Silva [32] defined that the context N_1 represents the first nucleotide after the target codon. Following this notation, we defined that the context ${}_1N$ represents the last nucleotide before the target codon. We devised a formula calculating R value for the context $N_1(xyz \sim n)$ and the context ${}_1N(n \sim xyz)$ depending on the formula previously reported [20, 34]. Consider

$$R\left(xyz{\sim}n\right) = \frac{F\left(xyz\right)}{F\left(xyz\right)F\left(n\right)},$$
$$R\left(n{\sim}xyz\right) = \frac{F\left(n{\sim}xyz\right)}{F\left(xyz\right)F\left(n\right)}, \tag{3}$$

where $F(xyz)$ is the frequency of the codon xyz and $F(n)$ is the frequency of nucleotide n in the N_1 or ${}_1N$ context. $F(xyz \sim n)$ and $F(n \sim xyz)$ are the frequency of a codon with the n context. It is noted that x, y, z, and n are the nucleotides (a, u, g, or c) and the codon is composed of xyz. Here and elsewhere the tilde character (∼) separates codons (italic) or oligonucleotides (nonunderlined) from their mononucleotide context.

2.5. Calculation of the Relative Abundance of Mononucleotide and Dinucleotides in the nsp1α Gene. To investigate whether the N_1 and ${}_1N$ contexts are shaped by randomness or not, we calculated the frequencies of each nucleotide $F(n)$ and dinucleotide $F(xy)$, where n, x, y, and z are each one of the four nucleotides (a, u, c, and g). Then we calculated the relative abundances (r value) of the mononucleotide and dinucleotides with a single nucleotide context: $r(n \sim x) = F(n \sim x)/[F(n)F(x)]$, for mononucleotide x with context n; $r(xy \sim n) = F(xy \sim n)/[F(xy)F(n)]$, for dinucleotide$xy$ with context n.

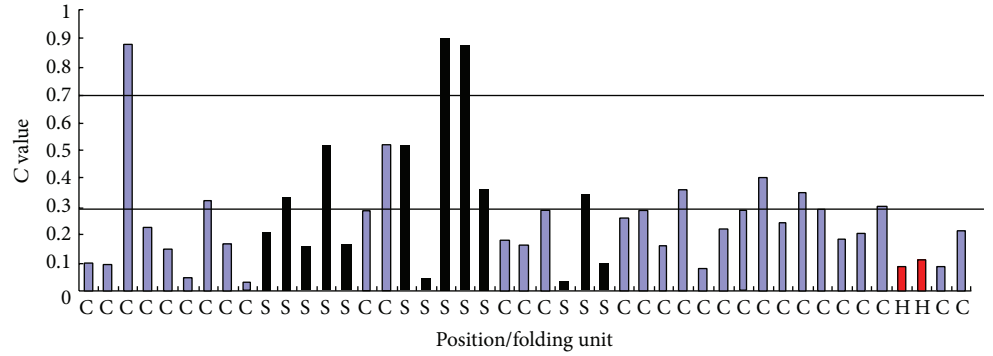
2.6. Statistic Analysis. One-way analysis of variance, namely, one-way ANOVA, is a technique used to compare means of two or more samples. In this study, the ANOVA test is applied for identifying whether the overall tRNA abundance of positions of a specific folding unit is different from other specific folding units or not. In addition, the ANOVA test is also employed to estimate whether the frequencies of codon usage in a specific folding unit are different from other specific folding units or not. This statistic analysis is carried out by the software SPSS 11.5.

3. Results

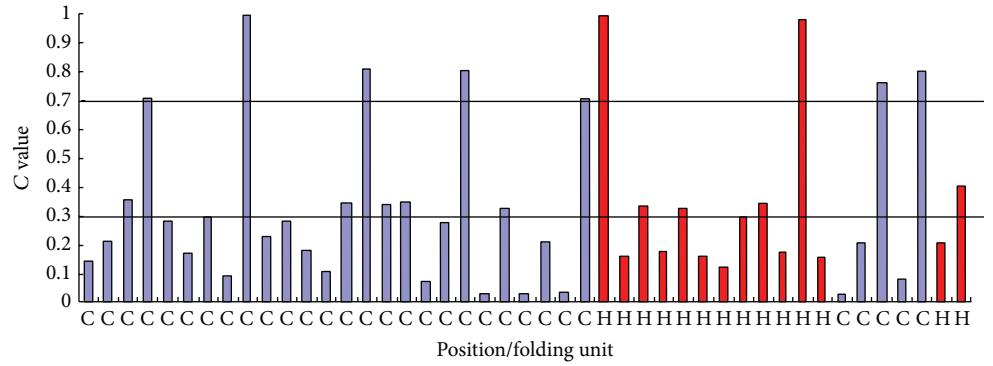
3.1. The Overall tRNA Abundance for Each Codon Position of the nsp1α Gene. Based on the C values, the tRNA abundance for each codon position along the PRRSV nsp1α gene was mapped. The translation speed for the synthesis of the nsp1α is not stable in the pigs (Figure 1). The codon positions with the C values much less than 0.30 have a tendency to cluster in *nsp1α* gene, including the positions 4–6, 8–10, 22–25, 27–30, 32–34, 38–40, 42–47, 50–53, 55–58, 68–70, 77–79, 83–85, 110–112, 119–122, 126–128, 139–141, 157–160, and 171–173. However, the codon positions with the C values much greater than 0.70 have few chances to cluster in *nsp1α* gene which is translated in the pigs. Due to most codon positions with C values much less than 0.70 existing in the target gene, these positions within the *nsp1α* might reduce the translation rate of this protein when the *nsp1α* was scanned by the ribosomes in pig cells. It is noted that there are no significant differences ($P > 0.05$) of the overall tRNA abundance for the codon positions in the regions of the three specific folding units of the nsp1α. This result suggests that the fluctuation of the overall tRNA abundance pairing to each codon position along this *nsp1α* might not regulate the formation of the specific folding units but decrease the scanning speed of ribosomes in the pig cells.

3.2. The Relationship between the Synonymous Codon Usage Bias and the Structure of the nsp1α. Based on the P values for the synonymous codons which are involved in the formation of the specific folding units in the nsp1α, we found the link between SCUD and the specific folding unit ($P = 2.75 \times 10^{-11}$). In detail, the synonymous codons have a strong propensity toward shaping the α-helix unit, including AUC for Ile, GUA for Val, AGC for Ser, AAG for Lys, and AUG for Met (Table 1). Turning to the effects of SCUB on shaping the β-strand unit, there are UUA for Leu, AUA for Ile, GUG for Val, UCA and AGU for Ser, ACA for Thr, UAC for Tyr, CGC for Arg, and two synonymous codons for His (Table 1). It is interesting that there are no codons which have a strong tendency to exist in the coil of the nsp1α (Table 1). As for the codons which have a strong tendency (P value > 1.0) to exist in the nsp1α, all of them strongly tend to exist only in the α-helix or the β-strand of this protein.

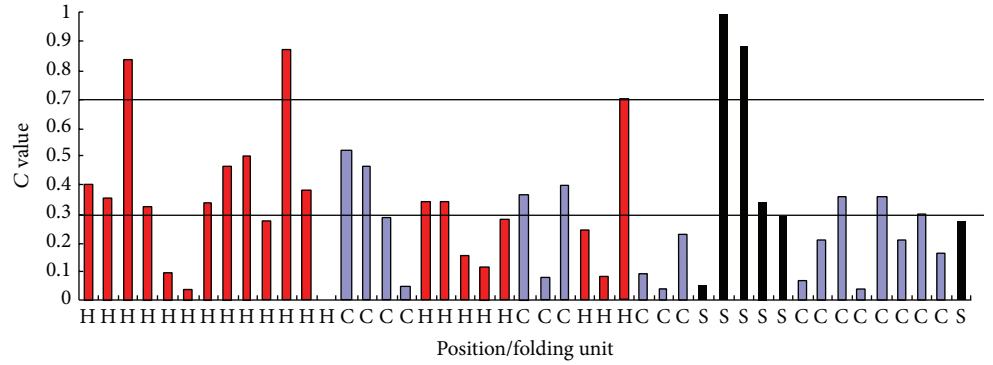
3.3. The Relative Abundance of the Codon with N_1 Context in the nsp1α Gene. As for the codons which have a strong tendency to exist in the specific folding unit of the nsp1α, their R values, the relative abundance of codons with N_1 contexts, were calculated from the 191 nsp1α genes (Table 2). The data show that the occurrence of the codon with N_1 context or ${}_1N$ context is not random, and many codons with N_1 context or ${}_1N$ context have a strong tendency to



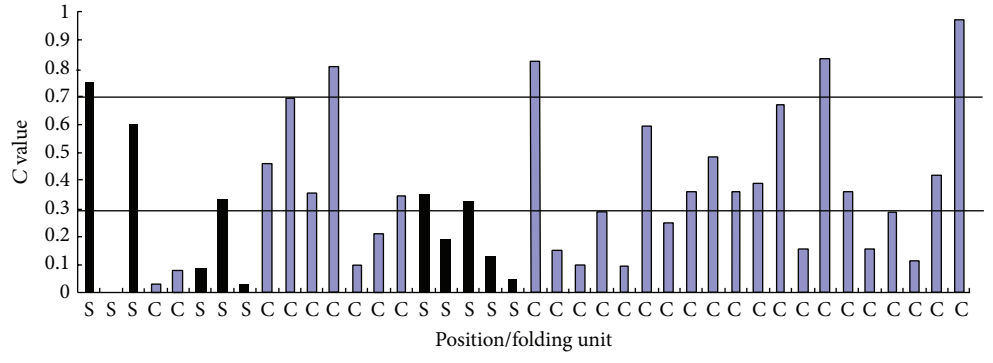
(a)



(b)



(c)



(d)

Figure 1: The overall tRNA abundance for each codon position of the PRRSV *nsp1α* gene. The black bar corresponds to the *β*-strand region in the *nsp1α*; the blue bar corresponds to the coil region in the *nsp1α*; the red bar corresponds to helix region in the *nsp1α*. (a) The codon positions range from the first position to the 45th position of the *nsp1α*. (b) The codon positions range from the 46th position to the 90th position of the *nsp1α*. (c) The codon positions range from the 91th position of the *nsp1α*. (d) The codon positions range from the 136th to the 175th position of the *nsp1α*.

Table 1: The relationship between the synonymous codon usage and the formation of the specific secondary structure unit.

Amino acid	Codon	[a]P value	[b]P value	[c]P value
Phe	UUU	0.40	0.07	−0.21
Phe	UUC	−0.92	−0.40	0.26
Leu	UUA	*−1.07*	**1.47**	*−1.22*
Leu	UUG	−0.16	−0.53	0.15
Leu	CUU	0.68	[d]—	−0.06
Leu	CUC	*−5.53*	*−5.32*	0.48
Leu	CUA	0.17	*−1.37*	0.13
Leu	CUG	0.00	*−4.15*	0.24
Ile	AUU	0.99	0.92	*−5.46*
Ile	AUC	**1.50**	*−1.17*	[d]—
Ile	AUA	0.69	**1.12**	*−2.46*
Val	GUU	0.58	0.28	−0.45
Val	GUC	0.32	0.20	−0.21
Val	GUA	**1.27**	0.02	*−2.17*
Val	GUG	−0.52	**1.21**	−0.73
Ser	UCU	0.76	[d]—	−0.12
Ser	UCC	*−4.45*	*−3.14*	0.47
Ser	UCA	[d]—	**1.76**	[d]—
Ser	UCG	[d]—	[d]—	[d]—
Ser	AGU	*−4.79*	**1.35**	−0.61
Ser	AGC	**1.24**	−0.25	*−1.53*
Pro	CCU	−0.48	*−2.90*	0.33
Pro	CCC	*−2.85*	−0.28	0.33
Pro	CCA	−0.32	*−4.40*	0.31
Pro	CCG	−0.23	*−2.69*	0.28
Thr	ACU	−0.08	0.58	−0.21
Thr	ACC	−0.57	0.53	−0.05
Thr	ACA	0.58	**1.28**	*−5.24*
Thr	ACG	*−4.40*	[d]—	0.48
Ala	GCU	0.55	*−2.56*	0.00
Ala	GCC	−0.07	0.10	−0.01
Ala	GCA	0.78	0.16	−0.61
Ala	GCG	0.76	1.15	*−4.18*
Tyr	UAU	0.95	*−1.60*	−0.39
Tyr	UAC	0.21	**1.37**	*−2.22*
His	CAU	*−4.21*	**1.75**	*−4.18*
His	CAC	[d]—	**1.75**	*−3.81*
Gln	CAA	0.62	−0.82	−0.15
Gln	CAG	*−2.74*	0.63	0.07
Asn	AAU	*−4.47*	−0.58	0.38
Asn	AAC	[d]—	0.18	0.25
Lys	AAA	−0.90	[d]—	0.39
Lys	AAG	**1.18**	[d]—	−0.69
Asp	GAU	[d]—	*−3.17*	0.48
Asp	GAC	[d]—	[d]—	0.48
Glu	GAA	0.93	[d]—	−0.28
Glu	GAG	*−1.15*	*−5.28*	0.41
Cys	UGU	[d]—	*−4.75*	0.48
Cys	UGC	*−5.41*	0.04	0.28
Arg	CGU	*−1.82*	0.28	0.18
Arg	CGC	[d]—	**1.52**	*−1.03*
Arg	CGA	0.83	*−2.13*	−0.22
Arg	CGG	−0.04	0.16	−0.04
Arg	AGA	**1.37**	*−4.30*	*−1.32*
Arg	AGG	[d]—	0.76	0.03
Gly	GGU	0.52	[d]—	0.04
Gly	GGC	−0.34	*−4.61*	0.32
Gly	GGA	*−4.76*	*−4.55*	0.48
Gly	GGG	[d]—	*−4.62*	0.48
Met	AUG	**1.01**	−0.35	−0.72
Trp	UGG	0.45	0.66	−0.61

[a]*represents* α-helix.
[b]*represents* β-strand.
[c]*represents* The coil.
[d]*represents* The corresponding codon is not selected in the specific secondary structure unit.
Italic indicates that the corresponding codon has a weak bias to be selected in a specific secondary structure unit.
Bold indicates that the corresponding codon has a tendency to be selected in a specific secondary structure unit.

exist in the specific folding units of the nsp1α. Based on the data of SCUB in the specific folding units (Table 1), the corresponding codon with N_1 context was found to have a trend to exist in the specific folding unit of the nsp1α. In detail, the codons with N_1 context or ${}_1N$ context (GUA~A, AGC~A, AAG~C, AGA~C, A~AUA, U~AGC, U~AAG, C~AAG, G~AGA, and U~AUG) have an obvious trend to exist in the helix unit of the nsp1α. Some codons with N_1 context or ${}_1N$ context have a strong tendency to exist in the β-strand of the nsp1α, including UUA~A, AUA~G, GUG~U, UCA~C, AGU~G, ACA~C, UAC~U, UAC~C, CAU~G, CAC~G, CGC~U, G~UUA, U~AUA, U~GUG, G~GUG, U~UCA, C~AGU, C~ACA, C~UAC, G~UAC, U~CAU, U~CAC, and U~CGC.

In order to identify the roles of nucleotide compositions (dinucleotide with N_1 context and mononucleotide with N_1 context) in shaping the codon with N_1 context or ${}_1N$ context, the R values for these interesting codons with N_1 context or ${}_1N$ context which have a strong tendency to exist in the helix or the β-strand were compared with the r values for the dinucleotide/mononucleotide with N_1 context (Tables 3 and 4). The R value for the target codon with N_1 context is higher than the corresponding dinucleotide/mononucleotide with N_1 context or ${}_1N$ context. For example, as for GUA which tends to exist in the helix of the nsp1α gene, GUA~A has a tendency to exist in the helix unit, because the R value (1.4751) of GUA~A for the helix unit is higher than the R value for GUA~A for the β-strand and the coil (Table 2) and higher than the r value for UA~A and the r value for A~A (Tables 3 and 4). As for UUA which tends to exist in the β-strand of this gene, UUA~A has a tendency to exist in the strand unit, because the R value (4.9268) for UUA~A is higher than the R values for UUA~A of the helix and the coil and higher than the r values for UA~A and A~A (Tables 3 and 4). As for AGC which tends to exist in the helix of this gene, U~AGC has

Table 2: Relative abundance of codons with N_1 context or $_1N$ context in the PRRSV *nsp1α* gene.

Codon-context (*xyz-n*)	*R* value[1]	*R* value[2]	*R* value[3]	Codon-context (*xyz-n*)	*R* value[1]	*R* value[2]	*R* value[3]
UUA~A	0.0000	4.9268	0.3345	A~UUA	0.0000	0.0401	3.6799
UUA~U	0.2911	0.0000	0.0000	U~UUA	0.1456	0.0617	0.2945
UUA~C	2.8120	0.0000	0.0000	C~UUA	2.9124	0.0000	0.0000
UUA~G	0.0000	0.0000	3.7713	G~UUA	0.0000	3.7019	0.0000
AUC~A	0.0000	0.0000	0.0000	A~AUC	0.0000	0.0000	2.4430
AUC~U	0.0000	0.0000	3.1392	U~AUC	0.0000	0.2233	0.0116
AUC~C	0.0000	0.0000	0.0000	C~AUC	0.0000	3.4849	0.0439
AUC~G	0.0000	3.7945	0.4601	G~AUC	0.0000	0.0000	1.5563
AUA~A	0.0000	0.0000	0.0000	A~AUA	0.0000	0.0000	0.0000
AUA~U	0.0000	0.0000	0.0000	U~AUA	4.3672	3.4158	2.2091
AUA~C	3.0128	0.0000	0.5557	C~AUA	0.0000	0.3703	0.0000
AUA~G	0.0000	3.7945	3.5998	G~AUA	0.0000	0.0000	1.5428
GUA~A	1.4751	0.0000	0.0934	A~GUA	2.9502	3.2845	0.0000
GUA~U	1.0918	3.7954	0.0822	U~GUA	0.0000	0.0000	0.0000
GUA~C	0.7532	0.0000	4.2386	C~GUA	0.7532	1.2342	0.0000
GUA~G	0.9274	0.0000	0.0000	G~GUA	0.9274	0.0000	4.1141
GUG~A	0.0510	0.0000	0.0000	A~GUG	1.4283	0.0000	0.3526
GUG~U	0.2014	2.7198	0.0000	U~GUG	0.0629	1.8132	0.0239
GUG~C	1.4413	0.1874	4.4454	C~GUG	0.0174	0.4872	0.0000
GUG~G	1.7318	0.8833	0.0000	G~GUG	2.7367	1.4824	3.7249
UCA~A	0.0000	0.0000	0.0000	A~UCA	0.0000	0.0000	0.0000
UCA~U	0.0000	0.0000	0.0000	U~UCA	0.0000	3.7954	0.0000
UCA~C	0.0000	3.7027	0.0000	C~UCA	0.0000	0.0000	0.0000
UCA~G	0.0000	0.0000	0.0000	G~UCA	0.0000	0.0000	0.0000
AGU~A	0.0000	0.0000	0.0000	A~AGU	0.0000	0.0000	0.0000
AGU~U	0.0000	0.0000	0.0000	U~AGU	2.4694	0.0000	0.0000
AGU~C	0.0473	0.0000	0.0000	C~AGU	1.2619	3.7027	4.4454
AGU~G	3.6513	3.7945	4.1141	G~AGU	0.0583	0.0000	0.0000
AGC~A	2.9502	0.0000	0.0000	A~AGC	0.0000	0.0000	0.0000
AGC~U	0.0000	0.0000	0.0000	U~AGC	5.9004	0.0000	0.0000
AGC~C	0.0000	0.0000	0.0000	C~AGC	0.0000	3.7027	0.0000
AGC~G	1.8548	3.7945	0.0000	G~AGC	0.0000	0.0000	0.0000
ACA~A	0.0000	0.0000	0.0000	A~ACA	0.0000	0.0000	0.0000
ACA~U	0.0000	0.0000	0.0000	U~ACA	4.3672	0.0000	3.4431
ACA~C	0.0000	3.7027	0.0383	C~ACA	0.0000	3.7027	0.1150
ACA~G	3.7096	0.0000	4.0786	G~ACA	0.0000	0.0000	0.0000
UAC~A	5.6595	0.0000	3.9513	A~UAC	0.0000	0.0502	0.0000
UAC~U	0.0000	1.4455	0.0000	U~UAC	1.2478	0.8812	1.4434
UAC~C	0.1230	1.4253	0.0000	C~UAC	2.1520	1.3876	2.6300
UAC~G	0.0000	0.8887	0.0646	G~UAC	0.0000	1.4529	0.0000
CAU~A	0.0000	0.0000	0.0000	A~CAU	3.9336	0.0000	0.0000
CAU~U	0.0000	0.0000	0.0000	U~CAU	0.0000	3.7043	0.0000
CAU~C	1.0043	0.0000	0.0000	C~CAU	0.0000	0.0889	4.4454
CAU~G	2.4730	3.7945	4.1141	G~CAU	1.2365	0.0000	0.0000
CAC~A	0.0000	0.0000	0.0000	A~CAC	0.0000	0.0000	0.0000
CAC~U	0.0000	0.2138	0.0000	U~CAC	4.3672	3.5816	0.0000
CAC~C	0.0000	0.0000	0.0000	C~CAC	0.0000	0.2086	0.0000
CAC~G	3.7096	3.5807	0.0000	G~CAC	0.0000	0.0000	0.0000
AAG~A	0.0000	0.0000	3.4986	A~AAG	0.0000	0.0000	3.3416

Table 2: Continued.

Codon-context (*xyz-n*)	*R* value[1]	*R* value[2]	*R* value[3]	Codon-context (*xyz-n*)	*R* value[1]	*R* value[2]	*R* value[3]
AAG~U	0.0546	0.0000	0.0000	U~AAG	1.4739	0.0000	0.0000
AAG~C	2.9751	0.0000	0.0000	C~AAG	1.9960	0.0000	0.0000
AAG~G	0.0000	0.0000	0.5286	G~AAG	0.0000	0.0000	0.6895
CGC~A	1.3112	0.0000	0.0000	A~CGC	2.8704	0.0000	0.0000
CGC~U	0.1213	3.7954	0.0000	U~CGC	0.1180	2.3721	0.0000
CGC~C	0.0837	0.0000	0.0000	C~CGC	1.3843	1.3885	0.0000
CGC~G	2.6791	0.0000	0.0000	G~CGC	0.1003	0.0000	0.0000
AGA~A	0.0000	0.0000	1.9453	A~AGA	0.0831	0.0000	2.0578
AGA~U	0.1230	0.0000	0.1188	U~AGA	0.6766	0.0000	0.0099
AGA~C	2.8855	0.0000	0.0249	C~AGA	0.1273	0.0000	2.1293
AGA~G	0.0522	0.0000	1.9591	G~AGA	2.9259	0.0000	0.0230
AUG~A	0.0000	0.0000	2.1721	A~AUG	0.0000	0.0000	3.9812
AUG~U	0.0000	0.0000	0.0000	U~AUG	4.1823	3.6955	0.0000
AUG~C	0.0000	0.0000	0.0000	C~AUG	0.1275	0.0974	0.0000
AUG~G	0.0000	3.7945	1.8881	G~AUG	0.0000	0.0000	0.0340

[1] *represents* The relative abundance of codons with N_1 context in the helix unit of the PRRSV *nsp1α*.
[2] *represents* The relative abundance of codons with N_1 context in the β-strand unit of the PRRSV *nsp1α*.
[3] *represents* The relative abundance of codons with N_1 context in the coil unit of the PRRSV *nsp1α*.

a tendency to exist in the helix rather than in the strand or the coil, because the *R* value (5.9004) for U~AGC is higher than the *R* value of U~AGC of the strand and the coil and higher than the *r* values for U~AG and U~A (Tables 3 and 4). As for UUA which tends to exist in the strand of this gene, G~UUA has a tendency to exist in the strand unit, because the *R* value (3.7019) for G~UUA of the strand unit is higher than the *R* values for G~UUA of the helix and the coil and higher than *r* values for G~UU and G~U (Tables 3 and 4). Based on the standard mentioned above, GUA~A, AGC~A, AAG~C, AGA~C, A~GUA, U~AGC, U~AAG, C~AAG, G~AGA, and U~AUG have a strong trend to exist in the helix of PRRSV nsp1*α* gene and UUA~A, AUA~G, GUG~U, UCA~C, ACA~C, UAC~U, UAC~C, CAU~G, CGC~U, G~UUA, U~GUG, U~UCA, C~AGU, C~ACA, G~UAC, U~CAU, and U~CGC have a strong tendency to exist in the β-strand of the nsp1*α* gene.

4. Discussion

In this study, we have mapped the fluctuation of the overall tRNA abundance for each codon position along the PRRSV nsp1*α* gene and estimated the correlation between the synonymous codon usage and different folding units of the nsp1*α*. The performance of mapping the fluctuation of the overall tRNA abundance for each codon position along the target gene likely reflects the translation speed of ribosomes scanning caused by the tRNA abundance of the pigs to some degree, since the tRNA abundance plays an important role in the ribosome scanning along the target coding sequence [35, 36]. The previous report showed that the α-helix is preferentially coded by translationally fast mRNA regions while the slow segments often encode β-strands and coil regions [37]. In the study, no linkage between the fluctuation of the overall tRNA abundance pairing to the codon positions along the nsp1*α* gene and the specific folding units might suggest that the process of translation fine-tunes is not performed by variation of translation speed for each codon position along the *nsp1α*. The fine-tuning *in vivo* protein folding exists in the gene, and this regularity is largely believed to occur in a cotranslational process [38]. However, the PRRSV nsp1*α* derives from the posttranslational processing of the pp1a [10, 39]. The process of the cleavage of the nsp1*α* from the pp1a polyprotein of PRRSV performed by the posttranslation might be free from the fluctuation of tRNA abundance pairing to the each codon position along the nsp1*α* gene. As for the ribosomes scanning the nsp1*α* gene, there is no significant link between the fluctuation of the overall tRNA abundance and the specific folding units, and the translation elongation rate of this gene is not high. These results suggest that the low tRNA abundance controls the ribosomal traffic along the translated message to achieve the effective synthesized product of the PRRSV pp1a. The low translational elongation at the translation beginning step directs the target gene to generate the corresponding protein effectively [40].

Turning to the role of the synonymous codon usage in the formation of the specific units of the nsp1*α*, there is significant relationship between the synonymous codon usage bias and the specific folding units in the target protein. The synonymous codons assist messenger RNA to carry the information of the specific folding units, and a single codon or a contiguous nucleotide region plays roles in shaping the specific folding units [24, 25, 41, 42]. As for the PRRSV *nsp1α*, there is no synonymous codon which tends to exist in coil unit. However, many synonymous codons exist in the α-helix and β-strands regions of this gene, and no synonymous codon has a strong tendency to be selected by both the α-helix and the β-strands in the PRRSV *nsp1α* simultaneously. These results indicated that SCUB might play roles in shaping this

Table 3: Relative abundance of dinucleotides with N_1 context or $_1N$ context in the PRRSV *nsp1α* gene.

Dinucleotides with N_1 context ($xy{\sim}n$)	r value	Dinucleotides with $_1N$ context ($n{\sim}xy$)	r value
UC~A	0.6130	A~AU	1.4456
UC~U	1.3115	U~AU	0.8552
UC~C	1.2805	C~AU	0.6485
UC~G	0.5669	G~AU	1.0047
UA~A	0.6177	A~GU	1.2963
UA~U	1.0269	U~GU	1.1651
UA~C	1.7402	C~GU	0.5072
UA~G	0.4351	G~GU	1.1064
UG~A	1.2060	A~UC	0.5880
UG~U	0.8195	U~UC	1.0330
UG~C	1.0921	C~UC	1.1230
UG~G	0.8019	G~UC	1.0286
CA~A	1.8155	A~AG	1.2415
CA~U	0.4556	U~AG	0.2825
CA~C	0.9184	C~AG	0.9946
CA~G	0.8963	G~AG	1.5135
GU~A	0.3165	A~AC	1.5521
GU~U	0.9679	U~AC	0.9913
GU~C	0.8485	C~AC	0.8942
GU~G	1.5896	G~AC	0.6067
GC~A	1.2887	A~UA	0.6720
GC~U	1.0849	U~UA	0.9776
GC~C	1.3411	C~UA	1.6741
GC~G	0.3298	G~UA	0.5155
CA~A	1.8155	A~CA	0.9721
CA~U	0.4556	U~CA	0.4819
CA~C	0.9184	C~CA	1.3762
CA~G	0.8963	G~CA	1.0994
AC~A	0.9465	A~AA	1.1734
AC~U	1.2243	U~AA	0.3547
AC~C	0.8848	C~AA	1.7821
AC~G	0.9524	G~AA	1.4708
AU~A	0.5596	A~CG	1.6312
AU~U	0.9124	U~CG	0.7431
AU~C	0.6579	C~CG	1.2830
AU~G	1.7819	G~CG	0.4693
AG~A	1.0440	A~UU	0.8843
AG~U	1.3708	U~UU	1.5257
AG~C	0.7761	C~UU	1.2791
AG~G	0.8164	G~UU	1.2724
GC~A	1.2887		
GC~U	1.0849		
GC~C	1.3411		
GC~G	0.3298		

Table 4: Relative abundance of mononucleotide with N_1 context in PRRSV *nsp1α* gene.

Mononucleotide with N_1 context	$r(x{\sim}n)$
A~A	1.0664
A~U	0.7354
A~C	1.0751
A~G	0.9433
U~A	0.6124
U~U	0.7588
U~C	0.8228
U~G	1.4182
C~A	1.0468
C~U	1.1225
C~C	0.8788
C~G	0.6277
G~A	1.0557
G~U	0.9975
G~C	0.8930
G~G	0.7782

protease with natural properties for the life-cycle of PRRSV. SCUB for formation of the specific folding units of the PRRSV nsp1*α* is influenced by the natural selection. As an example of the role for natural selection, the expressivity of genes is an important factor in shaping SCUB, both for prokaryotic and for eukaryotic organisms [18, 22, 43, 44]. Although the link between the SCUB and the formation of the specific folding units was reported [25, 35, 37, 38, 45], the role of CDCB in formation of specific folding units is not clear. In this study, we found that CDCB plays a role in the formation of specific folding units in the PRRSV nsp1*α*. The synonymous usage bias and CDCB, which play important roles in achieving accuracy and efficiency in protein synthesis, are particular manifestations of coding sequence nonrandomness [23, 46, 47]. Spatial interaction of ribosomal proteins with codon-anticodon RNA pairs inside the A and P sites of the ribosome could be preferable for particular codons with context [20, 48].

Conflict of Interests

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

Authors' Contribution

Yao-zhong Ding, Ya-nan You, and Dong-jie Sun contributed to the original draft of the paper and approved the final version. Hao-tai Chen, Yong-lu Wang, Hui-yun Chang, Li Pan, Yu-zhen Fang, Zhong-wang Zhang, Peng Zhou, Jian-liang Lv, Xinsheng Liu, Jun-jun Shao, Fu-rong Zhao, and Tong Lin downloaded the sequences and analyzed the data. Laszlo Stipkovits, Zygmunt Pejsak, Yong-guang Zhang, and Jie Zhang provided suggestive information to the Discussion

and revised the paper. All authors read and approved the final version.

Acknowledgments

This work was supported in part by Grants from International Science and Technology Cooperation Program of China (no. 2012DFG31890) and Gansu Provincial Funds for Distinguished Young Scientists (1111RJDA005). This study was also supported by National Natural Science foundation of China (no. 31172335 and no. 31072143).

References

[1] J. G. Cho and S. A. Dee, "Porcine reproductive and respiratory syndrome virus," *Theriogenology*, vol. 66, no. 3, pp. 655–662, 2006.

[2] K.-K. Conzelmann, N. Visser, P. van Woensel, and H.-J. Thiel, "Molecular characterization of porcine reproductive and respiratory syndrome virus, a member of the arterivirus group," *Virology*, vol. 193, no. 1, pp. 329–339, 1993.

[3] J. J. M. Meulenberg, M. M. Hulst, E. J. De Meijer et al., "Lelystad virus, the causative agent of porcine epidemic abortion and respiratory syndrome (PEARS), is related to LDV and EAV," *Virology*, vol. 192, no. 1, pp. 62–72, 1993.

[4] R. Allende, T. L. Lewis, Z. Lu et al., "North American and European porcine reproductive and respiratory syndrome viruses differ in non-structural protein coding regions," *Journal of General Virology*, vol. 80, part 2, pp. 307–315, 1999.

[5] E. M. Bautista, S. M. Goyal, I. J. Yoon, H. S. Joo, and J. E. Collins, "Comparison of porcine alveolar macrophages and CL 2621 for the detection of porcine reproductive and respiratory syndrome (PRRS) virus and anti-PRRS antibody," *Journal of Veterinary Diagnostic Investigation*, vol. 5, no. 2, pp. 163–165, 1993.

[6] J. E. Collins, D. A. Benfield, W. T. Christianson et al., "Isolation of swine infertility and respiratory syndrome virus (isolate ATCC VR-2332) in North America and experimental reproduction of the disease in gnotobiotic pigs," *Journal of Veterinary Diagnostic Investigation*, vol. 4, no. 2, pp. 117–126, 1992.

[7] Y. S. Liu, J. H. Zhou, H. T. Chen et al., "Analysis of synonymous codon usage in porcine reproductive and respiratory syndrome virus," *Infection, Genetics and Evolution*, vol. 10, no. 6, pp. 797–803, 2010.

[8] C. J. Nelsen, M. P. Murtaugh, and K. S. Faaberg, "Porcine reproductive and respiratory syndrome virus comparison: divergent evolution on two continents," *Journal of Virology*, vol. 73, no. 1, pp. 270–280, 1999.

[9] E. J. Snijder, A. L. M. Wassenaar, L. C. Van Dinten, W. J. M. Spaan, and A. E. Gorbalenya, "The arterrvirus Nsp4 protease is the prototype of a novel group of chymotrypsin-like enzymes, the 3C-like serine proteases," *Journal of Biological Chemistry*, vol. 271, no. 9, pp. 4864–4871, 1996.

[10] Y. Fang and E. J. Snijder, "The PRRSV replicase: exploring the multifunctionality of an intriguing set of nonstructural proteins," *Virus Research*, vol. 154, no. 1-2, pp. 61–76, 2010.

[11] D. D. Nedialkova, A. E. Gorbalenya, and E. J. Snijder, "Arterivirus Nsp1 modulates the accumulation of minus-strand templates to control the relative abundance of viral mRNAs," *PLOS Pathogens*, vol. 6, no. 2, Article ID e1000772, 2010.

[12] J. A. den Boon, K. S. Faaberg, J. J. M. Meulenberg et al., "Processing and evolution of the N-terminal region of the arterivirus replicase ORF1a protein: identification of two papainlike cysteine proteases," *Journal of Virology*, vol. 69, no. 7, pp. 4500–4505, 1995.

[13] E. J. Snijder, A. L. M. Wassenaar, and W. J. M. Spaan, "The 5′ end of the equine arteritis virus replicase gene encodes a papainlike cysteine protease," *Journal of Virology*, vol. 66, no. 12, pp. 7040–7048, 1992.

[14] T. Dokland, "The structural biology of PRRSV," *Virus Research*, vol. 154, no. 1-2, pp. 86–97, 2010.

[15] M. V. Kroese, J. C. Zevenhoven-Dobbe, J. N. A. Bos-de Ruijter et al., "The nsp 1α and nsp 1β papain-like autoproteinases are essential for porcine reproductive and respiratory syndrome virus RNA synthesis," *Journal of General Virology*, vol. 89, no. 2, pp. 494–499, 2008.

[16] S. Yuna, X. Fei, G. Yu et al., "Crystal structure of porcine reproductive and respiratory syndrome virus leader protease Nsp1α," *Journal of Virology*, vol. 83, no. 21, pp. 10931–10940, 2009.

[17] X. Shi, J. Chen, G. Xing et al., "Amino acid at position 176 was essential for porcine reproductive and respiratory syndrome virus (PRRSV) non-structural protein 1α (nsp1α) as an inhibitor to the induction of IFN-β," *Cellular Immunology*, vol. 280, no. 2, pp. 125–131, 2012.

[18] I. Bahir, M. Fromer, Y. Prat, and M. Linial, "Viral adaptation to host: a proteome-based analysis of codon usage and amino acid preferences," *Molecular Systems Biology*, vol. 5, article 311, 2009.

[19] T. F. Clarke IV and P. L. Clark, "Rare codons cluster," *PLoS ONE*, vol. 3, no. 10, Article ID e3412, 2008.

[20] A. Fedorov, S. Saxonov, and W. Gilbert, "Regularities of context-dependent codon bias in eukaryotic genes," *Nucleic Acids Research*, vol. 30, no. 5, pp. 1192–1197, 2002.

[21] J. Zhou, Z. Gao, J. Zhang et al., "The analysis of codon bias of foot-and-mouth disease virus and the adaptation of this virus to the hosts," *Infection, Genetics and Evolution*, vol. 14, no. 1, pp. 105–110, 2013.

[22] J. Zhou, J. Su, H. Chen et al., "Clustering of low usage codons in the translation initiation region of hepatitis C virus," *Infection, Genetics and Evolution*, vol. 18, pp. 8–12, 2013.

[23] H. Akashi, "Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy," *Genetics*, vol. 136, no. 3, pp. 927–935, 1994.

[24] I. Weygand-Durasevic and M. Ibba, "Cell biology: new roles for codon usage," *Science*, vol. 329, no. 5998, pp. 1473–1474, 2010.

[25] J. h. Zhou, Y. N. You, H. T. Chen et al., "The effects of the synonymous codon usage and tRNA abundance on protein folding of the 3C protease of foot-and-mouth disease virus," *Infection, Genetics and Evolution*, vol. 16, pp. 270–274, 2013.

[26] T. Zhou, M. Weems, and C. O. Wilke, "Translationally optimal codons associate with structurally sensitive sites in proteins," *Molecular Biology and Evolution*, vol. 26, no. 7, pp. 1571–1580, 2009.

[27] Y. Guisez, J. Robbens, E. Remaut, and W. Fiers, "Folding of the MS2 coat protein in *Escherichia coli* is modulated by translational pauses resulting from mRNA secondary structure and codon usage: a hypothesis," *Journal of Theoretical Biology*, vol. 162, no. 2, pp. 243–252, 1993.

[28] X. Tao and D. Dafu, "The relationship between synonymous codon usage and protein structure," *FEBS Letters*, vol. 434, no. 1-2, pp. 93–96, 1998.

[29] M. Gouy, "Codon contexts in enterobacterial and coliphage genes," *Molecular Biology and Evolution*, vol. 4, no. 4, pp. 426–444, 1987.

[30] S. Karlin and J. Mrázek, "What drives codon choices in human genes?" *Journal of Molecular Biology*, vol. 262, no. 4, pp. 459–472, 1996.

[31] E. G. Shpaer, "Constraints on codon context in *Escherichia coli* genes. Their possible role in modulating the efficiency of translation," *Journal of Molecular Biology*, vol. 188, no. 4, pp. 555–564, 1986.

[32] O. G. Berg and P. J. N. Silva, "Codon bias in Escherichia coli: The influence of codon context on mutation and selection," *Nucleic Acids Research*, vol. 25, no. 7, pp. 1397–1404, 1997.

[33] J. D. Thompson, D. G. Higgins, and T. J. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic Acids Research*, vol. 22, no. 22, pp. 4673–4680, 1994.

[34] S. Karlin and C. Burge, "Dinucleotide relative abundance extremes: a genomic signature," *Trends in Genetics*, vol. 11, no. 7, pp. 283–290, 1995.

[35] J. L. Parmley and M. A. Huynen, "Clustering of codons with rare cognate tRNAs in human genes suggests an extra level of expression regulation," *PLoS Genetics*, vol. 5, no. 7, Article ID e1000548, 2009.

[36] E. P. C. Rocha, "Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization," *Genome Research*, vol. 14, no. 11, pp. 2279–2286, 2004.

[37] T. A. Thanaraj and P. Argos, "Protein secondary structural types are differentially coded on messenger RNA," *Protein Science*, vol. 5, no. 10, pp. 1973–1983, 1996.

[38] A. A. Komar, "A pause for thought along the co-translational folding pathway," *Trends in Biochemical Sciences*, vol. 34, no. 1, pp. 16–24, 2009.

[39] D. van Aken, J. Zevenhoven-Dobbe, A. E. Gorbalenya, and E. J. Snijder, "Proteolytic maturation of replicase polyprotein pp1a by the nsp4 main proteinase is essential for equine arteritis virus replication and includes internal cleavage of nsp7," *Journal of General Virology*, vol. 87, no. 12, pp. 3473–3482, 2006.

[40] T. Tuller, A. Carmi, K. Vestsigian et al., "An evolutionarily conserved mechanism for controlling the efficiency of protein translation," *Cell*, vol. 141, no. 2, pp. 344–354, 2010.

[41] D. B. Carlini, Y. Chen, and W. Stephan, "The relationship between third-codon position nucleotide content, codon bias, mRNA secondary structure and gene expression in the drosophilid alcohol dehydrogenase genes Adh and Adhr," *Genetics*, vol. 159, no. 2, pp. 623–633, 2001.

[42] R. Saunders and C. M. Deane, "Synonymous codon usage influences the local protein structure observed," *Nucleic Acids Research*, vol. 38, no. 19, pp. 6719–6728, 2010.

[43] M. D. Ermolaeva, "Synonymous codon usage in bacteria," *Current Issues in Molecular Biology*, vol. 3, no. 4, pp. 91–97, 2001.

[44] J. B. Plotkin, H. Robins, and A. J. Levine, "Tissue-specific codon usage and the expression of human genes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 34, pp. 12588–12591, 2004.

[45] G. Zhang and Z. Ignatova, "Generic algorithm to predict the speed of translational elongation: implications for protein biogenesis," *PLoS ONE*, vol. 4, no. 4, Article ID e5036, 2009.

[46] A. Eyre-Walker, "Synonymous codon bias is related to gene length in Escherichia coli: selection for translational accuracy?" *Molecular Biology and Evolution*, vol. 13, no. 6, pp. 864–872, 1996.

[47] J. Precup and J. Parker, "Missense misreading of asparagine codons as a function of codon identity and context," *The Journal of Biological Chemistry*, vol. 262, no. 23, pp. 11351–11355, 1987.

[48] R. Green and H. F. Noller, "Ribosomes and translation," *Annual Review of Biochemistry*, vol. 66, pp. 679–716, 1997.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 471323, 9 pages
http://dx.doi.org/10.1155/2014/471323

Research Article

Cell Type-Dependent RNA Recombination Frequency in the Japanese Encephalitis Virus

Wei-Wei Chiang,[1] Ching-Kai Chuang,[1,2] Mei Chao,[1,3] and Wei-June Chen[1,4]

[1] *Division of Microbiology, Graduate Institute of Biomedical Sciences, Chang Gung University, Kwei-San, Tao-Yuan 33332, Taiwan*
[2] *Department of Plant Pathology, University of Kentucky, Lexington, KY 40546-0312, USA*
[3] *Department of Microbiology and Immunology, Chang Gung University, Kwei-San, Tao-Yuan 33332, Taiwan*
[4] *Department of Public Health and Parasitology, Chang Gung University, Kwei-San, Tao-Yuan 33332, Taiwan*

Correspondence should be addressed to Wei-June Chen; wjchen@mail.cgu.edu.tw

Received 10 May 2014; Accepted 2 July 2014; Published 22 July 2014

Academic Editor: Hongwei Wang

Japanese encephalitis virus (JEV) is one of approximately 70 flaviviruses, frequently causing symptoms involving the central nervous system. Mutations of its genomic RNA frequently occur during viral replication, which is believed to be a force contributing to viral evolution. Nevertheless, accumulating evidences show that some JEV strains may have actually arisen from RNA recombination between genetically different populations of the virus. We have demonstrated that RNA recombination in JEV occurs unequally in different cell types. In the present study, viral RNA fragments transfected into as well as viral RNAs synthesized in mosquito cells were shown not to be stable, especially in the early phase of infection possibly via cleavage by exoribonuclease. Such cleaved small RNA fragments may be further degraded through an RNA interference pathway triggered by viral double-stranded RNA during replication in mosquito cells, resulting in a lower frequency of RNA recombination in mosquito cells compared to that which occurs in mammalian cells. In fact, adjustment of viral RNA to an appropriately lower level in mosquito cells prevents overgrowth of the virus and is beneficial for cells to survive the infection. Our findings may also account for the slower evolution of arboviruses as reported previously.

1. Introduction

Japanese encephalitis (JE) is an important mosquito-borne viral disease, occasionally causing encephalitic symptoms [1]. Nowadays, it is extensively distributed in most Asian countries and was also recently reported from Australia [2]. The JE virus (JEV) is one of some 70 members of the genus *Flavivirus* belonging to the family Flaviviridae [3], the genome of which contains a linear, single-stranded positive-sense RNA (~11 kb long) that encodes 3 structural proteins including nucleocapsid (C), membrane (preM/M), and envelope (E) proteins, as well as 7 nonstructural proteins (NS1, NS2A, NS2B, NS3, NS4A, NS4B, and NS5) [4]. Due to lack of a proofreading mechanism and an inability to repair errors during RNA synthesis, spontaneous mutations frequently occur which contribute to the formation of genetically diversified populations or so-called "quasispecies" in flaviviruses including the JEV [5].

Maintenance of genetic diversity theoretically reduces the rapid loss of fitness via Muller's ratcheting during viral passage from one host to another [6], which provides benefits to a virus that is adapting to a new niche or selective regimen of its environment [7]. Possibly, this feature differentially occurs in different types of host cells [8]. In addition to gene mutations [9], RNA recombination, at least in some cases, can also serve as a factor helping a virus escape from accumulated deleterious effects in a viral population [10]. In other words, RNA recombination may serve as an alternate means to generate genetic changes [11] and likely produces a new form of RNA comprising genetic information from multiple sources [12].

The viral RNA recombination was first reported in the poliovirus, a picornavirus [13], and subsequently in a variety of viruses that infect humans, animals, plants, and bacteria [14–18]. Therefore, a new virus may be generated through RNA recombination between different strains.

Among arboviruses, at least the western equine encephalitis virus is believed to be a recombinant virus that arose from distant viral progenitors, including an eastern equine encephalitis virus-like virus and a Sindbis-like virus [19]. As a result, the ability to form unpredictable recombinant strains or species between virus populations is of considerable concern [20], particularly the possibility of RNA recombination occurring from cocirculated live-attenuated vaccine strains and wild viruses during synthesis of new RNAs [21, 22].

Flaviviruses naturally comprise multiple genotypes or strains [23, 24], making them likely to undergo RNA recombination. The first RNA recombination of the JEV was proposed based on a bioinformatics analysis [17]. Furthermore, RNA recombination was found to occur unequally in mosquito and mammalian cells [25]. Herein, we provided evidences of RNA recombination of the JEV that occurs at a lower frequency in mosquito cells, which may, at least partly, contribute to evolution of the virus [26].

2. Materials and Methods

2.1. Viruses and Cell Lines. Three strains of the JEV, including Nakayama (the vaccine strain), T1P1-S1 (a small plaque clone from the T1P1 strain) [27], and CJN-S1 (a small plaque clone from the CJN strain, a kind gift from Dr. M. H. Ho, Academia Sinica, Taipei, Taiwan), were used in this study. Of these, further purification via the plaque-picking method to select T1P1-S1 and CJN-S1 strains was implemented as part of the present study [27]. The viruses were propagated in C6/36 mosquito cells and titrated in baby hamster kidney- (BHK-) 21 cells. Both cell lines were maintained as previously described [27].

2.2. Virus Titration. Virus titers were determined by means of a plaque assay of BHK-21 cells following descriptions in our previous report [27]. Calculation of virus titers was based on the number of formed plaques, expressed as plaque-forming units (pfu)/mL.

2.3. Reverse Transcriptase-Polymerase Chain Reaction (RT-PCR). To detect viral infection in cells, extracted RNA was applied to perform RT with the reverse primer at 42°C for 30 min to generate complementary (c) DNA. PCR cycling was then carried out using the forward primer which was subsequently run to amplify a gene fragment with a size of 529 bp under the following conditions: 25 cycles of 95°C for 30 s, 60°C for 30 s, and 72°C for 1 min. The primers used to amplify specific regions are presented in individual sections below. All procedures in this portion of the study followed our previous description [25].

2.4. Assay for Coinfection and RNA Recombination of Viral Strains. Coinfection of JEV strains was verified by a method described in our previous report [25]. In brief, extracted viral RNAs were applied to perform the RT-PCR with the primer pair, 10-36F (5′-CTGTGTGAACTT CTTGGCTTAGTATCG-3′) and 850-877R (5′-CAGTTTTCATGAGATATCGTGTGTGGC-3′). Fragments (868 bp) amplified from JEV strains simultaneously infecting BHK-21 or C6/36 cells were subjected to restriction fragment length polymorphism (RFLP) with the restriction enzyme RsaI to verify coinfection. A pattern showing fragments of 219, 401, and 248 bp represented T1P1-S1 infection, while that showing fragments of 219 and 649 bp represented CJN-S1 infection. Those exhibiting all size of fragments indicated that both viral strains had coinfected a single cell. In addition, RFLP using specific restriction enzymes as shown in our previous report [25] was used to verify RNA recombination between viral strains in a single cell. In some experiments for assay of RNA recombination, RFLP was carried out by using cells cultured in the presence of an exoribonuclease inhibitor (3′-phosphoadenosine-5′-phosphate, PAP) (Sigma-Aldrich, St. Louis, MO, USA).

2.5. Construction of the Plasmid p(+)T1P1-5′3′-Untranslated Region- (UTR-) I. In order to evaluate RNA recombination between genomic RNA and a transfected RNA sequence, the p(+)T1P1-5′3′-UTR-1 plasmid was constructed as described here. Viral RNA derived from the T1P1 strain of the JEV was used as a template to generate DNA fragments corresponding to the 5′- or 3′-end of genomic-sense RNA. To prepare the 5′-end sequence, a primer (5′-CTGCCAAGCATCCAGCCAAGTA-3′, complementary to nt 895~916 of the 5′-end of the T1P1 genome) was used for RT to synthesize the first-strand cDNA. Subsequently, another primer (5′-TAATACGACTCACTATAGAGAAGTTTATCTGTGTG-3′) containing a partial sequence of the T7 polymerase promoter used as a tag (italicized) at the 5′-end (nt 1~18) of the T1P1 5′-end sequence was used in the PCR to amplify a 934 bp DNA fragment. In the meantime, the primer 5′-GTGTTCTTCCTCACCACCAGCTAC-3′ (nt 10,946~10,969 at the 3′-end of the T1P1 genome) was used for RT to generate cDNA. Another primer (5′-GAAAATTATGTTGACTAC-3′, corresponding to the sequence nt 10,320~10,337) was subsequently used for the PCR under conditions described above to amplify a 650 bp DNA fragment. Both types of PCR products were separately digested with the restriction enzyme, AatII; the resultant DNA fragments were ligated to form subgenomic DNA which contained both 5′-end (nt 1~599) and 3′-end (nt 10,367~10,969) sequences. Subsequently, the subgenomic DNAs were cloned into pGEM-T (Promega, Madison, WI, USA) to form a plasmid designated p(+)T1P1-5′3′-UTR-I which contained an insert of a 1202 bp fragment.

2.6. Construction of the p(+)5′3′-UTR-II Plasmid. In order to see the stability of viral fragments in host cells, the p(+)5′3′-UTR-II plasmid was constructed. To construct the plasmid, the pT1P1-5′3′-UTR was used as a template, and the PCR was performed under conditions described above with the primers 5′-TAATACGACTCA-CTATAGAGAAGTTTATCTGTGTG-3′ (the italics indicate a partial T7 polymerase promoter sequence) and 5′-AAGATATCGTGTTCTTCCTCAC CACC-3′ (the italics indicate an EcoRV restriction enzyme site). The PCR

products were digested with SpeI and AatII to delete a fragment from nt 178~599; the resultant DNA fragments were then treated with Klenow Fragment enzyme (Fermentas, Hanover, MD, USA) and ligated to form subgenomic DNA which only contained the 5′- and 3′-UTRs of the T1P1 genome. Subsequently, the subgenomic DNAs were cloned into pGEM-T (Promega, Fitchburg, WI, USA) to form plasmids designated p(+)5′3′-UTR-II.

2.7. Preparation of the Positive (+) and Negative (−) Sense 5′-End RNA Sequences and Derived dsRNA. Both (+) and (−) sense 5′-end RNA sequences were prepared from the pT1P1-5′3′-UTR-II plasmid. In preparation of the (+) sense 5′-end RNA sequence, the plasmid was linearized by NdeI and transcribed with T7 RNA polymerase using an in vitro transcription system (Fermentas). The RNA products (599 bp) were extracted with phenol-chloroform, precipitated in ethanol, and then stored in a deep freezer until used for transfection. To prepare the (−) sense 5′-end RNA sequence, the plasmid was first linearized, and the 3′-end sequence of the subgenomic DNA was deleted with NdeI. The resultant linear forms of the plasmid were religated and then redigested with SacII. The products were transcribed with T7 RNA polymerase using an in vitro transcription system (Fermentas) to generate the (−) sense 5′-end RNA sequence which was harvested as done for the (+) sense 5′-end RNA sequence. To prepare dsRNA, positive- and negative-stranded RNA described from pT1P1-5′3′-UTR-II were mixed together, incubated at 95°C for 5 min and then 4°C for 10 min. Ultimately, 2 μL RNase was added to cleave single-stranded RNA that failed to anneal in the mixture. The product was used to evaluate degradation of dsRNA fragments in host cells.

2.8. Transfection of dsRNA or the (+) sense 5′-end RNA Sequence and Viral Infection in Cells. Transfection of dsRNA or (+) sense 5′-end RNA-I prepared from the plasmids (+) pT1P1-5′3′-UTR-II was carried out in BHK-21 and C6/36 cells. At 5 h posttransfection (hpt), cells were infected with the Nakayama strain of the JEV, at an MOI of 5. The detailed procedure followed a previous description, from which efficacy of transfection was demonstrated [25].

2.9. Assessment of RNA Stability by an RT-PCR. Sequences derived from (+)5′3′-UTR-II RNA were transfected into cells either treated or untreated with an exoribonuclease inhibitor (3′-phosphoadenosine-5′-phosphate, PAP) (Sigma-Aldrich, St. Louis, MO, USA) and incubated for 5 h. RNA was extracted with the TRIzol reagent (5 PRIME, Gaithersburg, MD, USA), and then DNAse (Promega) was added to delete interference of genomic DNA. RT was subsequently run in a mixture containing 4 μg RNA, 1 μL 100 mM random hexamer primer, and 1 μL 10 mM dNTP, and double-distilled (dd) H_2O water was added to bring the volume up to 12 μL. This was heated at 65°C for 5 min, allowed to stand at 4°C for 2 min, and then 4 μL 5x first-strand buffer, 2 μL dithiothreitol (DTT), 1 μL of an RNase inhibitor (RNase OUTTM; Invitrogen, Carlsbad, CA, USA) were added. After incubation at room temperature for 5 min, 1 μL of reverse transcriptase M-MLV (Invitrogen) was added and allowed to react for 1 h at 37°C, followed by 15 min at 75°C. The cDNA produced was then used for the subsequent PCR under the conditions described above. The primers used for the PCR included the 5′-UTR (5′-AGAAGTTTATCTGTGTGAAC-3′) and 3′-UTR (5′-AGATCCTGTGTTCTTCC-3), which generated PCR products predicted to be 907 bp. To assess integration of dsRNA, the same cDNA and primer pair described above were used to amplify a fragment 807 bp long.

2.10. Assay for RNA Recombination from Transfected as well as Infected Cells. BHK-21 and C6/36 cells transfected with transcribed RNA fragments were then infected by the Nakayama strain of the JEV. Total RNA extracted from cells that had been transfected with (+) sense 5′-end RNA-1 was run for RT using a primer (850-877R: 5′-TCAGTTTTC-ATGAGATATCGTGTGTGGC-3′) complementary to the sequence of nt 850~877. Amplification using the forward primer (RVF1: 5′-GCGGGATTTAATACGACTCACTAT-AG-3′) which is a partial sequence of the plasmid that serves as a tag and the reverse primer (RVR1/nt 516~538: 5′-CTGCAATATCCGTATTGTTGAC-3′) produced a specific region comprised of 564 nt. The reverse primer used here was specific for the Nakayama strain. As a result, the fragments amplified by this primer pair must represent a strain of genetic recombination.

2.11. Measurement of Viral RNA Accumulated in Cells Infected by the JEV. Viral replication was validated by RNA accumulation through a real-time RT-PCR with cDNAs reverse-transcribed from extracted RNA of infected (at an MOI of 1) or uninfected C6/36 and BHK-21 cells. The primer pairs TS1-F/TS1R (5′-TGTGGCTTGCGAGCTTGGCAG-3′/5′-ACATGTAGCCGACGTCGATT-3′) and CJN1-F/CJN1R (5′-TGTGGCTTGCGAGCTTGGCTA-3′/5′-ACATGTAGCCGACGTCTATC-3′) were used to amplify specific regions of the T1P1-S1 and CJN-S1 strains, respectively. Levels of 18S rRNA designed from the genome of C6/36 or BHK-21 cells were also amplified as an internal control as our previous report [25]. Results are expressed as the relative quantities, so fold change was used to represent the amount of viral RNA that accumulated at each time point of infection. To monitor synthesis of viral RNA including positive and negative strands in a time course in C6/36 cells, viral RNA extracted from infected cells (0~15 hpi) was used to run RT-PCR as the procedures described previously [27]. As above, 18S rRNA designed from the genome of C6/36 cells was also amplified as an internal control. The amplified cDNA fragment was then identified by running the PCR product on a 2% (w/v) agarose gel.

2.12. Statistical Analysis. Yates' chi-square test was used to assess the frequency of RNA recombination in cells coinfected by two virus strains or transfected by viral RNA fragments.

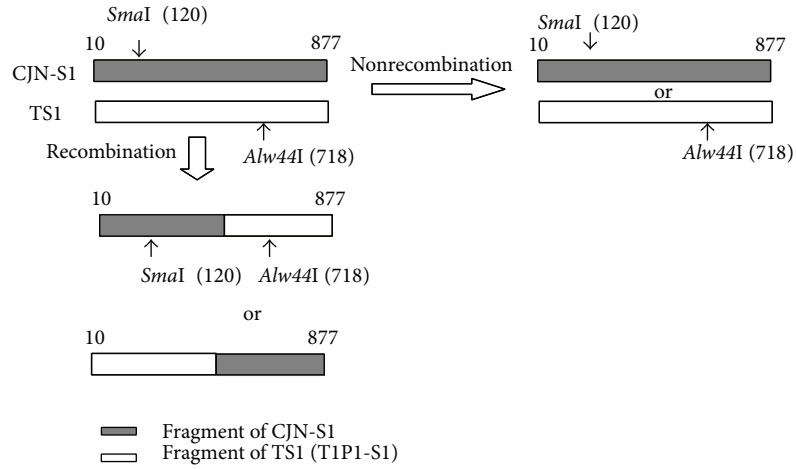


Figure 1: The schematic sketch designed to identify RNA recombination between viral strains. A fragment (868 bp) comprised of the C/preM junction (nt 10~877) of viral RNA extracted from coinfected BHK-21 or C6/36 cells was amplified, cloned, and then used for an RFLP analysis with *Sma*I or *Alw44*I. Two and one recombinant form(s) were, respectively, identified in selected samples from BHK-21 and C6/36 cells, when they were coinfected with the T1P1-S1 and CJN-S1 strains of the Japanese encephalitis virus.

Table 1: Identification of RNA recombination of the Japanese encephalitis virus based on a fragment (868 bp) comprised of the C/preM junction (nt 10~877) of viral RNA extracted from coinfected BHK-21 or C6/36 cells using an RFLP analysis with restriction enzymes *SmaI* or *Alw44I*.

Treatment	BHK-21 cells		C6/36 cells	
	Number of detection	Number of recombination	Number of detection	Number of recombination
Coinfected viral genomic RNA	98	20 (20.4%)	38	5 (13.1%)
Mixed RNA*	44	2 (4.5%)	39	3 (7.7%)
Stastistical analysis**	$P < 0.05$		$P > 0.05$	

*Mixed RNA was a mixture of RNAs separately extracted from T1P1-S1 and CJN-S1 strains of the Japanese encephalitis virus, being used as the internal control.
**Yates' chi-square test was used to assess the difference of RNA recombination in cells coinfected by two virus strains at 5% level of significance.

3. Results

3.1. RNA Recombination in BHK-21 Cells and C6/36 Cells. Viral RNA extracted from single infectious centers (ICs) which were randomly selected and picked out from infected BHK-21 or C6/36 cells was subjected to an *Rsa*I RFLP assay as described in our previous report. The result reveals that different strains of the JEV can coinfect a single BHK-21 or C6/36 cell. The C/preM junction comprising 868 nucleotides (nt 10~877) of viral RNA extracted from BHK-21 or C6/36 cells coinfected with the TaP1-S1 and CJN-S1 strains was cloned and used for the *Sma*I-*Alw44*I RFLP analysis (Figure 1). The recombinant forms of the viral genome were actually identified in BHK-21 and C6/36 cells, when they were coinfected by the 2 strains of the JEV. Totally, 20 recombination clones (20.4%) were found from 98 clones coinoculated with the 2 strains in BHK-21 cells while being 5 out 38 (13.1%) in C6/36 cells (Table 1). Probability of occurring RNA recombination was significantly different, compared with the mixed RNA control, in BHK-21 cells while being nonsignificant in C6/36 cells (Table 1). In other words, the frequency of RNA recombination is significantly higher in BHK-21 cells than in C6/36 cells.

3.2. Recombination between Genomic RNA and a Transfected RNA Fragment of the Virus. A 564 bp fragment was significantly amplified in BHK-21 and C6/36 cells which were infected by the JEV (Nakayama strain) following transfection with the (+)5′3′-UTR-I RNA plasmid, although a light band was also shown in the control group that contained a mixture of RNAs extracted from transfected cells.A specific fragment of viral RNA (529 bp) was amplified as an internal control in all groups with viral infection. In addition, no fragment presenting an artifact of RNA recombination was shown in the control groups of mock treatment (neither infection nor transfection), transfection with only the (+)5′3′-UTR RNA-I plasmid, or infection with only a single strain. An image-density analysis revealed recombination in BHK-21 cells to be 10.7-fold higher than that of the control group, while it was 7.73-fold higher in C6/36 cells, suggesting that RNA recombination may occur in both mammalian and mosquito cells. However, a slightly lower frequency of RNA recombination was eventually shown in mosquito cells (Figure 2).

3.3. Enzymatic Effect on RNA Stability Modulates RNA Recombination between Genomic RNA and a Transfected RNA Fragment of the Virus. The RNA recombination rate was shown to have increased to a higher level in BHK-21 cells

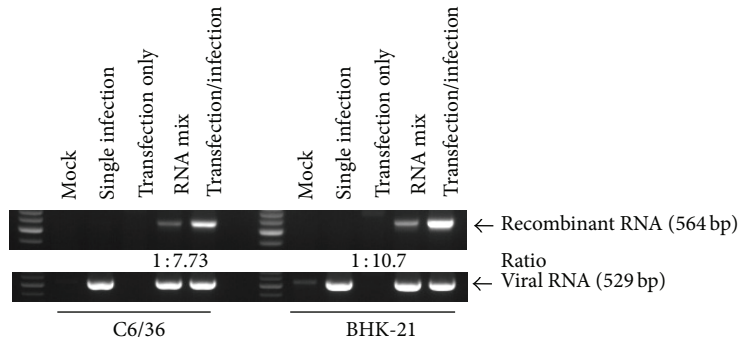


Figure 2: RNA recombination between genomic RNA and a transfected RNA sequence. No fragment was seen in tests with RNA extracted from cells following mock treatment (with neither infection nor transfection), virus infection only, or transfection only. Although amplification of a 564 bp fragment showing RNA recombination was present in the control group which contained a mixture of RNAs extracted from infected and transfected cells, RNA recombination was significantly elevated in BHK-21 and C6/36 cells infected by the Japanese encephalitis virus (Nakayama strain) following transfection with the (+)5′3′-UTR-I plasmid RNA. According to the image-density analysis, it seems that RNA recombination occurred less frequently in mosquito cells. A specific fragment of viral RNA (529 bp) was used as an internal control in all groups with viral infection.

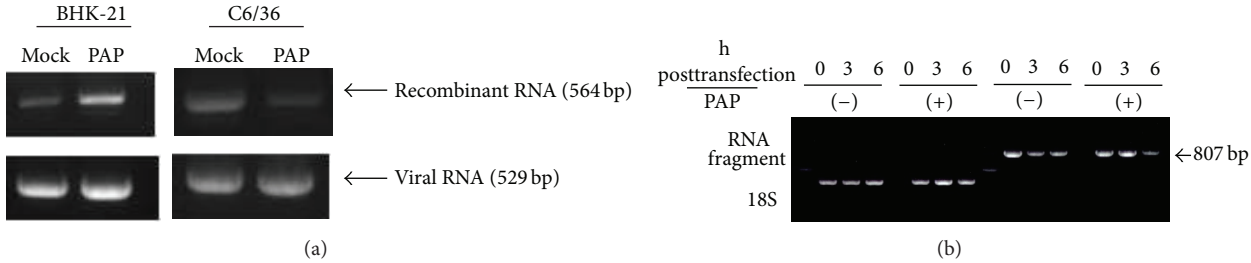


Figure 3: Status of RNA recombination after inhibition by exoribonuclease with PAP (3′-phosphoadenosine-5′-phosphate, an inhibitor of exoribonuclease). (a) The RNA recombination rate increased to a higher level in BHK-21 cells after treatment with PAP, compared to that of untreated cells. In contrast, no effect of PAP on increasing RNA recombination of the virus was shown in C6/36 cells despite a very low level of RNA recombination still being observed. Viral RNA was not affected after treatment with PAP, suggesting exoribonuclease-mediated degradation of transfected RNA fragments might increase RNA recombination of the virus strains, especially in mammalian cells. (b) Treatment with PAP in C6/36 cells did not cause degradation of the transfected (+) RNA fragment at 3 h until 6 h after transfection at which a partial effect appeared.

treated with PAP, the inhibitor of exoribonuclease, compared to untreated cells. In contrast, no effect of PAP on increasing RNA recombination was seen in C6/36 cells; only a low level of RNA recombination was found in this test (Figure 3(a)). Looking at transfected viral fragment (+) RNA in C6/36 cells treated with PAP, enzymatic cleavage by exoribonuclease did not occur at 3 h after transfection while it evidently decreased preservation of such RNA fragment at 6 h after transfection in mosquito cells (Figure 3(b)). Viral genomic RNA was not affected when treated with PAP in both cell types (Figure 3(a)), implying that the transfected viral RNA fragment may not be further degraded by exoribonuclease mostly in mammalian cells; which leads to a higher possibility of occurring RNA recombination in such cells.

3.4. Assessment to the Enzymatic Effect on RNA Recombination in Mosquito Cells with Coinfection by Two Different Virus Strains. When we coinfected T1P1-S1 and CJN-S1 strains of JEV into C6/36 cells and treated with PAP, only 1 of 30 clones occurred RNA recombination while 4 out of 31 clones occurred in the control group (without treatment with PAP). The RNA recombination rate did not change significantly (P value = 0.370; Yates' chi-square test) in coinfected C6/36 cells and even their function of exoribonuclease was inhibited and thus unable to dissolve viral RNA (Table 2). The result implicated that the low level of viral RNA at the early phase of infection may not be fully exoribonuclease-mediated but, as above, is probably contributed by the RNAi-dependent effect.

3.5. Fate of Transfected dsRNA Fragments in Mosquito Cells. The dsRNA intermediates are generally formed during virus replication in host cells, however, which may be cleaved in invertebrate cells. Through an RT-PCR, a corresponding segment of RNA (807 bp) was detected in C6/36 cells immediately after transfection (0 hpt) with a fragment of dsRNA derived from (+) or (−) 5′3′-UTR RNA; however, it had faded by 3 and 6 hpt (Figure 4). This suggests that transfected dsRNAs may have been cleaved and presumably generated short interfering (si)RNAs which were not shown on the gel. It suggested that a part of viral RNAs may be degraded at the early phase of infection, likely to modulate virus growth, in mosquito cells.

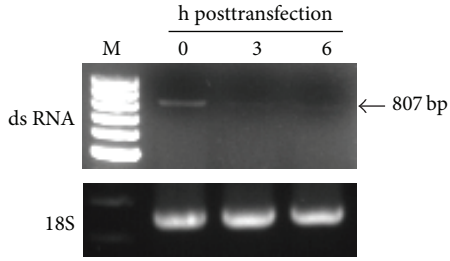


Figure 4: Degradation of double-stranded (ds) RNA fragments transfected into mosquito cells. A fragment (807 bp) of viral RNA extracted from C6/36 cells was detected through an RT-PCR at 0 h after transfection (hpt) with dsRNA derived from (+) or (−) 5′3′-UTR RNA. The transfected dsRNA had faded at 3 and 6 h after transfection, suggesting that dsRNAs may have been cleaved, and thus generated undetectable short interfering RNAs.

Table 2: Detection of RNA recombination in C6/36 cells simultaneously infected by T1P1-S1 and CJN-S1 in the presence of PAP (3′-phosphoadenosine-5′-phosphate, an inhibitor of exoribonuclease).

Recombination	PAP treatment		
	−	+	Total
+	4	1	6
−	27	29	85
Total	**31**	**30**	**91**

Yates' chi-square test was used to assess the difference of RNA recombination (P value = 0.370).

3.6. Differential RNA Accumulation of Japanese Encephalitis Virus during Early Infection. Appropriate accumulation of viral RNA in host cells is essential for prosperous production of progeny virions. According to the results, RNAs of both the T1P1-S1 and CJN-S1 strains accumulated more slowly in C6/36 cells than BHK-21 cells (Figure 5(a)). Specifically, the RNA amount of the T1P1-S1 strain remained at the baseline level until 12 hpi (3.81-fold change), compared with an increase of 169.72-fold at 24 hpi in C6/36 cells. In contrast, T1P1-S1 RNA, respectively, increased to 3.09-, 28.99-, 429.05-, 4396.07-, and 5487.75-fold, at 3, 6, 9, 12, and 24 hpi in BHK-21 cells. Similarly, the RNA amount of CJN-S1 also accumulated more slowly in C6/36 cells than in BHK-21 cells. The RNA amount remained at the baseline level until 12 hpi (2.36-fold increase) and subsequently increased to 152.32-fold at 24 hpi in C6/36 cells. In contrast, the RNA amount of CJN-S1 RNA, respectively, increased by 16.64-, 111.43-, and 554.87-fold at 9, 12, and 24 hpi, despite it having been unchanged at 6 hpi (1.35-fold change) in BHK-21 cells. The result revealed that progeny RNA of the virus is delayed to accumulate in mosquito compared to mammalian cells, especially at the early phase of infection. The stability of viral RNA is crucial for the productivity of the progeny virions, which was evaluated after transfection of an RNA fragment prepared from (+)5′3′-UTR-II into either BHK-21 or C6/36 cells. Results showed that transfected fragments had not significantly degraded even at 3 or 6 hpt in BHK-21 cells, while those in C6/36 cells had more obviously degraded (Figure 5(b)), implying that different outcomes of RNA existed in the 2 cell types especially in the early phase of infection.

4. Discussion

RNA viruses generate new genetic strains with approximately 6 orders of magnitude higher rates of nucleotide substitutions compared to DNA viruses [28]. Thus, the rate of spontaneous mutations is a critical parameter modeling the genetic structure of viral populations [29]. The primary variation following a mutation may provide for further evolutionary processes, for example, selection and/or recombination [30]. Those in turn lead to the generation of viral strains which are more adept and fit in nature. RNA recombination is now believed to be a strategy for the evolution of many viruses [12], for instance, the poliovirus [31], hepatitis C virus [32], hepatitis D virus [33], and norovirus [34]. A variety of flaviviruses including dengue virus and JEV were also reported to carry out RNA recombination according to bioinformatics inferences [17, 35] and experimental demonstration [25].

Currently, 2 possible mechanisms are reported to lead to the occurrence of recombination [36]: a copy-choice mechanism and a breakage and rejoining mechanism. Of these, the former apparently occurs more commonly as it has been shown in the poliovirus [37], coronaviruses [38], and plant viruses [39]. This mechanism of viral RNA recombinations can further be divided into 3 types: precisely homologous, imprecisely (aberrantly) homologous, and nonhomologous [16]. Among these, precisely homologous recombination through a template-switching (copy-choice) mechanism is probably most common [40]. As in our previous report, different strains of the JEV can coinfect host cells derived from mosquitoes or mammals [25], which actually generates recombinant forms of the virus [30].

In this study, we infected host cells with Nakayama strains of the JEV, followed by transfection of the (+)5′3′-UTR-I RNA fragment. The result was parallel to our previous observation [25], showing imbalanced RNA recombination between BHK-21 and C6/36 cells. Looking at RNA accumulation of JEV in host cells, it takes longer, at least a 24 h difference, in mosquito cells to reach the level of that in mammalian cells. The stability of viral RNA was also shown by degradation of transfected single-stranded RNA fragments, either positive or negative sense, particularly in mosquito cells. It implicated that viral RNA is less stable at least at the early phase of infection by JEV in mosquito cells, which may result in delayed growth of the virus. Since transfected RNA fragments were degraded in both C6/36 cells and BHK-21 cells, RNase cleavage may be actually involved in viral RNA degradation to form small RNAs [41]. However, degradation of transfected RNA fragments in C6/36 cells was partially ameliorated by treatment with PAP, suggesting that viral RNAs are not completely degraded by the RNase cleavage pathway [42]. Perhaps RNA interference (RNAi) plays an important role in the related events [43].

Generally double-stranded replicative-form RNA (dsRF-RNA) accumulates to provide an immediate signal which activates specific transcription factors such as type-I interferon (IFN) [44] and facilitates the triggering of intracellular

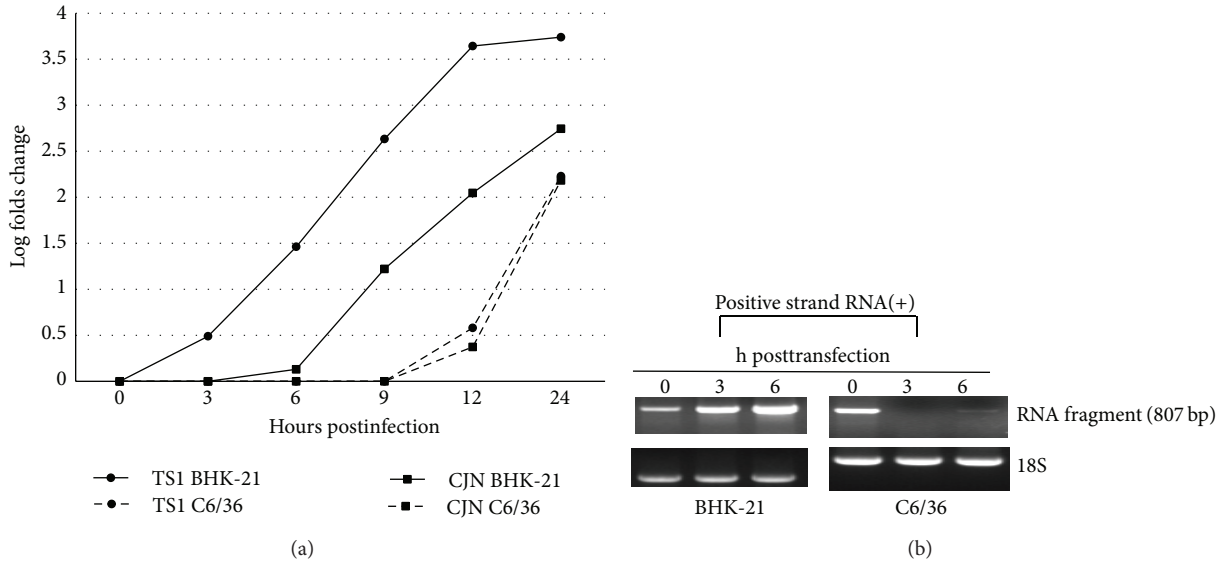


Figure 5: Viral RNA, either T1P1-S1 or CJN-S1, accumulated in C6/36 cells more slowly than in BHK-21 cells. (a) The RNA amount of T1P1-S1 remained at the baseline level until 12 h after infection (hpi) (3.81-fold change), which increased to 169.72-fold at 24 hpi in C6/36 cells. In contrast, T1P1-S1 RNA, respectively, increased to 3.09-, 28.99-, 429.05-, 4396.07-, and 5487.75-fold, at 3, 6, 9, 12, and 24 hpi in BHK-21 cells. The RNA amount of CJN-S1 also accumulated more slowly in C6/36 cells than BHK-21 cells, which remained at the baseline level until 12 hpi (2.36-fold increase) and had increased to 152.32-fold by 24 hpi in C6/36 cells. Although the amount of CJN-S1 RNA did not evidently increase until 6 hpi (1.35-fold change), it increased to 16.64-, 111.43-, and 554.87-fold at 9, 12, and 24 hpi, respectively, in BHK-21 cells. (b) Stability of viral RNA was evaluated after a fragment of (+)5′3′-UTR-II RNA was transfected into BHK-21 or C6/36 cells. Transfected fragments were insignificantly degraded even at 3 or 6 h after transfection in BHK-21 cells while more obvious degradation appeared in C6/36 cells.

innate immunity in mammalian cells [45]. On the other hand, dsRNAs formed in invertebrate cells are usually cleaved to be siRNA that consequently degrades viral RNAs [46], leading to RNAi-mediated innate immunity [47]. Small RNAs ranging from 10 to 24 mer have been identified in C6/36 cells infected by West Nile virus [43]. We have also detected normal expression of Dicer-2 in C6/36 cells infected by JEV for 12 h although it was almost half of inhibition at 6 hpi (data not shown). As a result, dsRF-RNA of the JEV may have a great potential for viral RNA degradation at least in the early phase of infection in mosquito cells. This adjustment of RNA amount is believed to be the way for a delay in RNA accumulation and thus a lower frequency of RNA recombination of the JEV. In contrast, dsRNAs are recognized as a central component of IFN and therefore are incapable of mediating RNAi in mammalian cells [48]. Eventually, our results have shown that protection of RNA from RNase cleavage increases the efficiency of RNA recombination particularly in mammalian cells.

RNA recombination creates advantageous genotypes by evolutionary jumps [30], which permits the removal of deleterious genes based on the notion of "Muller's ratchet" from the host cell, usually mammalian cells [12]. Notably, viral RNAs usually accumulated at a lower amount in mosquito cells through RNase cleavage as well as RNA-mediated pathways, leading to stagnancy of RNA recombination which may brake evolution of the JEV and probably most, if not all, arboviruses which are maintained in nature by alternate cycles involving mosquitoes and vertebrates [28].

Conflict of Interests

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

Authors' Contribution

Wei-Wei Chiang and Ching-Kai Chuang contributed equally to this work.

Acknowledgments

The authors thank Yi-Hsuan Chiang, Chao-Fu Yang, and Tien-Huang Chen for their technical assistance in this study. This work was supported by a grant from Chang Gung Memorial Hospital (CMRPD 190161~3) and partly from the Ministry of Science and Technology, Executive Yuan, Taiwan (NSC100-2313-B-182-001-MY3).

References

[1] Y. F. Tseng, C. C. Wang, S. K. Liao, C. K. Chuang, and W. J. Chen, "Autoimmunity-related demyelination in infection by Japanese encephalitis virus," *Journal of Biomedical Science*, vol. 18, article 20, 2011.

[2] T. Solomon, "Flavivirus encephalitis," *The New England Journal of Medicine*, vol. 351, no. 4, pp. 370–409, 2004.

[3] B. D. Lindenbach and C. M. Rice, "Flaviviridae: the viruses and their replication," in *Fields Virology*, D. M. Knipe and P.

M. Howley, Eds., vol. 1, pp. 991–1041, Lippincott Williams & Wilkins, Philadelphia, Pa, USA, 4th edition, 2001.

[4] H. Sumiyoshi, C. Mori, I. Fuke et al., "Complete nucleotide sequence of the Japanese encephalitis virus genome RNA," *Virology*, vol. 161, no. 2, pp. 497–510, 1987.

[5] E. Domingo, "Quasispecies theory in virology," *Journal of Virology*, vol. 76, no. 1, pp. 463–465, 2002.

[6] E. Duarte, D. Clarke, A. Moya, E. Domingo, and J. Holland, "Rapid fitness losses in mammalian RNA virus clones due to Muller's ratchet," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 89, no. 13, pp. 6015–6019, 1992.

[7] W. L. Schneider and M. J. Roossinck, "Genetic diversity in RNA virus quasispecies is controlled by host-virus interactions," *Journal of Virology*, vol. 75, no. 14, pp. 6566–6571, 2001.

[8] N. Vasilakis, E. R. Deardorff, J. L. Kenney, S. L. Rossi, K. A. Hanley, and S. C. Weaver, "Mosquitoes put the brake on arbovirus evolution: experimental evolution reveals slower mutation accumulation in mosquito than vertebrate cells," *PLoS Pathogens*, vol. 5, no. 6, Article ID e1000467, 2009.

[9] W.-J. Chen, H.-R. Wu, and S.-S. Chiou, "E/NS1 modifications of dengue 2 virus after serial passages in mammalian and/or mosquito cells," *Intervirology*, vol. 46, no. 5, pp. 289–295, 2003.

[10] L. Chao and T. T. Tran, "The advantage of sex in the RNA virus φ6," *Genetics*, vol. 147, no. 3, pp. 953–959, 1997.

[11] C. K. Chuang and W. J. Chen, "Genetic evolution of Japanese encephalitis virus," in *Flavivirus Encephalitis*, D. Růžek, Ed., pp. 383–404, InTech Open Access, Rijeka, Croatia, 2011.

[12] M. Worobey and E. C. Holmes, "Evolutionary aspects of recombination in RNA viruses," *Journal of General Virology*, vol. 80, no. 10, pp. 2535–2543, 1999.

[13] P. D. Cooper, "A genetic map of poliovirus temperature-sensitive mutants," *Virology*, vol. 35, no. 4, pp. 584–596, 1968.

[14] L. Mindich, X. Qiao, S. Onodera, P. Gottlieb, and J. Strassman, "Heterologous recombination in the double-stranded RNA bacteriophage φ6," *Journal of Virology*, vol. 66, no. 5, pp. 2605–2610, 1992.

[15] M. Figlerowicz and J. J. Bujarski, "RNA recombination in brome mosaic virus, a model plus strand RNA virus," *Acta Biochimica Polonica*, vol. 45, no. 4, pp. 847–868, 1998.

[16] M. Alejska, A. Kurzyńska-Kokorniak, M. Broda, R. Kierzek, and M. Figlerowicz, "How RNA viruses exchange their genetic material," *Acta Biochimica Polonica*, vol. 48, no. 2, pp. 391–407, 2001.

[17] S. S. Twiddy and E. C. Holmes, "The extent of homologous recombination in members of the genus Flavivirus," *Journal of General Virology*, vol. 84, no. 2, pp. 429–440, 2003.

[18] E. R. Chare and E. C. Holmes, "A phylogenetic survey of recombination frequency in plant RNA viruses," *Archives of Virology*, vol. 151, no. 5, pp. 933–946, 2006.

[19] C. S. Hahn, S. Lustig, E. G. Strauss, and J. H. Strauss, "Western equine encephalitis virus is a recombinant virus," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 85, no. 16, pp. 5997–6001, 1988.

[20] H. J. G. Tolou, P. Couissinier-Paris, J. P. Durand et al., "Evidence for recombination in natural populations of dengue virus type 1 based on the analysis of complete genome suquences," *Journal of General Virology*, vol. 82, no. 6, pp. 1283–1290, 2001.

[21] E. C. Holmes, M. Worobey, and A. Rambaut, "Phylogenetic evidence for recombination in dengue virus," *Molecular Biology and Evolution*, vol. 16, no. 3, pp. 405–409, 1999.

[22] P. Becher, M. Orlich, and H.-J. Thiel, "RNA recombination between persisting pestivirus and a vaccine strain: generation of cytopathogenic virus and induction of lethal disease," *Journal of Virology*, vol. 75, no. 14, pp. 6256–6264, 2001.

[23] A. T. Ciota, A. O. Lovelace, S. A. Jones, A. Payne, and L. D. Kramer, "Adaptation of two flaviviruses results in differences in genetic heterogeneity and virus adaptability," *Journal of General Virology*, vol. 88, no. 9, pp. 2398–2406, 2007.

[24] T. Kurosu, "Quasispecies of dengue virus," *Tropical Medicine and Health*, vol. 39, supplement 4, pp. 29–36, 2011.

[25] C. K. Chuang and W. J. Chen, "Experimental evidence that RNA recombination occurs in the Japanese encephalitis virus," *Virology*, vol. 394, no. 2, pp. 286–297, 2009.

[26] S. C. Weaver, R. Rico-Hesse, and T. W. Scott, "Genetic diversity and slow rates of evolution in new world alphaviruses," *Current Topics in Microbiology and Immunology*, vol. 176, pp. 99–117, 1992.

[27] S.-S. Chiou and W.-J. Chen, "Mutations in the NS3 gene and 3$\prime$-NCR of Japanese encephalitis virus isolated from an unconventional ecosystem and implications for natural attenuation of the virus," *Virology*, vol. 289, no. 1, pp. 129–136, 2001.

[28] G. M. Jenkins, A. Rambaut, O. G. Pybus, and E. C. Holmes, "Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis," *Journal of Molecular Evolution*, vol. 54, no. 2, pp. 156–165, 2002.

[29] J. W. Drake and J. J. Holland, "Mutation rates among RNA viruses," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 24, pp. 13910–13913, 1999.

[30] E. Simon-Loriere and E. C. Holmes, "Why do RNA viruses recombine?" *Nature Reviews Microbiology*, vol. 9, no. 8, pp. 617–626, 2011.

[31] N. Ledinko, "Genetic recombination with poliovirus type 1. Studies of crosses between a normal horse serum-resistant mutant and several guanidine-resistant mutants of the same strain," *Virology*, vol. 20, no. 1, pp. 107–119, 1963.

[32] J. Cristina and R. Colina, "Evidence of structural genomic region recombination in Hepatitis C virus," *Virology Journal*, vol. 3, article 53, 2006.

[33] T. Wang and M. Chao, "RNA recombination of hepatitis delta virus in natural mixed-genotype infection and transfected cultured cells," *Journal of Virology*, vol. 79, no. 4, pp. 2221–2229, 2005.

[34] J. Rohayem, J. Münch, and A. Rethwilm, "Evidence of recombination in the norovirus capsid gene," *Journal of Virology*, vol. 79, no. 8, pp. 4977–4990, 2005.

[35] J. Carney, J. M. Daly, A. Nisalak, and T. Solomon, "Recombination and positive selection identified in complete genome sequences of Japanese encephalitis virus," *Archives of Virology*, vol. 157, no. 1, pp. 75–83, 2012.

[36] M. M. C. Lai, "RNA recombination in animal and plant viruses," *Microbiological Reviews*, vol. 56, no. 1, pp. 61–79, 1992.

[37] K. Kirkegaard and D. Baltimore, "The mechanism of RNA recombination in poliovirus," *Cell*, vol. 47, no. 3, pp. 433–443, 1986.

[38] S. Makino, J. G. Keck, S. A. Stohlman, and M. M. C. Lai, "High-frequency RNA recombination of murine coronaviruses," *Journal of Virology*, vol. 57, no. 3, pp. 729–737, 1986.

[39] J. Sztuba-Solińska, A. Urbanowicz, M. Figlerowicz, and J. J. Bujarski, "RNA-RNA recombination in plant virus replication and evolution," *Annual Review of Phytopathology*, vol. 49, pp. 415–443, 2011.

[40] R. Wierzchoslawski and J. J. Bujarski, "Efficient in vitro system of homologous recombination in brome mosaic bromovirus," *Journal of Virology*, vol. 80, no. 12, pp. 6182–6187, 2006.

[41] K. Lin, H. Chang, and R. Chang, "Accumulation of a 3′-terminal genome fragment in Japanese encephalitis virus-infected mammalian and mosquito cells," *Journal of Virology*, vol. 78, no. 10, pp. 5133–5138, 2004.

[42] S. Slomovic and G. Schuster, "Exonucleases and endonucleases involved in polyadenylation-assisted RNA decay," *Wiley Interdisciplinary Reviews: RNA*, vol. 2, no. 1, pp. 106–123, 2011.

[43] D. E. Brackney, J. C. Scott, F. Sagawa et al., "C6/36 *Aedes albopictus* cells have a dysfunctional antiviral RNA interference response," *PLoS Neglected Tropical Diseases*, vol. 4, no. 10, article e856, 2010.

[44] G. R. Stark, I. M. Kerr, B. R. G. Williams, R. H. Silverman, and R. D. Schreiber, "How cells respond to interferons," *Annual Review of Biochemistry*, vol. 67, pp. 227–264, 1998.

[45] I. Jensen and B. Robertsen, "Effect of double-stranded RNA and interferon on the antiviral activity of Atlantic salmon cells against infectious salmon anemia virus and infectious pancreatic necrosis virus," *Fish & Shellfish Immunology*, vol. 13, no. 3, pp. 221–241, 2002.

[46] N. J. Caplen, Z. Zheng, B. Falgout, and R. A. Morgan, "Inhibition of viral gene expression and replication in mosquito cells by dsRNA-triggered RNA interference," *Molecular Therapy*, vol. 6, no. 2, pp. 243–251, 2002.

[47] C. D. Blair, "Mosquito RNAi is the major innate immune pathway controlling arbovirus infection and transmission," *Future Microbiology*, vol. 6, no. 3, pp. 265–277, 2011.

[48] M. P. Gantier and B. R. G. Williams, "The response of mammalian cells to double-stranded RNA," *Cytokine and Growth Factor Reviews*, vol. 18, no. 5-6, pp. 363–371, 2007.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 736798, 7 pages
http://dx.doi.org/10.1155/2014/736798

Research Article

Computational Evidence of NAGNAG Alternative Splicing in Human Large Intergenic Noncoding RNA

Xiaoyong Sun,[1] Simon M. Lin,[2] and Xiaoyan Yan[3]

[1] *Agricultural Big-Data Research Center, College of Information Science and Engineering, Shandong Agricultural University, Taian, Shandong 271018, China*
[2] *Biomedical Informatics Research Center, Marshfield Clinic Research Foundation, Marshfield, WI 54449, USA*
[3] *Affiliated Hospital of Shandong University of Traditional Chinese Medicine, No. 42 Wenhua West Road, Jinan, Shandong 250011, China*

Correspondence should be addressed to Xiaoyong Sun; johnsunx1@gmail.com and Simon M. Lin; linmd.simon@mcrf.mfldclin.edu

Received 4 February 2014; Revised 8 May 2014; Accepted 21 May 2014; Published 5 June 2014

Academic Editor: Shiwei Duan

NAGNAG alternative splicing plays an essential role in biological processes and represents a highly adaptable system for posttranslational regulation of gene function. NAGNAG alternative splicing impacts a myriad of biological processes. Previous studies of NAGNAG largely focused on messenger RNA. To the best of our knowledge, this is the first study testing the hypothesis that NAGNAG alternative splicing is also operative in large intergenic noncoding RNA (lincRNA). The RNA-seq data sets from recent deep sequencing studies were queried to test our hypothesis. NAGNAG alternative splicing of human lincRNA was identified while querying two independent RNA-seq data sets. Within these datasets, 31 NAGNAG alternative splicing sites were identified in lincRNA. Notably, most exons of lincRNA containing NAGNAG acceptors were longer than those from protein-coding genes. Furthermore, presence of CAG coding appeared to participate in the splice site selection. Finally, expression of the isoforms of NAGNAG lincRNA exhibited tissue specificity. Together, this study improves our understanding of the NAGNAG alternative splicing in lincRNA.

1. Introduction

The NAGNAG alternative splicing mechanism is a process which facilitates alternative protein expression from a single gene. Analysis of deep RNA-sequencing data by Bradley et al. (2012) confirmed that NAGNAG is highly regulated [1]. NAGNAG alternative splicing specifically targets inclusion or exclusion of three nucleotides at $3'$ splice sites (Figure 1), thus effecting a change in one or two amino acids encoded in the final protein [2–9]. Such amino acid substitutions have been shown to affect protein function and interfere with signaling [10], affect cellular localization [11], and impact on DNA and protein binding [12–14] in both plants and mammals. A role for NAGNAG alternative splicing was shown in human Stargardt disease [15] and has been implicated in other disease processes including cancer [16].

Large intergenic noncoding RNAs (lincRNAs) have traditionally been defined as long noncoding transcripts greater than 200 nucleotides in length. Overlapping isoforms of lincRNA have been reported previously and may include protein-coding genes [17]. Recently, while exploring the dynamic profiles of NAGNAG acceptors in Arabidopsis, we identified two isoforms originating from the same NAGNAG acceptors but located in noncoding RNA [18]. To date, previous studies have assumed NAGNAG acceptors function through the classical mRNA paradigm based on observation of altered coding for one or two amino acids in the protein-coding gene. Based on this observation of NAGNAG acceptors in Arabidopsis, we proposed an expanded paradigm and hypothesized that NAGNAG alternative splicing mechanism also exists in lincRNA.

Bioinformatics has become a powerful tool for the study of alternative splicing and its functional consequence. To date, bioinformatic analyses have produced evidence of alternative splicing in approximately 80% of human genes [19]. Bioinformatic approaches have been invaluable for exploring comparative genomics across species and such studies have produced important insights into regulatory mechanisms

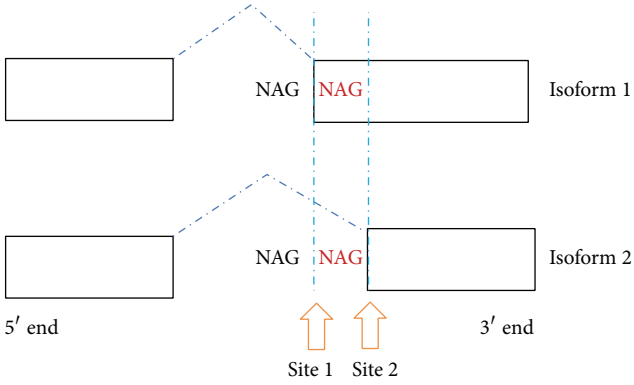


Figure 1: NAGNAG alternative splicing can result in two isoforms. The NAGNAG acceptors at the 3′-end can be either at site 1 or site 2, are three nucleotides apart, and exhibit the "NAGNAG" motif signature.

governing splicing and its role in evolution and adaptation. Single base-pair resolution offered by deep RNA sequencing motivated us to find further direct evidence of NAGNAG alternative splicing in lincRNA. To accomplish this goal we applied computational approaches to two public datasets of deeply-sequenced human tissue genomic data whose content included previously annotated lincRNA. By aligning the two RNA-seq data sets and systematically screening, identifying, and quantifying the NAGNAG alternative splicing of lincRNA, 31 NAGNAG alternative splicing events in lincRNA were defined. Importantly, tissue-specific patterns of expression for NAGNAG isoforms in lincRNA were observed.

2. Methods

2.1. Data. RNA-seq data sets were downloaded from NCBI SRA (accession number for data sets 1 and 2: E-MTAB-513 and GSE30554). These RNA-seq data were generated by sequencing 8 individual human tissues and mixture of 16 tissues (Illumina Body Map) using the Illumina HiSeq 2000 (Illumina, Inc.) platform. Each sample was deeply sequenced with more than 200 million reads and annotated for lincRNA. We only kept the high-quality reads using FastX quality filter with the following criteria: minimum of 20 Phred score over at least 80% of the sequence read.

2.2. Alignment, Screening, and Quantification. Annotations of human lincRNA were obtained from Human lincRNA Catalog hosted at Broad Institute [20]. All RNA-seq datasets were aligned to lincRNA with tophat [21] using the "-max-multihits 1", which only permits unique mapping. The anchor length of the software was set at 8 nt and the mismatch number in these regions at 0 nt to avoid alignment bias. After the data were aligned, sequence postprocessing tool (SAMtools) was used to store, sort, and index the binary SAM data (bam files) with respect to sequence alignment (http://samtools.sourceforge.net) [22].

To identify lincRNA containing NAGNAG alternative splicing sites, we screened the lincRNA sequences using the classical expression of the "NAGNAG" motif. Alignment of RNA-seq reads to the NAGNAG splicing junctions was used to confirm and validate the existence of the splice sites. We required at least four junction reads with the same 5′ splice sites, stipulating that two needed to match the first NAGNAG splice site (site 1) while the other two were required to match the second NAGNAG splice site (site 2) [23, 24].

The sequences for splice sites and the 30 bp exonic and intronic flanking sequences were extracted based on hg19 genome sequence with Bioconductor package Biostrings (R package version 2.22.0). Sequence logos were drawn by WebLogo with default parameters as described previously [25]. Two flanking sequences of the NAGNAG acceptors, including 30 bp from intron and 30 bp from exon, were extracted and screened for the potential patterns. The ratio of isoform expression at two alternative splice sites (site 1 and site 2) was calculated as log(read counts at side 1âĄĎread counts at side 2). NAGNAG acceptors were grouped into four categories based on this ratio and the strand information. If the expression of isoform 1 was more than that of isoform 2, ratio > 0; otherwise, ratio < 0.

To quantify RNA expression levels, all RNA-seq counts were normalized using reads per million (RPM). The expression level of NAGNAG isoforms in lincRNA was calculated by read counts through Bioconductor package Rsamtools (R package version 1.6.3) and IRanges (R package version 1.12.6). Duplicate reads were kept for quantification purpose. NAGNAG motifs were only designated as NAGNAG acceptors if two splice sites exhibited more than 2 reads in at least two samples. To avoid ambiguity, we discarded those NAGNAG acceptors located in the overlapping area between lincRNAs and annotated genes.

2.3. Quantification of Tissue-Specific NAGNAG Acceptors. To analyze the relationship between the ratio of two NAGNAG splice sites and the tissues, we used Bioconductor package limma through the linear model:

$$Y_{ijk} = \alpha_i + \beta_j + \varepsilon_{ijk}, \quad (1)$$

where Y represents the log ratio of two NAGNAG splice sites from the same NAGNAG acceptor, with NAGNAG acceptor i, tissue j, and sample k; α represents the main effect of ith NAGNAG acceptor; β represents the main effect of jth tissue; ε represents the measurement error. The NAGNAG acceptors were selected using false discovery rate (FDR)-adjusted P values < 0.05.

3. Results

Two novel observations were documented. First, mapping of unique reads to the potential NAGNAG alternative splicing sites in human lincRNA demonstrated existence of NAGNAG alternative splicing in lincRNA (Table 1). Of the 1320 lincRNAs containing the NAGNAG motif, presence of NAGNAG acceptors was confirmed with RNA-seq data in 30 lincRNAs. These 31 NAGNAG acceptors originate from 30 transcripts. Interestingly, linc-POLR3G-10 exhibited two NAGNAG acceptors located in two distinct transcripts: TCONS_00010012 and TCONS_00010010. Presence of two NAGNAG acceptors was identified in the upstream region

Table 1: NAGNAG acceptors in lincRNA confirmed by RNA-seq.

Transcript ID	linc name	chr	Site 1	Site 1 existence	Site 2	Site 2 existence	Strand	Neighbouring gene
TCONS_00000929	linc-CMPK1-3	chr1	47645537	Data 1, 2	47645540	Data 1, 2	+	CMPK1
TCONS_00001552	linc-CTBS-1	chr1	85084564	Data 1	85084567	Data 1	−	CTBS
TCONS_00002502	linc-CRP-1	chr1	159746542	Data 1	159746545	Data 1	−	CRP
TCONS_00002232	linc-IARS2-3	chr1	219414541	Data 1, 2	219414544	Data 1	+	IARS2
TCONS_00018502	linc-GDF10-1	chr10	48515547	Data 1	48515550	Data 1	−	GDF10
TCONS_00021357	linc-BEST3-1	chr12	70124473	Data 1	70124476	Data 2	−	BEST3
TCONS_00020623	linc-TMEM132C-14	chr12	126580786	Data 1, 2	126580789	Data 1	+	TMEM132C
TCONS_00023051	linc-DIO3-8	chr14	101363930	Data 2	101363933	Data 2	+	DIO3
TCONS_00023721	linc-ANP32A-1	chr15	69753046	Data 2	69753049	Data 1, 2	−	ANP32A
TCONS_00023791	linc-FAM174B-1	chr15	93325371	Data 1, 2	93325374	Data 1	−	FAM174B
TCONS_00023799	linc-RGMA-7	chr15	95753867	Data 1	95753870	Data 1	−	RGMA
TCONS_00024399	linc-CHD9-6	chr16	51806287	Data 1	51806290	Data 1	+	CHD9
TCONS_00025631	linc-NR1D1-1	chr17	38277663	Data 1	38277666	Data 1, 2	−	NR1D1
TCONS_00025146	linc-VEZF1-1	chr17	56066627	Data 1, 2	56066630	Data 1, 2	−	VEZF1
TCONS_00026560	linc-NETO1-1	chr18	71351479	Data 1, 2	71351482	Data 1, 2	−	NETO1
TCONS_00027051	linc-ZNF227-1	chr19	44700207	Data 1	44700210	Data 1	+	ZNF227
TCONS_00004960	linc-ITGA4-2	chr2	181940923	Data 1	181940926	Data 1	+	ITGA4
TCONS_00003507	linc-GPR55-1	chr2	231856751	Data 1, 2	231856754	Data 2	−	GPR55
TCONS_00029585	linc-RASD2-1	chr22	35850418	Data 1	35850421	Data 1	+	RASD2
TCONS_00005471	linc-TMEM14E-2	chr3	153103176	Data 1	153103179	Data 1	−	TMEM14E
TCONS_00007527	linc-SPATA18-1	chr4	52912845	Data 1, 2	52912848	Data 1, 2	+	SPATA18
TCONS_00009387	linc-OSMR-1	chr5	38792356	Data 1	38792359	Data 1	+	OSMR
TCONS_00010010	linc-POLR3G-10	chr5	87581558	Data 1, 2	87581561	Data 1, 2	+	POLR3G
TCONS_00010012	linc-POLR3G-10	chr5	87583253	Data 1, 2	87583256	Data 1	+	POLR3G
TCONS_00009724	linc-LYSMD3-2	chr5	90610061	Data 1, 2	90610064	Data 1, 2	−	LYSMD3
TCONS_00010581	linc-MGAT1-2	chr5	180257403	Data 1	180257406	Data 1	−	MGAT1
TCONS_00012396	linc-PSMG4-1	chr6	3257557	Data 1	3257560	Data 1	+	PSMG4
TCONS_00011322	linc-FAM135A-1	chr6	71104930	Data 1	71104933	Data 1	+	FAM135A
TCONS_00012862	linc-FAM20C-2	chr7	153409	Data 1	153412	Data 1	+	FAM20C
TCONS_00014103	linc-SEPT7-1	chr7	35756638	Data 1	35756641	Data 1, 2	+	SEPT7
TCONS_00014833	linc-UTP23-3	chr8	112757671	Data 1, 2	112757674	Data 2	+	UTP23

of the fourth and fifth exons of this 5-exon gene. In addition, 8 NAGNAG acceptors were identified within the overlapping regions between lincRNA and protein-coding RNA but were not further considered in this study (see Supplementary Data 1 in Supplementary Material available online at http://dx.doi.org/10.1155/2014/736798).

Most exons in lincRNA containing NAGNAG acceptors exceeded protein-coding genes in length (Wilcoxon rank sum test, P value $< 2.2e - 16$). The average exon length of protein-coding genes ranged between 306 ± 702 bp and the average neighbouring intron length ranged between 6092 ± 19983 bp (Supplementary Figure S1), compared to the average exon and intron length of lincRNA which ranged between 349 ± 630 bp and 8476 ± 19751 bp, respectively. Most tandem acceptors of lincRNA occurred at the furthest exon, that is, second exon occurring in the lincRNA (mean: 2.52; sd: 0.71) whereas those found in protein-coding genes were found centrally located among all of the exons occurring in the gene (mean: 10.7; sd: 8.8). The most prevalent triplet found among the lincRNA sequences was CAG for both splice sites, with GAG present at lowest frequency (Supplementary Table S1). CAGCAG and CAGAAG combinations occurred at highest frequency. Positive correlation with the expression level was found when CAG was encoded relative to splice site selection. Specifically, a predilection for the first splice site was noted when CAG was encoded at the first NAG site (ratio > 0, Figure 2). Alternatively, when CAG was located at the second NAG position or was absent from the splice site altogether, the second NAG was favoured for splicing (ratio < 0, Figure 2).

The second novel observation was demonstration of tissue-specific properties by 6 NAGNAG acceptors in lincRNA (FDR adjusted P value < 0.05). Figure 3 shows that 6 of 31 NAGNAG acceptors exhibited statistically significant differences in expression levels across diverse tissues. Specifically, as seen in Figure 3, the first NAG splice site is specifically targeted by the NAGNAG acceptor: chr5:87583253-87583256_+ from TCONS_00010012. Presence of these splice sites was associated with a clear expression pattern in several tissues including lymph node, lung, and kidney, and this

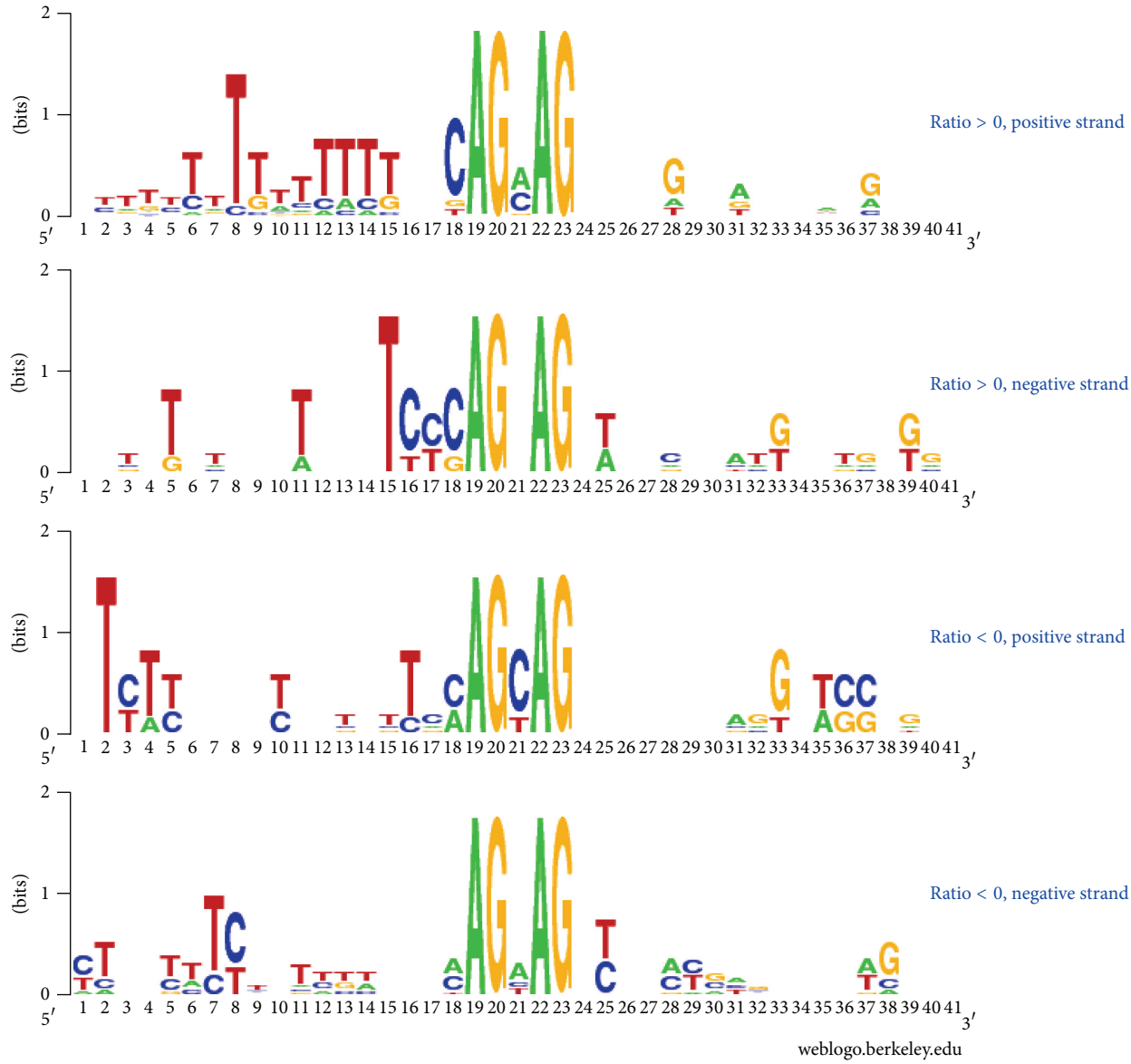


Figure 2: Sequence logos for 30 bp flanking sequences for 3′ splice sites. The logos are divided into four groups based on the chromosome strand and ratio of read counts of site 1 to site 2.

signature was remarkably consistent. Moreover, a similar pattern for the alternative splice sites was noted and the second NAG splice site was specifically targeted by NAGNAG acceptors: chr15:95753867-95753870_-. This distinctive expression pattern was clearly evident in ovary. Twenty-five of NAGNAG acceptors were notably absent or exhibited no difference in expression pattern across most tissues.

4. Discussion

Splice sites are pivotal factors in the splicing process [26]. NAGNAG alternative splicing was identified in the past decade and is characterized by inclusion or exclusion of three nucleotides at 3′ splice sites, resulting in substitutions in one or two amino acids in the protein products. Previous studies have shown that this type of alternative splicing is highly regulated and related to proteome evolution [1]. Functionally, NAGNAG alternative splicing in mRNA results in various isoforms which generate alternative proteins following translation.

To the best of our knowledge, the present study provides the first evidence that NAGNAG alternative splicing can be observed not only in mRNA but also in lincRNA. Although alternative splicing of lincRNA was reported previously [20], the report of NAGNAG alternative splicing is novel. Following analysis of two RNA-seq data sets including annotations for lincRNA, we identified 31 NAGNAG acceptors in lincRNA. These 31 NAGNAG acceptors originated from 30 transcripts. Interestingly, a role for "CAG" sequence was suggested in splice site selection with CAG being the most prevalent triplet found among the lincRNA sequences for both splice sites. GAG was present at lowest frequency and

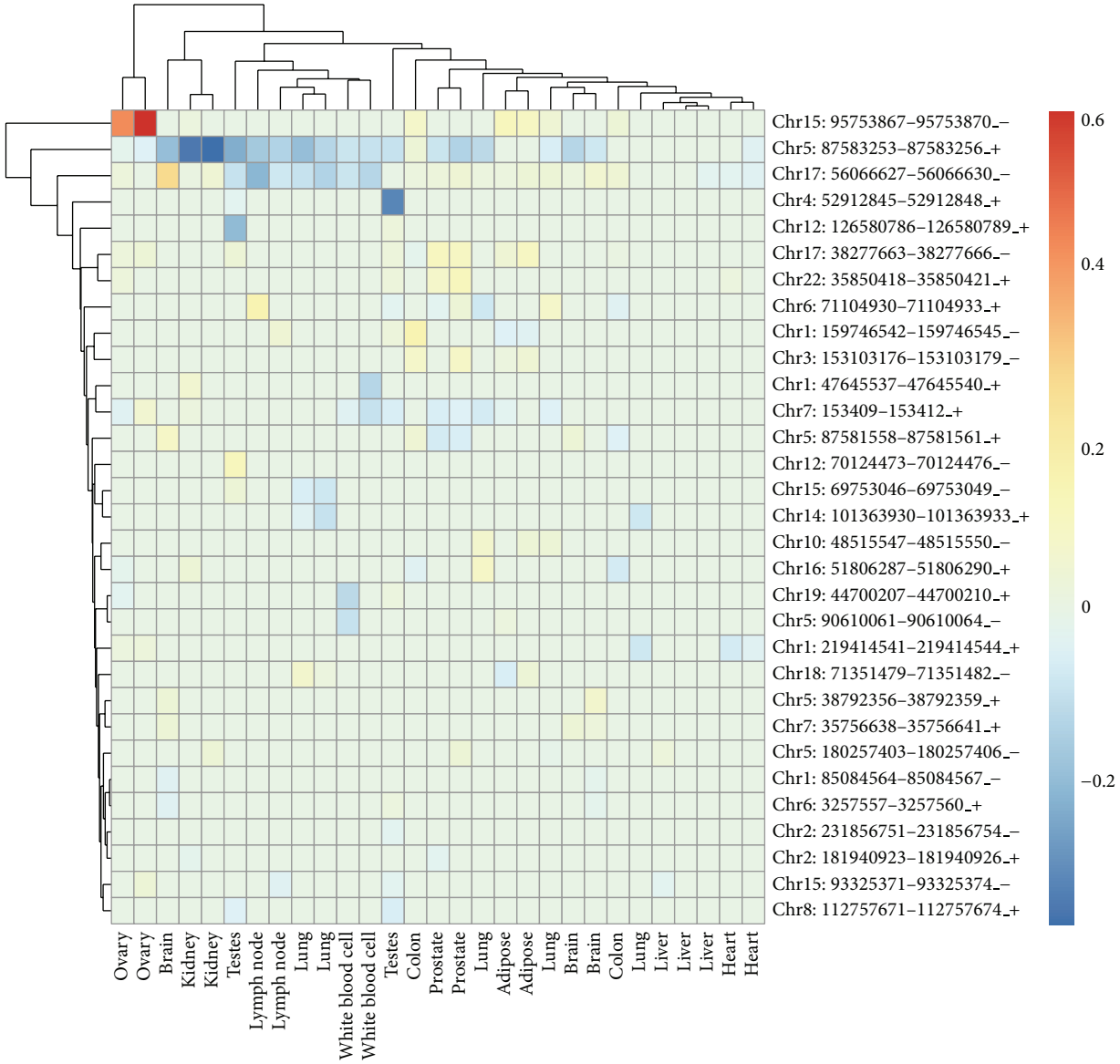


Figure 3: Heat map for the ratio of the NAGNAG isoforms at the two alternative splice sites (site 1 and site 2). Row represents 31 NAGNAG acceptors while column represents various tissues. Ratio > 0: site 2 is preferred. Ratio < 0: site 1 is preferred.

CAGCAG and CAGAAG combinations occurred at highest frequency. A predilection for the first splice site was noted when CAG was encoded at the first NAG site. The second NAG was favoured for splicing when CAG was located at the second NAG position or was absent altogether. This finding is consistent with the previous reports about mRNA [27].

Traditionally, lincRNA has been defined as stretches of DNA transcripts exceeding 200 base pairs in length which do not encode putative functional protein products [28]. lincRNA has been posited to play a role in splicing processes [29] and has been reported to contain predominately two exons [30]. In the current study, most exons from lincRNA containing NAGNAG acceptors exceeded protein-coding genes in length. Most tandem acceptors of lincRNA identified in the present study occurred at the furthest exon, that is, the second exon occurring in the lincRNA. By contrast those found in protein-coding genes have generally been found centrally located among all of the exons occurring in the gene.

The mechanism of this NAGNAG alternative splicing is not completely understood. Hiller and colleagues [3] suggested that these NAGNAG acceptors are not random noise because some fraction of NAGNAG acceptors is tissue-specific, although this theory was not universally shared by others [6, 8]. However, Bradley et al. provided solid evidence in support of tissue specificity based on RNA-seq analysis of 16 human and 8 mouse tissues wherein they demonstrated that at least 25% of NAGNAG acceptors in mRNA were regulated in a tissue-specific manner [1]. This percentage exceeded earlier estimates for tissue specificity [27]. Analysis of our selected datasets revealed low levels of consistent tissue-specific patterns relative to NAGNAG acceptors in lincRNA. Among 19% of NAGNAG acceptors that exhibited

distinct differences in expression levels of certain tissues, targeting of specific splicing pattern among two NAGNAG acceptors was noted.

There are some limitations of this computational study. First, use of annotation data was limited to the Human lincRNA Catalog at Broad Institute [20], although other annotations of human lincRNA are also available [30]. More information about lincRNA will help to identify more NAGNAG alternative splicing. Second, biological significance and potential disease impact of NAGNAG alternative splicing was only projected computationally, and awaits confirmation through further proteomic studies. For example, results of gene ontology analysis by application for genes targeted by NAGNAG acceptors in lincRNA indicated that these genes were all functionally engaged in transcription regulation (ANP32A, CHD9, NR1D1, POLR3G, VEZF1, ZNF227) and signalling (CRP, CTBS, FAM174B, FAM20C, GDF10, ITGA4, NETO1, RGMA, TMEM132C, OSMR). Further, analysis for potential disease association of the neighbouring genes revealed that these genes represented candidate genes associated with risk for many important diseases, including hypertension, obesity, and cancer, among others (see Supplementary Table S2 for a complete list).

Importantly, bioinformatics analysis has proved to be an invaluable tool in the investigation of the role of alternative splicing from numerous perspectives including microarray analysis, alternative splicing prediction utilizing comparative genomic approaches, identification and depiction of isoform and splicing patterns, definition of regulation of alternative splicing, delineation of functional impact, and its role in defining evolutionary and adaptive processes, among other investigations [19]. To delineate alternative splicing in lincRNA, further investigations are essential in unraveling their functional and regulatory roles through application of bioinformatic, genetic, and proteomic approaches. The evolutionary aspect of lincRNA NAGNAG alternative splicing across different species can also be studied in the future.

Conflict of Interests

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

Authors' Contribution

Xiaoyong Sun and Simon M. Lin designed the project, analyzed the data, and drafted the paper. Xiaoyan Yan participated in data analysis and performed the gene ontology analysis. All authors read, wrote, and approved the paper.

Acknowledgments

The authors would like to thank Dr. Zhaoyuan Hou for helpful discussion, Dr. Ingrid Glurich for editing the paper, and Dr. Steven Schrodi for reviewing the paper. They also thank National Supercomputer Center in Jinan for technical support. The project described was supported by Start-up Grant from Shandong Agricultural University to Xiaoyong Sun and partially supported by the Clinical and Translational Science Award (CTSA) program, through the NIH National Center for Advancing Translational Sciences (NCATS), Grant UL1TR000427. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

[1] R. K. Bradley, J. Merkin, N. J. Lambert, and C. B. Burge, "Alternative splicing of RNA triplets is often regulated and accelerates proteome evolution," *PLoS Biology*, vol. 10, no. 1, Article ID e1001229, 2012.

[2] L. Li and G. A. Howe, "Alternative splicing of prosystemin pre-mRNA produces two isoforms that are active as signals in the wound response pathway," *Plant Molecular Biology*, vol. 46, no. 4, pp. 409–419, 2001.

[3] M. Hiller, K. Huse, K. Szafranski et al., "Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity," *Nature Genetics*, vol. 36, no. 12, pp. 1255–1257, 2004.

[4] C. W. Sugnet, W. J. Kent, M. Ares Jr., and D. Haussler, "Transcriptome and genome conservation of alternative splicing events in humans and mice," *Pacific Symposium on Biocomputing*, pp. 66–77, 2004.

[5] K. W. Tsai and W. C. Lin, "Quantitative analysis of wobble splicing indicates that it is not tissue specific," *Genomics*, vol. 88, no. 6, pp. 855–864, 2006.

[6] T. Chern, E. van Nimwegen, C. Kai et al., "A simple physical model predicts small exon length variations," *PLoS Genetics*, vol. 2, no. 4, article e45, 2006.

[7] K. Iida, M. Shionyu, and Y. Suso, "Alternative splicing at NAGNAG acceptor sites shares common properties in land plants and mammals," *Molecular Biology and Evolution*, vol. 25, no. 4, pp. 709–718, 2008.

[8] R. Sinha, S. Nikolajewa, K. Szafranski et al., "Accurate prediction of NAGNAG alternative splicing," *Nucleic Acids Research*, vol. 37, no. 11, pp. 3569–3579, 2009.

[9] R. Sinha, A. D. Zimmer, K. Bolte et al., "Identification and characterization of NAGNAG alternative splicing in the moss *Physcomitrella patens*," *BMC Plant Biology*, vol. 10, article 76, 2010.

[10] G. Condorelli, R. Bueno, and R. J. Smith, "Two alternatively spliced forms of the human insulin-like growth factor I receptor have distinct biological activities and internalization kinetics," *The Journal of Biological Chemistry*, vol. 269, no. 11, pp. 8510–8516, 1994.

[11] K. Tadokoro, M. Yamazaki-Inoue, M. Tachibana et al., "Frequent occurrence of protein isoforms with or without a single amino acid residue by subtle alternative splicing: the case of Gln in DRPLA affects subcellular localization of the products," *Journal of Human Genetics*, vol. 50, no. 8, pp. 382–394, 2005.

[12] K. J. Vogan, D. A. Underhill, and P. Gros, "An alternative splicing event in the Pax-3 paired domain identifies the linker region as a key determinant of paired domain DNA-binding activity," *Molecular and Cellular Biology*, vol. 16, no. 12, pp. 6677–6686, 1996.

[13] Z. J. Lorković, R. Lehner, C. Forstner, and A. Barta, "Evolutionary conservation of minor U12-type spliceosome between plants and humans," *RNA*, vol. 11, no. 7, pp. 1095–1107, 2005.

[14] M. Hiller, K. Szafranski, K. Huse, R. Backofen, and M. Platzer, "Selection against tandem splice sites affecting structured protein regions," *BMC Evolutionary Biology*, vol. 8, no. 1, article 89, 2008.

[15] A. Maugeri, M. A. van Driel, D. J. R. van de Pol et al., "The 2588G → C mutation in the ABCR gene is a mild frequent founder mutation in the western European population and allows the classification of ABCR mutations in patients with Stargardt disease," *The American Journal of Human Genetics*, vol. 64, no. 4, pp. 1024–1035, 1999.

[16] L. Hui, X. Zhang, X. Wu et al., "Identification of alternatively spliced mRNA variants related to cancers by genome-wide ESTs alignment," *Oncogene*, vol. 23, no. 17, pp. 3013–3023, 2004.

[17] P. Kapranov, A. T. Willingham, and T. R. Gingeras, "Genome-wide transcription and the implications for genomic organization," *Nature Reviews Genetics*, vol. 8, no. 6, pp. 413–423, 2007.

[18] Y. Shi, G. Sha, and X. Sun, "Genome-wide study of NAGNAG alternative splicing in *Arabidopsis*," *Planta*, vol. 239, no. 1, pp. 127–138, 2014.

[19] C. Lee and Q. Wang, "Bioinformatics analysis of alternative splicing," *Briefings in Bioinformatics*, vol. 6, no. 1, pp. 23–33, 2005.

[20] M. Cabili, C. Trapnell, L. Goff et al., "Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses," *Genes and Development*, vol. 25, no. 18, pp. 1915–1927, 2011.

[21] C. Trapnell, L. Pachter, and S. L. Salzberg, "TopHat: discovering splice junctions with RNA-Seq," *Bioinformatics*, vol. 25, no. 9, pp. 1105–1111, 2009.

[22] H. Li, B. Handsaker, A. Wysoker et al., "The sequence alignment/map format and SAMtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009.

[23] A. Ameur, A. Wetterbom, L. Feuk, and U. Gyllensten, "Global and unbiased detection of splice junctions from RNA-seq data," *Genome Biology*, vol. 11, no. 3, article R34, 2010.

[24] J. W. Nam and D. P. Bartel, "Long noncoding RNAs in *C. elegans*," *Genome Research*, vol. 22, no. 12, pp. 2529–2540, 2012.

[25] G. E. Crooks, G. Hon, J. M. Chandonia, and S. E. Brenner, "WebLogo: a sequence logo generator," *Genome Research*, vol. 14, no. 6, pp. 1188–1190, 2004.

[26] E. L. Huttlin, M. P. Jedrychowski, J. E. Elias et al., "A tissue-specific atlas of mouse protein phosphorylation and expression," *Cell*, vol. 143, no. 7, pp. 1174–1189, 2010.

[27] M. Akerman and Y. Mandel-Gutfreund, "Alternative splicing regulation at tandem 3′ splice sites," *Nucleic Acids Research*, vol. 34, no. 1, pp. 23–31, 2006.

[28] T. R. Mercer, M. E. Dinger, and J. S. Mattick, "Long non-coding RNAs: insights into functions," *Nature Reviews Genetics*, vol. 10, no. 3, pp. 155–159, 2009.

[29] K. L. Fox-Walsh, Y. Dou, B. J. Lam, S. Hung, P. F. Baldi, and K. J. Hertel, "The architecture of pre-mRNAs affects mechanisms of splice-site pairing," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 45, pp. 16176–16181, 2005.

[30] T. Derrien, R. Johnson, G. Bussotti et al., "The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression," *Genome Research*, vol. 22, no. 9, pp. 1775–1789, 2012.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 286419, 10 pages
http://dx.doi.org/10.1155/2014/286419

Research Article

iCTX-Type: A Sequence-Based Predictor for Identifying the Types of Conotoxins in Targeting Ion Channels

Hui Ding,[1] En-Ze Deng,[1] Lu-Feng Yuan,[1] Li Liu,[2] Hao Lin,[1,3] Wei Chen,[3,4] and Kuo-Chen Chou[3,5]

[1] *Key Laboratory for Neuro-Information of Ministry of Education, Center of Bioinformatics, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China*
[2] *Laboratory of Theoretical Biophysics, School of Physical Science and Technology, Inner Mongolia University, Hohhot 010021, China*
[3] *Gordon Life Science Institute, Boston, MA 02478, USA*
[4] *Department of Physics, School of Sciences Center for Genomics and Computational Biology, Hebei United University, Tangshan 063000, China*
[5] *Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah 21589, Saudi Arabia*

Correspondence should be addressed to Hao Lin; hlin@gordonlifescience.org and Wei Chen; greatchen@heuu.edu.cn

Received 13 March 2014; Revised 22 April 2014; Accepted 7 May 2014; Published 1 June 2014

Academic Editor: Shiwei Duan

Conotoxins are small disulfide-rich neurotoxic peptides, which can bind to ion channels with very high specificity and modulate their activities. Over the last few decades, conotoxins have been the drug candidates for treating chronic pain, epilepsy, spasticity, and cardiovascular diseases. According to their functions and targets, conotoxins are generally categorized into three types: potassium-channel type, sodium-channel type, and calcium-channel types. With the avalanche of peptide sequences generated in the postgenomic age, it is urgent and challenging to develop an automated method for rapidly and accurately identifying the types of conotoxins based on their sequence information alone. To address this challenge, a new predictor, called iCTX-Type, was developed by incorporating the dipeptide occurrence frequencies of a conotoxin sequence into a 400-D (dimensional) general pseudoamino acid composition, followed by the feature optimization procedure to reduce the sample representation from 400-D to 50-D vector. The overall success rate achieved by iCTX-Type via a rigorous cross-validation was over 91%, outperforming its counterpart (RBF network). Besides, iCTX-Type is so far the only predictor in this area with its web-server available, and hence is particularly useful for most experimental scientists to get their desired results without the need to follow the complicated mathematics involved.

1. Introduction

Being peptides consisting of about 10 to 30 amino acid residues, conotoxins are toxins secreted by cone snails for capturing prey and securing themselves. This kind of toxins can bind to various targets, such as G protein-coupled receptors (GPCRs), nicotinic acetylcholine, and neurotensin receptors. In particular, they display extremely high specificity and affinity for ion channels. Ion channels represent a class of membrane spanning protein pores that mediate the flux of ions in a variety of cell types. There are over 300 types of ion channels in a living cell [1]. Many crucial functions in life, such as heartbeat, sensory transduction, and central nervous system response, are controlled by cell signaling via various ion channels. Ion channel dysfunction may lead to a number of diseases, such as epilepsy, arrhythmia, and type II diabetes. These kinds of diseases are primarily treated with the drugs that modulate the ion channels concerned. Ion channels are also the important targets for treating virus diseases (see, e.g., [2–4]). Owing to their importance to human being's life, ion channels have become the 2nd most frequent targets for drug development, just next to GPCRs (G protein-coupled receptors) [5]. The following three kinds of ion channels are usually the targets by conotoxins: potassium (K) channel (Figure 1), sodium (Na) channel (Figure 2), and calcium (Ca) channel (Figure 3). Based on their functions

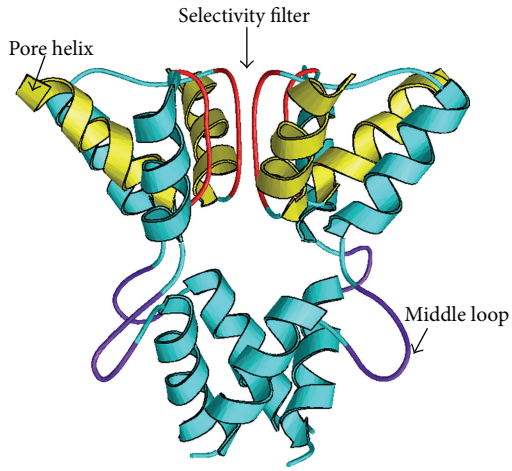


Figure 1: A ribbon drawing to show the human potassium (K) channel. Reproduced from Chou [6] with permission.

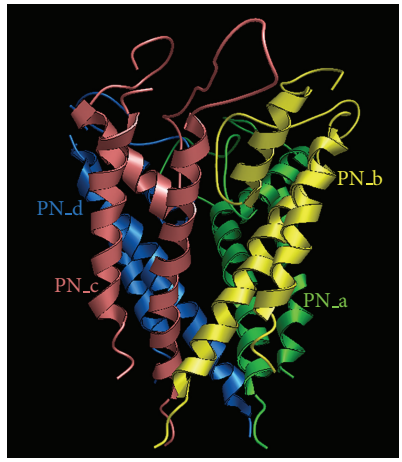


Figure 2: A ribbon drawing to show the human sodium (Na) channel. Reproduced from Chou [6] with permission.

and targeting objects, conotoxins can be classified into the following three types: (i) K-channel-targeting type; (ii) Na-channel-targeting type; and (iii) Ca-channel-targeting type.

Although conotoxins are lethally venomous because of blocking the transmission of nerve impulses, they have been widely used to treat chronic pain, epilepsy, spasticity, and cardiovascular diseases. Therefore, conotoxins have been regarded as important pharmacological tools for neuroscience research.

It has been estimated that there are more than 100,000 kinds of conotoxins secreted by over 700 kinds of *Conus* in the world [8]. However, relatively much fewer conotoxins (about 3,000 peptides) have been experimentally confirmed and reported in literature and databases. Moreover, the records about the functions of conotoxins in public databases are no more than 300 items. Hence, developing a computational method to predict the functions of conotoxins has become a challenging task.

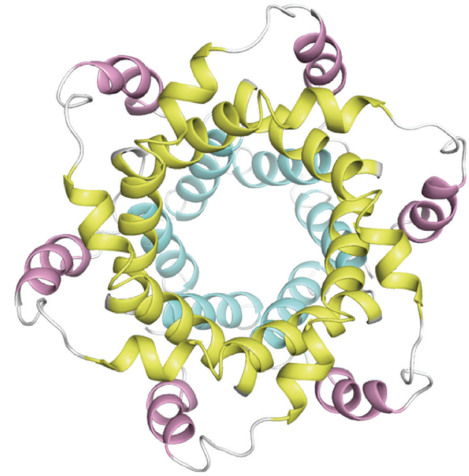


Figure 3: A ribbon drawing to show the calcium (Ca) channel from hepatitis C virus. Reproduced from [4] with permission.

In a pioneer work, Mondal et al. [9] proposed a method for predicting conotoxin superfamilies by using the pseudoamino acid composition approach [10, 11]. Subsequently, a series of studies have been reported in predicting conotoxin superfamilies (see, for example, [12–15]). All these methods yielded quite encouraging results, and each of them did play a role in stimulating the development of this area. However, none of these methods can be used to predict the types of conotoxins defined according to their targeting ion-channels. For instance, both delta-conotoxin-like Ac6.1 (UniProt accession number: P0C8V5) [16] and omega-conotoxin-like Ai6.2 [17] (UniProt accession number: P0CB10) belong to the conotoxin O1 superfamily. However, the former targets the voltage-gated sodium channels, while the latter targets the voltage-gated calcium channels.

To deal with this problem, recently, a method was developed [7] to identify conotoxins among the aforementioned three types by using their sequence information alone. However, further work is needed in this regard due to the following reasons. (i) The prediction quality can be further improved. (ii) No web server for the prediction method in [7] was provided, and hence its usage is quite limited, especially for the majority of experimental scientists.

The present study was devoted to develop a new predictor for identifying the conotoxins' types from the above two aspects.

As elaborated in a comprehensive review [18] and conducted by a series of recent publications [19–28], to establish a really useful statistical predictor for a biological system, we need to consider the following procedures: (i) construct or select a valid benchmark dataset to train and test the predictor; (ii) formulate the biological samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (iii) introduce or develop a powerful algorithm (or engine) to operate the prediction; (iv) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; (v) establish a user-friendly web server for

the predictor that is accessible to the public. In what follows, let us describe how to deal with these procedures one by one.

2. Materials and Methods

2.1. Benchmark Dataset. The sequences of conotoxins and their functions were collected from the UniProt [29]. To ensure its quality, the benchmark dataset was constructed strictly according to the following criteria. (i) Included were only those peptides annotated with "conotoxin" and with the keyword of potassium, calcium, or sodium in their functional ontologies. (ii) Included were only those conotoxins with clear functional annotations based on experiment results. In other words, we excluded those annotated with "uncertain," "predicted," or "inferred from homology" because of lacking confidence. (iii) Excluded were those that were annotated with "immature" due to the incompleteness. (iv) Excluded were also those that contained any invalid amino acid codes, such as "B," "X," and "Z". After going through the above procedures, we obtained 195 conotoxins, of which 37 belonged to the K-channel-targeting type, 86 to the Na-channel-targeting type, and 72 to the Ca-channel-targeting type.

As elaborated in a comprehensive review [18], a benchmark dataset containing many redundant samples with high similarity would lack statistical representativeness. A predictor, if trained and tested by a benchmark dataset with many homologous sequences, might yield misleading results with overestimated accuracy [30]. To remove the homologous sequences from the benchmark dataset, a cutoff threshold of 25% was recommended [31] to exclude those protein/peptide sequences from the benchmark datasets that had ≥25% pairwise sequence identity to any other sample in the same subset. However, in this study we did not use such a stringent criterion because the currently available data did not allow us to do so. Otherwise, the numbers of peptides for some subsets would be very few to have statistical significance. As a compromise, we set the cutoff threshold at 80% and used the CD-HIT software [32] to remove those conotoxin samples that had ≥80% sequence identity to any other in a same subset. After such a screening procedure, we obtained 112 conotoxin samples for the benchmark dataset $\mathbb{S}$, as formulated as follows:

$$\mathbb{S} = \mathbb{S}_{\mathrm{K}} \cup \mathbb{S}_{\mathrm{Na}} \cup \mathbb{S}_{\mathrm{Ca}}, \tag{1}$$

where the subset $\mathbb{S}_{\mathrm{K}}$ contains 24 conotoxin samples of K-channel-targeting type, $\mathbb{S}_{\mathrm{Na}}$ contains 43 samples of Na-channel-targeting type, and $\mathbb{S}_{\mathrm{Ca}}$ contains 45 samples of Ca-channel-targeting type, while the symbol $\cup$ represents the union in the set theory. The codes of 112 conotoxins and their sequences are given in Supporting Information S1 (see Supplementary Material available online at http://dx.doi.org/10.1155/2014/286419).

Likewise, we also constructed an independent dataset $\mathbb{S}^{\mathrm{Ind}}$ as formulated by

$$\mathbb{S}^{\mathrm{Ind}} = \mathbb{S}^{\mathrm{Ind}}_{\mathrm{K}} \cup \mathbb{S}^{\mathrm{Ind}}_{\mathrm{Na}} \cup \mathbb{S}^{\mathrm{Ind}}_{\mathrm{Ca}}, \tag{2}$$

where $\mathbb{S}^{\mathrm{Ind}}_{\mathrm{K}}$ contains 12 K-conotoxins, $\mathbb{S}^{\mathrm{Ind}}_{\mathrm{Na}}$ contains 37 Na-conotoxins, and $\mathbb{S}^{\mathrm{Ind}}_{\mathrm{Ca}}$ contains 21 Ca-conotoxins. None of the samples in the independent dataset occurs in the dataset $\mathbb{S}$ of (1), and their detailed sequences are given in Supporting Information S2.

For simplicity, hereafter, let us use "K-conotoxin," "Na-conotoxin," and "Ca-conotoxin" to represent K-channel-targeting type conotoxin, Na-channel-targeting type conotoxin, and Ca-channel-targeting type conotoxin, respectively.

2.2. The Dipeptide Mode of Pseudoamino Acid Composition. Given a conotoxin peptide **P** with L amino acids, how do we translate it into a mathematical expression for statistical prediction? This is one of the first important problems to develop a sequence-based predictor for identifying the type of a conotoxin. The most straightforward way to formulate the sample of a conotoxin peptide **P** with L residues is to use its entire amino acid sequence, as can be formulated by

$$\mathbf{P} = \mathrm{R}_1\mathrm{R}_2\mathrm{R}_3\mathrm{R}_4 \cdots \mathrm{R}_{\mathrm{L}}, \tag{3}$$

where R_1 represents the 1st residue of the conotoxin peptide and R_2 the 2nd residue of the peptide and so forth. Subsequently, we can utilize various sequence similarity search based tools, such as BLAST [33], to perform statistical prediction. Although this kind of sequence model was very straightforward and intuitive, unfortunately, it failed to work when a query conotoxin peptide did not have significant similarity to any of the peptide sequences in the training dataset. Thus, investigators turned to use vectors to represent the peptide samples. Another reason for them to do so is that the statistical samples in vector format are much easier to be handled than in sequence format by many existing operation engines, such as the correlation angle approach [34], covariance discriminant (CD) [27, 35–37], neural network [38–40], optimization approach [41], support vector machine (SVM) [22, 23, 42, 43], random forest [44, 45], conditional random field [20], nearest neighbor (NN) [46, 47]; K-nearest neighbor (KNN) [30], OET-KNN [48–50], fuzzy K-nearest neighbor [25, 51–55], ML-KNN algorithm [56], and SLLE algorithm [36].

The simplest vector used to represent a peptide or protein sample is its amino acid composition (AAC), as given as follows:

$$\mathbf{P} = [f_1 \;\; f_2 \;\; \cdots \;\; f_{20}]^{\mathbf{T}}, \tag{4}$$

where f_i ($i = 1, 2, \ldots, 20$) is the normalized occurrence frequency of the ith type of native amino acid in the peptide chain and **T** is the transpose operator. The AAC model was used by many in predicting various contributes of proteins (see, e.g., [41, 57–59]). However, as we can see from (4), when using AAC to represent a peptide or protein sample, all its sequence order information would be completely lost and hence limit the prediction quality.

How can we formulate a peptide or protein sequence with a vector yet still keep considerable sequence order information? As reported in many recent publications, in order to incorporate the sequence order information, the pseudoamino acid composition [10, 11] or Chou's PseAAC [60] was proposed. Since the concept of PseAAC was proposed in 2001 [10], it has been penetrating into almost all

the fields of protein attribute predictions (see, e.g., [61–78]). Recently, the concept of PseAAC was further extended to represent the feature vectors of DNA and nucleotides [19, 21, 23, 27, 79], as well as other biological samples (see, e.g., [80–82]). Because it has been widely and increasingly used, in addition to the web server "PseAAC" [83] built in 2008, recently three types of powerful open access software, called "PseAAC-Builder" [84], "propy" [85], and "PseAAC-General" [86], were established: the former two are for generating various modes of Chou's special PseAAC, while the 3rd one is for those of Chou's general PseAAC.

According to a comprehensive review [18], the general PseAAC is formulated by

$$\mathbf{P} = \left[\psi_1 \;\; \psi_2 \;\; \cdots \;\; \psi_u \;\; \cdots \;\; \psi_\Omega\right]^{\mathbf{T}}, \tag{5}$$

where the component ψ_u ($u = 1, 2, \ldots, \Omega$) and the dimension Ω will depend on how to extract the features from the peptide sequences concerned. For the current study, since the conotoxin sequences are not long (about 10–30 residues), we could just consider the sequence order information between two most contiguous amino acid residues. Thus, the dimension of the vector $\mathbf{P}$ in (5) is $\Omega = 20 \times 20 = 400$ and each of the components therein is given by

$$\psi_u = \begin{cases} f(\mathrm{AA}) & \text{when } u = 1 \\ f(\mathrm{AC}) & \text{when } u = 2 \\ \quad\vdots & \quad\vdots \\ f(\mathrm{AY}) & \text{when } u = 20 \\ f(\mathrm{CA}) & \text{when } u = 21 \\ \quad\vdots & \quad\vdots \\ f(\mathrm{YW}) & \text{when } u = 399 \\ f(\mathrm{YY}) & \text{when } u = 400, \end{cases} \tag{6}$$

where A, C, ..., W, Y are, respectively, the single letter codes of 20 native amino acids, $f(\mathrm{AA})$ is the occurrence frequency for the dipeptide AA in the conotoxin sequence (see (3)), and $f(\mathrm{AC})$ is for the dipeptide AC and so forth. The formulation defined by (5)-(6) is actually the dipeptide mode of PseAAC, which can be automatically generated by the PseAAC server [83] for a given peptide or protein sequence.

2.3. Feature Selection. The original raw features usually contain the redundant information and noise that may negatively affect the prediction quality [87]. Using the feature selection techniques to optimize the feature set can not only enhance the prediction accuracy but also provide useful insights for in-depth understanding of the action mechanism of conotoxins. According to the feature selection algorithm [87], the F-score function is defined by

$$F(i) = \frac{\sum_{k=1}^{3}\left(\overline{f}_i^{\,k} - \overline{f}_i\right)^2}{\sum_{k=1}^{3}\left(1/\left(N_k - 1\right)\right)\sum_{j=1}^{N_k}\left(f_{ij}^k - \overline{f}_i^{\,k}\right)^2}, \tag{7}$$

where $\overline{f}_i^{\,k}$ is the average frequency of the ith feature in the kth dataset, $\overline{f}_i$ the average frequency of the ith feature in the all datasets concerned, f_{ij}^k is the frequencies of the ith feature of the jth sequence in the kth dataset, and N_k is the number of peptide samples in the kth dataset. The program called "fselect.py" was downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvmtools to calculate F-score defined in (7).

The larger the F-score is, the more likely it has a better discriminative capability [87]. Accordingly, we ranked the 400 dipeptides in (5) according to their F-scores. Subsequently, based on the ranked dipeptides, we performed the incremental feature selection (IFS) strategy to find an optimal subset of features that yielded the highest predictive accuracy. During the IFS procedure, the feature subset started with one feature with the highest F-score. A new feature subset was composed when one more feature with the second highest F-score was added. By adding these features sequentially from the higher to lower ranks, 400 feature sets would be obtained. The τth feature set can be formulated as

$$S_\tau = \{f_1, f_2, \ldots, f_\tau\}, \quad (1 \le \tau \le 400). \tag{8}$$

For each of the 400 feature sets, a prediction model based on the proposed predictive algorithm was constructed and examined with the jackknife cross-validation on the benchmark dataset. By doing so, we obtained an IFS curve in a 2D (dimensional) Cartesian coordinate system with index τ as the abscissa (or X-coordinate) and the overall accuracy as the ordinate (or Y-coordinate). The optimal feature set is expressed as

$$S_\Theta = \{f_1, f_2, \ldots, f_\Theta\}. \tag{9}$$

with which the IFS curve reached its peak. In other words, in the 2D coordinate system, when $X = \Theta$, the value of the overall accuracy was the maximum. Thus, we used the Θ features to build the final predictor.

2.4. Support Vector Machine (SVM). The classification algorithm used in this work was the support vector machine (SVM). The SVM has been widely used in the realm of bioinformatics (see, e.g., [19, 22, 23, 88–90]). Its basic principle is to transform the input vector into a high-dimension Hilbert space and seek a separating hyperplane with the maximal margin in this space by using the decision function:

$$f\left(\overrightarrow{X}\right) = \mathrm{sgn}\left(\sum_{i=1}^{N} y_i \alpha_i \cdot K\left(\overrightarrow{X}, \overrightarrow{X_i}\right) + b\right), \tag{10}$$

where $\overrightarrow{X_i}$ is the ith training vector, the y_i represents the type of the ith training vector, and $K(\overrightarrow{X}, \overrightarrow{X_i})$ is a kernel function which defines an inner product in a high dimensional feature space. Because of its effectiveness and speed in nonlinear classification process, the radial basis kernel function (RBF) $K(\overrightarrow{X_i}, \overrightarrow{X_j}) = \exp(-\gamma \| \overrightarrow{X_i} - \overrightarrow{X_j} \|^2)$ was used in the current work. The original SVM was designed for two-class problems. For multiclass problems, several strategies such as one-versus-rest (OVR), one-versus-one (OVO), and

DAGSVM have been applied to extend the traditional SVM. In the present study, we used the OVO strategy for multiclass prediction. The concrete SVM software (LibSVM) was downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm. A grid search method was used to optimize the regularization parameter C and kernel parameter via the jackknife cross-validation. The search spaces for C and γ are $[2^{15}, 2^{-5}]$ and $[2^{-5}, 2^{-15}]$ with steps of 2^{-1} and 2, respectively. For more details about SVM, see a monograph [91].

3. Results and Discussion

3.1. Test Method and Criteria. In statistical prediction, the independent dataset test, subsampling or K-fold crossover test and jackknife test are the three cross-validation methods often used to check a predictor for its accuracy [92]. However, among the three test methods, the jackknife test is deemed the least arbitrary that can always yield a unique result for a given benchmark dataset [18]. Accordingly, the jackknife test has been increasingly used and widely recognized by investigators to examine the quality of various predictors (see, e.g., [19, 21, 73, 75, 93–95]). Therefore, in this study we also adopted the jackknife test.

In addition to an objective test method, we also need a set of metrics to reasonably measure the test outcome. Here, let us use the criterion proposed in [96, 97] to develop a set of more intuitive and easier-to-understand metrics; that is, the correct rates Λ^{K} in predicting K-conotoxins, Λ^{Na} in predicting Na-conotoxins, and Λ^{Ca} in predicting Ca-conotoxins are defined by

$$\begin{aligned}
\Lambda^{K} &= \frac{N^{K} - N^{K}_{Na} - N^{K}_{Ca}}{N^{K}}, \quad \text{for the K-conotoxins} \\
\Lambda^{Na} &= \frac{N^{Na} - N^{Na}_{K} - N^{Na}_{Ca}}{N^{K}}, \quad \text{for the Na-conotoxins} \\
\Lambda^{Ca} &= \frac{N^{Ca} - N^{Ca}_{K} - N^{Ca}_{Na}}{N^{Ca}}, \quad \text{for the Ca-conotoxins,}
\end{aligned} \tag{11}$$

where N^{K} is the total number of the K-conotoxins investigated, while N^{K}_{Na} is the number of the K-conotoxins incorrectly predicted as the Na-conotoxins, and N^{K}_{Ca} is the number of the K-conotoxins incorrectly predicted as the Ca-conotoxins; N^{Na} is the total number of the Na-conotoxins investigated, while N^{Na}_{K} is the number of the Na-conotoxins incorrectly predicted as the K-conotoxins and N^{Na}_{Ca} is the number of the Na-conotoxins incorrectly predicted as the Ca-conotoxins; and N^{Ca} is the total number of the Ca-conotoxins investigated, while N^{Ca}_{Na} is the number of the Ca-conotoxins incorrectly predicted as the Na-conotoxins and N^{Ca}_{K} is the number of the Ca-conotoxins incorrectly predicted as the K-conotoxins. From (11), it follows that

$$\begin{aligned}
\text{OA} = \Lambda &= 1 - \frac{N^{K}_{Na} + N^{K}_{Ca} + N^{Na}_{K} + N^{Na}_{Ca} + N^{Ca}_{Na} + N^{Ca}_{K}}{N^{K} + N^{Na} + N^{Ca}} \\
\text{AA} &= \frac{\Lambda^{K} + \Lambda^{Na} + \Lambda^{Ca}}{3},
\end{aligned} \tag{12}$$

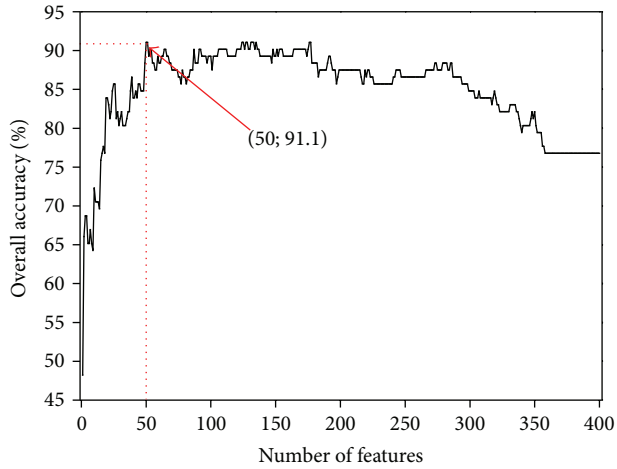


Figure 4: A plot to show the IFS curve, where the abscissa and ordinate axis denote the number of features and the overall accuracy, respectively. As shown in the figure, the value of the overall accuracy reached its peak (91.1%) when the top-ranked 50 dipeptide features were taken into account.

Table 1: List of the 50 optimal features or dipeptides derived according to (7)–(9) as elaborated in the Section 2.3.

AA	AS	CC	CH	CS	DH	DN	EN	GA	GH
GL	GT	GY	HA	HL	HS	IY	KD	KK	KM
KP	LN	LV	MC	MY	ND	NQ	NS	PI	QK
QT	RC	RD	RF	RN	RT	RW	SC	SG	TE
TF	TT	VV	WG	WI	YD	YH	YL	YT	YY

where OA stands for the overall accuracy and AA for the average accuracy.

3.2. The Optimal Features. As mentioned above, it would be no good for a sample vector to contain either too few or too many features. This is because the former would limit the prediction quality due to lack of information, while the latter would generate a lot of noise due to redundancy. Therefore, we should find a set of optimal features, for which there is minimal redundancy among themselves but maximal relevancy to the target to be predicted. In the present study, such an optimal feature-set is none but (9).

Shown in Figure 4 is the IFS curve for the value of OA against the number of the counted features, as described in Section 2.3. As can be seen from there, the value of OA reached its peak of 91.1% when the top-ranked 50 dipeptides (Table 1) were taken into account.

The predictor thus obtained via the aforementioned procedures is called "iCTX-Type," where "i" stands for "identify" and "CTX" for "conotoxin."

A comparison of the current predictor iCTX-Type with the one in [7] (i.e., to the best of our knowledge, it is the only existing predictor in this area) is given in Table 2, from which we can see the following. (i) For four of the five metrics defined in (10)-(11), iCTX-Type yielded higher scores than the method in [7]. Particularly, iCTX-Type achieved

Table 2: Comparison of the current method with the one in [7] by the jackknife test on the same benchmark dataset (Supporting Information S1) according to the metrics defined in (11)-(12).

Method	Number of features counted	Λ^{K} (%)	Λ^{Na} (%)	Λ^{Ca} (%)	AA (%)	OA (%)
RBF network[a]	70	91.7	88.4	88.9	89.7	89.3
iCTX-Type[b]	50	83.3	97.8	89.8	90.3	91.1

[a]See [7].
[b]This paper.

higher overall accuracy (OA) and average accuracy (AA). (ii) Compared with the method of [7] using 70 features, only 50 features were used in the present method (Table 1), indicating that the iCTX-Type is more efficient in excluding redundancy and noise as well as in capturing the core features.

To further verify the performance of the current predictor, iCTX-Type was also used to identify the samples in the independent dataset $\mathbb{S}^{Ind}$ (see Supporting Information S2), and the success rates (see (11)) thus obtained were 91.7%, 91.9%, and 90.5% for K-, Na-, and Ca-conotoxins, respectively. These results are fully consistent with those obtained by the jackknife test as given in Table 2, furtherindicating that the new predictor iCTX-Type is quite promising and holds a high potential to become a useful tool for in-depth studying ion channel-targeted conotoxins.

To enhance the value of its practical applications [98], a web server for the new iCTX-Type predictor was established as described below.

3.3. Web-Server Guide. For the convenience of the vast majority of experimental scientists, below a step-by-step guide is provided for how to use the web server to get the desired results without the need to follow the mathematic equations that were presented in this paper just for the integrity in developing the predictor.

Step 1. Open the web server at http://lin.uestc.edu.cn/server/iCTX-Type and you will see the top page of iCTX-Type on your computer screen, as shown in Figure 5. Click on the Read Me button to see a brief introduction about the predictor and the caveat when using it.

Step 2. Either type or copy/paste the query peptide sequences into the input box at the center of Figure 5. The input sequence should be in the FASTA format. A sequence in FASTA format consists of a single initial line beginning with a greater-than symbol ">" in the first column, followed by lines of sequence data. The words right after the ">" symbol in the single initial line are optional and only used for the purpose of identification and description. All lines should be no longer than 120 characters and usually do not exceed 80 characters. The sequence ends if another line starting with a ">" appears; this indicates the start of another sample sequence. Example sequences in FASTA format can be seen by clicking on the Example button right above the input box.

Step 3. Click on the Submit button to see the predicted result. For instance, when using the three peptide sequences as an input and clicking the Submit button, you will see the following shown on the screen of your computer: the outcome for the 1st query example is "Ca-conotoxin"; the outcome for the 2nd query sample is "K-conotoxin"; the outcome for the 3rd query sample is "Na-conotoxin." All these results are fully consistent with the experimental observations. It takes only a few seconds for the above computation before the predicted result appears on your computer screen; the more number of query sequences, the longer time it usually needs.

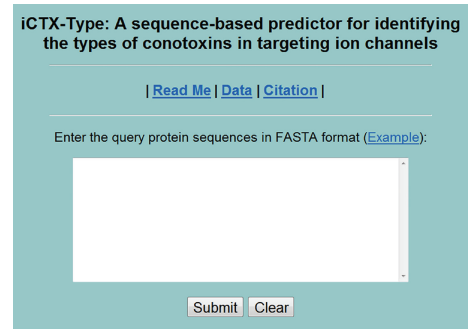


Figure 5: A screenshot to show the top page of the iCTX-Type web server. Its website address is http://lin.uestc.edu.cn/server/iCTX-Type.

Step 4. Click on the Data button to download the benchmark datasets used to train and test the iCTX-Type predictor.

Step 5. Click on the Citation button to find the relevant papers that document the detailed development and algorithm of iCTX-Type.

Caveats. The input query sequences must be formed by the single-letter codes of the 20 native amino acids; any other characters such as "B," "X," "U," and "Z" are invalid and should not be part of the peptide sequence.

4. Conclusion

It is anticipated that iCTX-Type may become a useful high throughput tool for both basic research and drug development, particularly for in-depth investigation into the mechanisms of ion-channels and developing new drugs to treat chronic pain, epilepsy, spasticity, and cardiovascular diseases, among others.

It is instructive to point out that since the binding of conotoxins to ion-channel is highly selective and specific, the information obtained by iCTX-Type in identifying the

types of conotoxins may be also very useful for designing ion channel inhibitors according to the Chou's distorted key theory as elaborated in [99] and briefed in a Wikipedia article at http://en.wikipedia.org/wiki/Chou's_distorted_key_theory_for_peptide_drugs.

Conflict of Interests

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

Acknowledgments

The authors wish to thank the anonymous reviewers for their constructive comments, which were very helpful for strengthening the presentation of this study. This work was supported by the National Nature Scientific Foundation of China (nos. 61202256, 61301260, and 61100092), the Nature Scientific Foundation of Hebei Province (no. C2013209105), and the Fundamental Research Funds for the Central Universities (nos. ZYGX2012J113 and ZYGX2013J102).

References

[1] I. S. Gabashvili, B. H. Sokolowski, C. C. Morton, and A. B. Giersch, "Ion channel gene expression in the inner ear," *Journal of the Association for Research in Otolaryngology*, vol. 8, no. 3, pp. 305–328, 2007.

[2] J. R. Schnell and J. J. Chou, "Structure and mechanism of the M2 proton channel of influenza A virus," *Nature*, vol. 451, no. 7178, pp. 591–595, 2008.

[3] R. M. Pielak, J. R. Schnell, and J. J. Chou, "Mechanism of drug inhibition and drug resistance of influenza A M2 channel," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 18, pp. 7379–7384, 2009.

[4] B. OuYang, S. Xie, M. J. Berardi et al., "Unusual architecture of the p7 channel from hepatitis C virus," *Nature*, vol. 498, no. 7455, pp. 521–525, 2013.

[5] X. Xiao, J. L. Min, and P. Wang, "Predict drug-protein interaction in cellular networking," *Current Topics in Medicinal Chemistry*, vol. 13, no. 14, pp. 1707–1712, 2013.

[6] K.-C. Chou, "Insights from modeling three-dimensional structures of the human potassium and sodium channels," *Journal of Proteome Research*, vol. 3, no. 4, pp. 856–861, 2004.

[7] L. F. Yuan, C. Ding, S. H. Guo, H. Ding, W. Chen, and H. Lin, "Prediction of the types of ion channel-targeted conotoxins based on radial basis function network," *Toxicology in Vitro*, vol. 27, no. 2, pp. 852–856, 2013.

[8] N. L. Daly and D. J. Craik, "Structural studies of conotoxins," *IUBMB Life*, vol. 61, no. 2, pp. 144–150, 2009.

[9] S. Mondal, R. Bhavna, R. Mohan Babu, and S. Ramakumar, "Pseudo amino acid composition and multi-class support vector machines approach for conotoxin superfamily classification," *Journal of Theoretical Biology*, vol. 243, no. 2, pp. 252–260, 2006.

[10] K.-C. Chou, "Prediction of protein cellular attributes using pseudo-amino acid composition," *Proteins*, vol. 43, no. 3, pp. 246–255, 2001, Erratum in: *Proteins*, vol. 44, no. 1, article 60, 2001.

[11] K. C. Chou, "Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes," *Bioinformatics*, vol. 21, no. 1, pp. 10–19, 2005.

[12] H. Lin and Q. Z. Li, "Predicting conotoxin superfamily and family by using pseudo amino acid composition and modified Mahalanobis discriminant," *Biochemical and Biophysical Research Communications*, vol. 354, no. 2, pp. 548–551, 2007.

[13] J. B. Yin, Y. X. Fan, and H. B. Shen, "Conotoxin superfamily prediction using diffusion maps dimensionality reduction and subspace classifier," *Current Protein and Peptide Science*, vol. 12, no. 6, pp. 580–588, 2011.

[14] S. Laht, D. Koua, L. Kaplinski, F. Lisacek, R. Stöcklin, and M. Remm, "Identification and classification of conopeptides using profile hidden Markov Models," *Biochimica et Biophysica Acta*, vol. 1824, no. 3, pp. 488–492, 2012.

[15] D. Koua, S. Laht, L. Kaplinski et al., "Position-specific scoring matrix and hidden Markov model complement each other for the prediction of conopeptide superfamilies," *Biochimica et Biophysica Acta*, vol. 1834, no. 4, pp. 717–724, 2013.

[16] K. H. Gowd, K. K. Dewan, P. Iengar, K. S. Krishnan, and P. Balaram, "Probing peptide libraries from Conus achatinus using mass spectrometry and cDNA sequencing: identification of δ and ω-conotoxins," *Journal of Mass Spectrometry*, vol. 43, no. 6, pp. 791–805, 2008.

[17] D. R. Hillyard, M. J. Mcintosh, R. M. Jones et al., "O-superfamily conotoxin peptides," Patent number JP2003533178, 2008.

[18] K.-C. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 273, no. 1, pp. 236–247, 2011.

[19] S. H. Guo, E. Z. Deng, L. Q. Xu, H. Ding, H. Lin, and W. Chen, "iNuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition," *Bioinformatics*, 2014.

[20] Y. Xu, J. Ding, and L. Y. Wu, "iSNO-PseAAC: predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition," *PLoS ONE*, vol. 8, no. 2, Article ID e55844, 2013.

[21] W. R. Qiu and X. Xiao, "iRSpot-TNCPseAAC: identify recombination spots with trinucleotide composition and pseudo amino acid components," *International Journal of Molecular Sciences*, vol. 15, no. 2, pp. 1746–1766, 2014.

[22] B. Liu, D. Zhang, R. Xu et al., "Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection," *Bioinformatics*, vol. 30, no. 4, pp. 472–479, 2014.

[23] W. Chen, P. M. Feng, H. Lin, and K. C. Chou, "iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition," *Nucleic Acids Research*, vol. 41, no. 6, article e68, 2013.

[24] J. L. Xiao, X. Min, and K.-C. Chou, "iEzy-Drug: a web server for identifying the interaction between enzymes and drugs in cellular networking," *BioMed Research International*, vol. 2013, Article ID 701317, 13 pages, 2013.

[25] X. Xiao, J. L. Min, and P. Wang, "iCDI-PseFpt: identify the channel-drug interaction in cellular networking with PseAAC and molecular fingerprints," *Journal of Theoretical Biology C*, vol. 337, pp. 71–79, 2013.

[26] Y. Xu, X. J. Shao, L. Y. Wu, N. Y. Deng, and K. C. Chou, "iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins," *PeerJ*, vol. 1, article e171, 2013.

[27] W. Chen, H. Lin, P. M. Feng, C. Ding, and Y. C. Zuo, "iNuc-PhysChem: a sequence-based predictor for identifying nucleosomes via physicochemical properties," *PLoS ONE*, vol. 7, no. 10, Article ID e47843, 2012.

[28] Y. Xu, X. Wen, X. J. Shao, and N. Y. Deng, "iHyd-PseAAC: predicting hydroxyproline and hydroxylysine in proteins by incorporating dipeptide position-specific propensity into pseudo amino acid composition," *International Journal of Molecular Sciences*, vol. 15, no. 5, pp. 7594–7610, 2014.

[29] T. U. Consortium, "Reorganizing the protein space at the Universal Protein Resource (UniProt)," *Nucleic Acids Research*, vol. 40, pp. D71–D75, 2012.

[30] K. C. Chou and H. B. Shen, "Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers," *Journal of Proteome Research*, vol. 5, no. 8, pp. 1888–1897, 2006.

[31] K. C. Chou and H. B. Shen, "Recent progress in protein subcellular location prediction," *Analytical Biochemistry*, vol. 370, no. 1, pp. 1–16, 2007.

[32] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, "CD-HIT: accelerated for clustering the next-generation sequencing data," *Bioinformatics*, vol. 28, no. 23, pp. 3150–3152, 2012.

[33] J. C. Wootton and S. Federhen, "Statistics of local complexity in amino acid sequences and sequence databases," *Computers and Chemistry*, vol. 17, no. 2, pp. 149–163, 1993.

[34] J. J. Chou, "A formulation for correlating properties of peptides and its application to predicting human immunodeficiency virus protease-cleavable sites in proteins," *Biopolymers*, vol. 33, no. 9, pp. 1405–1414, 1993.

[35] K. C. Chou, "Prediction of G-protein-coupled receptor classes," *Journal of Proteome Research*, vol. 4, no. 4, pp. 1413–1418, 2005.

[36] M. Wang, J. Yang, Z. J. Xu, and K.C. Chou, "SLLE for predicting membrane protein types," *Journal of Theoretical Biology*, vol. 232, no. 1, pp. 7–15, 2005.

[37] X. Xiao, P. Wang, and K.-C. Chou, "Predicting protein structural classes with pseudo amino acid composition: an approach using geometric moments of cellular automaton image," *Journal of Theoretical Biology*, vol. 254, no. 3, pp. 691–696, 2008.

[38] K. Y. Feng, Y. D. Cai, and K. C. Chou, "Boosting classifier for predicting protein domain structural class," *Biochemical and Biophysical Research Communications*, vol. 334, no. 1, pp. 213–217, 2005.

[39] Y. D. Cai and K. C. Chou, "Artificial neural network model for predicting α-turn types," *Analytical Biochemistry*, vol. 268, no. 2, pp. 407–409, 1999.

[40] T. B. Thompson, C. Zheng, and K.-C. Chou, "Neural network prediction of the HIV-1 protease cleavage sites," *Journal of Theoretical Biology*, vol. 177, no. 4, pp. 369–379, 1995.

[41] C. T. Zhang and K. C. Chou, "An optimization approach to predicting protein structural class from amino acid composition," *Protein Science*, vol. 1, no. 3, pp. 401–408, 1992.

[42] P. M. Feng, W. Chen, and H. Lin, "iHSP-PseRAAAC: identifying the heat shock protein families using pseudo reduced amino acid alphabet composition," *Analytical Biochemistry*, vol. 442, no. 1, pp. 118–125, 2013.

[43] X. Xiao, P. Wang, and K. C. Chou, "iNR-physchem: a sequence-based predictor for identifying nuclear receptors and their subfamilies via physical-chemical property matrix," *PLoS ONE*, vol. 7, no. 2, Article ID e30869, 2012.

[44] W. Z. Lin, J. A. Fang, X. Xiao, and K. C. Chou, "iDNA-prot: identification of DNA binding proteins using random forest with grey model," *PLoS ONE*, vol. 6, no. 9, Article ID e24756, 2011.

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

[46] Y. D. Cai and K. C. Chou, "Predicting subcellular localization of proteins in a hybridization space," *Bioinformatics*, vol. 20, no. 7, pp. 1151–1156, 2004.

[47] K. C. Chou and Y. D. Cai, "Prediction of protease types in a hybridization space," *Biochemical and Biophysical Research Communications*, vol. 339, no. 3, pp. 1015–1020, 2006.

[48] H. Shen and K. C. Chou, "Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo-amino acid composition to predict membrane protein types," *Biochemical and Biophysical Research Communications*, vol. 334, no. 1, pp. 288–292, 2005.

[49] K. C. Chou and H. B. Shen, "Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites," *Journal of Proteome Research*, vol. 6, no. 5, pp. 1728–1734, 2007.

[50] H. B. Shen and K. C. Chou, "A top-down approach to enhance the power of predicting human protein subcellular localization: Hum-mPLoc 2.0," *Analytical Biochemistry*, vol. 394, no. 2, pp. 269–274, 2009.

[51] T. L. Zhang, Y. S. Ding, and K. C. Chou, "Prediction protein structural classes with pseudo-amino acid composition: approximate entropy and hydrophobicity pattern," *Journal of Theoretical Biology*, vol. 250, no. 1, pp. 186–193, 2008.

[52] X. Xiao, P. Wang, and K. C. Chou, "GPCR-2L: predicting G protein-coupled receptors and their types by hybridizing two different modes of pseudo amino acid compositions," *Molecular BioSystems*, vol. 7, no. 3, pp. 911–919, 2011.

[53] H. B. Shen, J. Yang, and K. C. Chou, "Fuzzy KNN for predicting membrane protein types from pseudo-amino acid composition," *Journal of Theoretical Biology*, vol. 240, no. 1, pp. 9–13, 2006.

[54] X. Xiao, J. L. Min, and P. Wang, "iGPCR-Drug: a web server for predicting interaction between GPCRs and drugs in cellular networking," *PLoS ONE*, vol. 8, no. 8, Article ID e72234, 2013.

[55] X. Xiao, P. Wang, W.-Z. Lin, J.-H. Jia, and K.-C. Chou, "iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types," *Analytical Biochemistry*, vol. 436, no. 2, pp. 168–177, 2013.

[56] K. C. Chou, "Some remarks on predicting multi-label attributes in molecular biosystems," *Molecular Biosystems*, vol. 9, no. 6, pp. 1092–1100, 2013.

[57] H. Nakashima, K. Nishikawa, and T. Ooi, "The folding type of a protein is relevant to the amino acid composition," *Journal of Biochemistry*, vol. 99, no. 1, pp. 153–162, 1986.

[58] J. Cedano, P. Aloy, J. A. Pérez-Pons, and E. Querol, "Relation between amino acid composition and cellular location of proteins," *Journal of Molecular Biology*, vol. 266, no. 3, pp. 594–600, 1997.

[59] G.-P. Zhou, "An intriguing controversy over protein structural class prediction," *Protein Journal*, vol. 17, no. 8, pp. 729–738, 1998.

[60] S.-X. Lin and J. Lapointe, "Theoretical and experimental biology in one," *Journal of Biomedical Science and Engineering (JBiSE)*, vol. 6, pp. 435–442, 2013.

[61] X.-B. Zhou, C. Chen, Z.-C. Li, and X.-Y. Zou, "Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes," *Journal of Theoretical Biology*, vol. 248, no. 3, pp. 546–551, 2007.

[62] L. Nanni and A. Lumini, "Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization," *Amino Acids*, vol. 34, no. 4, pp. 653–660, 2008.

[63] D. N. Georgiou, T. E. Karakasidis, J. J. Nieto, and A. Torres, "Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 257, no. 1, pp. 17–26, 2009.

[64] M. Esmaeili, H. Mohabatkar, and S. Mohsenzadeh, "Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses," *Journal of Theoretical Biology*, vol. 263, no. 2, pp. 203–209, 2010.

[65] H. Mohabatkar, "Prediction of cyclin proteins using chou's pseudo amino acid composition," *Protein and Peptide Letters*, vol. 17, no. 10, pp. 1207–1214, 2010.

[66] S. S. Sahu and G. Panda, "A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction," *Computational Biology and Chemistry*, vol. 34, no. 5-6, pp. 320–327, 2010.

[67] H. Mohabatkar, M. Mohammad Beigi, and A. Esmaeili, "Prediction of GABA(A) receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine," *Journal of Theoretical Biology*, vol. 281, no. 1, pp. 18–23, 2011.

[68] M. Mohammad Beigi, M. Behjati, and H. Mohabatkar, "Prediction of metalloproteinase family based on the concept of Chou's pseudo amino acid composition using a machine learning approach," *Journal of Structural and Functional Genomics*, vol. 12, no. 4, pp. 191–197, 2011.

[69] S. Mei, "Multi-kernel transfer learning based on Chou's PseAAC formulation for protein submitochondria localization," *Journal of Theoretical Biology*, vol. 293, pp. 121–130, 2012.

[70] L. Nanni, S. Brahnam, and A. Lumini, "Wavelet images and Chou's pseudo amino acid composition for protein classification," *Amino Acids*, vol. 43, no. 2, pp. 657–665, 2012.

[71] L. Nanni, A. Lumini, D. Gupta, and A. Garg, "Identifying bacterial virulent proteins by fusing a set of classifiers based on variants of Chou's Pseudo amino acid composition and on evolutionary information," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, no. 2, pp. 467–475, 2012.

[72] M. K. Gupta, R. Niyogi, and M. Misra, "An alignment-free method to find similarity among protein sequences via the general form of Chou's pseudo amino acid composition," *SAR and QSAR in Environmental Research*, vol. 24, no. 7, pp. 597–609, 2013.

[73] Z. Hajisharifi, M. Piryaiee, M. Mohammad Beigi, M. Behbahani, and H. Mohabatkar, "Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test," *Journal of Theoretical Biology*, vol. 341, pp. 34–40, 2014.

[74] C. Huang and J. Yuan, "Using radial basis function on the general form of Chou's pseudo amino acid composition and PSSM to predict subcellular locations of proteins with both single and multiple sites," *Biosystems*, vol. 113, no. 1, pp. 50–57, 2013.

[75] C. Huang and J. Q. Yuan, "Predicting protein subchloroplast locations with both single and multiple sites via three different modes of Chou's pseudo amino acid compositions," *Journal of Theoretical Biology*, vol. 335, pp. 205–212, 2013.

[76] H. Mohabatkar, M. Mohammad Beigi, K. Abdolahi, and S. Mohsenzadeh, "Prediction of allergenic proteins by means of the concept of Chou's pseudo amino acid composition and a machine learning approach," *Medicinal Chemistry*, vol. 9, no. 1, pp. 133–137, 2013.

[77] A. N. Sarangi, M. Lohani, and R. Aggarwal, "Prediction of essential proteins in prokaryotes by incorporating various physico-chemical features into the general form of Chou's pseudo amino acid composition," *Protein and Peptide Letters*, vol. 20, no. 7, pp. 781–795, 2013.

[78] S. Wan, M. W. Mak, and S. Y. Kung, "GOASVM: a subcellular location predictor by incorporating term-frequency gene ontology into the general form of Chou's pseudo-amino acid composition," *Journal of Theoretical Biology*, vol. 323, pp. 40–48, 2013.

[79] W. Chen, T. Y. Lei, D. C. Jin, and H. Lin, "PseKNC: a flexible web-server for generating pseudo K-tuple nucleotide composition," *Analytical Biochemistry*, vol. 456, pp. 53–60, 2014.

[80] B.-Q. Li, T. Huang, L. Liu, Y.-D. Cai, and K.-C. Chou, "Identification of colorectal cancer related genes with mrmr and shortest path in protein-protein interaction network," *PLoS ONE*, vol. 7, no. 4, Article ID e33393, 2012.

[81] T. Huang, J. Wang, Y.-D. Cai, H. Yu, and K.-C. Chou, "Hepatitis C virus network based classification of hepatocellular cirrhosis and carcinoma," *PLoS ONE*, vol. 7, no. 4, Article ID e34460, 2012.

[82] Y. Jiang, T. Huang, L. Chen, Y. F. Gao, Y. Cai, and K.-C. Chou, "Signal propagation in protein interaction network during colorectal cancer progression," *BioMed Research International*, vol. 2013, Article ID 287019, 9 pages, 2013.

[83] H.-B. Shen and K.-C. Chou, "PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition," *Analytical Biochemistry*, vol. 373, no. 2, pp. 386–388, 2008.

[84] 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.

[85] D. S. Cao, Q. S. Xu, and Y. Z. Liang, "propy: a tool to generate various modes of Chou's PseAAC," *Bioinformatics*, vol. 29, no. 7, pp. 960–962, 2013.

[86] P. Du, S. Gu, and Y. Jiao, "PseAAC-General: fast building various modes of general form of Chou's pseudo-amino acid composition for large-scale protein datasets," *International Journal of Molecular Sciences*, vol. 15, no. 3, pp. 3495–3506, 2014.

[87] L. C. Chen YW, "Combining SVMs with various feature selection strategies," in *Feature Extraction*, I. Guyon, N. Nikravesh, S. Gunn, and L. Zadeh, Eds., pp. 315–324, Springer, Berlin, Germany, 2006.

[88] H. Lin, H. Ding, F.-B. Guo, and J. Huang, "Prediction of subcellular location of mycobacterial protein using feature selection techniques," *Molecular Diversity*, vol. 14, no. 4, pp. 667–671, 2010.

[89] K.-C. Chou and Y.-D. Cai, "Using functional domain composition and support vector machines for prediction of protein subcellular location," *The Journal of Biological Chemistry*, vol. 277, no. 48, pp. 45765–45769, 2002.

[90] Y.-D. Cai, G.-P. Zhou, and K.-C. Chou, "Support vector machines for predicting membrane protein types by using functional domain composition," *Biophysical Journal*, vol. 84, no. 5, pp. 3257–3263, 2003.

[91] N. Cristianini and J. Shawe-Taylor, *An Introduction of Support Vector Machines and Other Kernel-Based Learning Methodds*, Cambridge University Press, Cambridge, UK, 2000.

[92] K. C. Chou and C. T. Zhang, "Review: prediction of protein structural classes," *Critical Reviews in Biochemistry and Molecular Biology*, vol. 30, no. 4, pp. 275–349, 1995.

[93] G. P. Zhou and N. Assa-Munt, "Some insights into protein structural class prediction," *Proteins: Structure, Function and Genetics*, vol. 44, no. 1, pp. 57–59, 2001.

[94] K.-C. Chou, Z.-C. Wu, and X. Xiao, "ILoc-Hum: using the accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites," *Molecular BioSystems*, vol. 8, no. 2, pp. 629–641, 2012.

[95] K.-C. Chou, Z.-C. Wu, and X. Xiao, "iLoc-Euk: a multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins," *PLoS ONE*, vol. 6, no. 3, Article ID e18258, 2011.

[96] K.-C. Chou, "Using subsite coupling to predict signal peptides," *Protein Engineering*, vol. 14, no. 2, pp. 75–79, 2001.

[97] K. C. Chou, "Prediction of signal peptides using scaled window," *Peptides*, vol. 22, no. 12, pp. 1973–1979, 2001.

[98] K. C. Chou and H. B. Shen, "Review: recent advances in developing web-servers for predicting protein attributes," *Natural Science*, vol. 1, no. 2, pp. 63–92, 2009.

[99] K. C. Chou, "Review: prediction of human immunodeficiency virus protease cleavage sites in proteins," *Analytical Biochemistry*, vol. 233, no. 1, pp. 1–14, 1996.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 504727, 6 pages
http://dx.doi.org/10.1155/2014/504727

Research Article

An Association Study between Genetic Polymorphism in the Interleukin-6 Receptor Gene and Coronary Heart Disease

Jiangqing Zhou,[1] Xiaoliang Chen,[1,2] Huadan Ye,[2] Ping Peng,[1,2] Yanna Ba,[1,2] Xi Yang,[1] Xiaoyan Huang,[1] Yae Lu,[1] Xin Jiang,[1] Jiangfang Lian,[1] and Shiwei Duan[2]

[1] *Ningbo Medical Center, Lihuili Hospital, Ningbo University, 57 Xingning Road, Ningbo, Zhejiang 315211, China*
[2] *School of Medicine, Ningbo University, 818 Fenghua Road, Ningbo, Zhejiang 315211, China*

Correspondence should be addressed to Jiangfang Lian; hjmpin@163.com and Shiwei Duan; duanshiwei@nbu.edu.cn

Received 7 February 2014; Revised 9 April 2014; Accepted 16 April 2014; Published 26 May 2014

Academic Editor: Hongwei Wang

The goal of our study is to test the association of IL6R rs7529229 polymorphism with CHD through a case-control study in Han Chinese population and a meta-analysis. Our result showed there is a lack of association between IL6R rs7529229 polymorphism and CHD on both genotype and allele levels in Han Chinese ($P > 0.05$). However, a meta-analysis among 11678 cases and 12861 controls showed that rs7529229-C allele was significantly associated with a decreased risk of CHD, especially in Europeans ($P < 0.0001$, odds ratio = 0.93, 95% confidential interval = 0.89–0.96). Since there is significant difference among different populations, further studies are warranted to test the contribution of rs7529229 to CHD in other ethnic populations.

1. Introduction

Coronary heart disease (CHD) is one of the leading causes of human deaths in the developed and developing countries such as China [1]. As a complex disease, CHD results from the interaction between genetic and environmental factors. CHD is one of the most common manifestations of atherosclerosis that is related to inflammation [2]. CHD is regarded as a chronic inflammatory disease [3] that has been shown to be associated with the response to inflammatory signaling [4].

Interleukin-6 is an inflammatory cytokine [5], whose synthesis is stimulated by its binding to *IL6R*. *IL6R* signaling activates an intracellular signaling cascade leading to the inflammatory response [6] and thus has become an important therapeutic target for prevention of CHD [7, 8].

Human *IL6R* is located on 1q21, a susceptible locus for CHD. *IL6R* rs7529229 is a T/C variation associated with both *IL6R* level and a decreased risk of CHD events in Europeans [7]. Since there is a lack of evidence concerning its role in CHD in Han Chinese, the goal of our study was to replicate the association between *IL6R* rs7529229 polymorphism and CHD in Han Chinese. In addition, we performed a meta-analysis of the available case-control studies between rs7529229 of *IL6R* gene and CHD.

2. Methods

2.1. Sample Collection. A total of 459 unrelated individuals were selected between May 2011 and November 2013 from Ningbo Lihuili Hospital, Zhejiang, China. Of these, 263 patients had CHD (males: 181; females: 82; age: 61.04 ± 8.68 years) and 196 patients were non-CHD controls (males: 98; females: 98; age: 57.76 ± 7.97 years). The patients had been examined by standardized coronary angiography according to the Seldinger method [9] and were judged by at least two independent cardiologists. In CHD cases, patients ($n = 263$) were diagnosed with the angiographic evidence that coronary artery stenosis was greater than 50% in one or more major coronary arteries [10]. Gensini scoring system was used to determine the severity of CHD [11]. A total of 196 patients, who did not have detectable coronary stenosis and atherosclerotic vascular disease, were considered as controls. All individuals had no cardiomyopathy or congenital heart, liver, or renal diseases. All the samples were Han Chinese living in Ningbo of China. The blood samples were collected by the same investigators. Blood samples were collected in 3.2% citrate sodium-treated tubes and then stored at −80°C. The study protocol was approved by the Ethics Committee of

Lihuili Hospital in Ningbo and informed written consent was obtained from all subjects.

2.2. PCR Amplification and SNP Genotyping. Human genomic DNA was prepared from peripheral blood samples using the nucleic acid extraction automatic analyzer (Lab-Aid 820, Xiamen, China) and was quantified using the PicoGreen dsDNA Quantification kit (Molecular Probes Inc., Eugene, OR, USA). Amplification was performed on the ABI GeneAmp PCR System 9700 Dual 96-Well Sample Block Module (Applied Biosystems, Foster City, CA, USA). Genomic DNA was subjected to polymerase chain reac-tion (PCR) with primers specific to *IL6R* gene. The sequences of the two allele-specific primers were 5′-GCGGCA-GGGCGGCAATGTGGTCGTGGTGAGTTACC C-3′ and 5′-GATTACCGAATGTGGTCGTGGTGAGTTA CCT-3′. The sequence of a reverse primer was 5 -TTTCTATGA-TTCCCTTTCACAGAGGTTTGA-3′. The reaction was performed with an initial denaturation stage at 95°C for 30 sec, followed by 40 cycles at 95°C for 30 sec, 59°C for 30 sec, and 72°C for 30 sec, and a final extension at 72°C for 30 sec. Genotyping of the PCR products was performed on the Roche LightCycler 480 Fluorescence Real-Time PCR System (Roche, Rotkreuz, Switzerland) using melting temperature shift (Tm-shift) according to the manufacturer's instructions [12, 13]. Tm-shift method uses two allele-specific primers and one reverse primer to amplify the polymorphic region encoding the targeted variant, and genotypes can be determined by inspection of a melting curve [14, 15] (Figure 1). To verify the repeatability and stability of experiment, 5% of random samples and 18 control samples (including 9 negative and 9 positive controls) were used for quality control.

2.3. Retrieval of Published Studies. A search was performed for the publications from 2008 to 2013 in the electronic databases (including PubMed, EMbase, Web of Science, and Cochrane Library). The search keywords included "coronary heart disease" or "coronary artery disease" or "myocardial infarction" combined with "IL6R" or "interleukin-6 receptor" or "rs7529229" or "polymorphism" and "genetic association." We read the full-text articles to collect the relevant information. References listed on the retrieved articles and previous meta-analyses on this subject were searched to appraise other studies of potential relevance. The included studies for the meta-analysis need to be case-control design and need to have information consisting of ORs and their 95% CIs or genotyping data to measure the relative risk. Data extraction was carried out by at least two reviewers (Xiaoliang Chen and Ping Peng) on a standard protocol, and the consensus data were established by discussion. In the meta-analyses, the following data collection was included: name of the first author, publication year, country, ethnic population, study stage, numbers of individuals in the case and the control groups, OR, and 95% CI.

2.4. Statistical Analysis. Hardy-Weinberg equilibrium (HWE) was analyzed using the Arlequin software (v3.5)

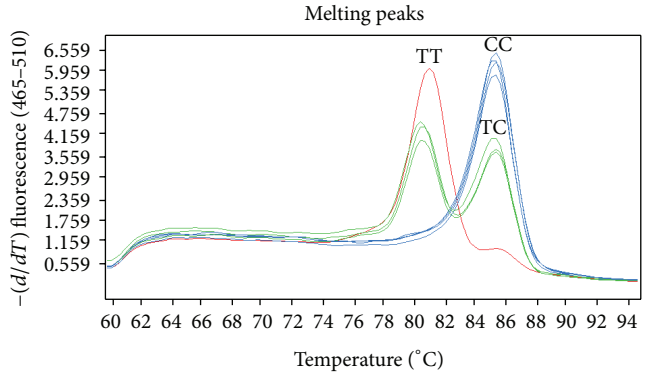


Figure 1: Melting temperature- (Tm-) shift method was used for SNP genotyping.

[16]. The statistical software package SPSS v.18.0 (SPSS, Chicago, IL, USA) was used for the following analyses. Continuous data were expressed as mean ± SD and Student t-test was employed to analyze differences between two study groups. χ^2 analysis was used to compare the categorical variables. Genotype and allele frequencies between CHD cases and healthy controls among the different subgroups were compared using χ^2 test. Association between Gensini scores and rs7529229 was compared using linear regression test. Mann-Whitney test was used for the association of Gensini scores with rs7529229 under the dominant and the recessive models. The power of the study was estimated by the Power and Sample Size Calculation software (v3.0.43) [17]. Meta-analysis was performed by Stata software version 11.0 (Stata Corporation, College Station, TX) and in accordance with Stroup's study [18]. Heterogeneity of the studies was evaluated by the I^2 statistic at the significant level of 0.05. The combined odds ratios (ORs) along with their 95% confidence intervals (CIs) were assessed with inverse-variance fixed-effect model (subtotal $I^2 = 20.4\%$, $P = 0.274$; overall $I^2 = 12.5\%$, $P = 0.441$) [19]. Subgroup meta-analysis was performed by ethnicity. Sensitivity analysis was conducted by omitting each study in turn. Funnel plots and Egger regression tests [20] were used to estimate the publication bias. A two-sided $P < 0.05$ was considered to be statistically significant.

3. Results

3.1. Basic Characteristics of the Study Population. As shown in Table 1, the prevalence of essential hypertension (EH), diabetes mellitus (DM), and smoking history was significantly higher in CHD patients than controls ($P < 0.05$). However, there were no significant differences between the two groups for a series of biochemical parameters, including high-density lipoprotein cholesterol (HDL-C), triglycerides, low-density lipoprotein cholesterol (LDL-C), and total cholesterol.

3.2. Genotype and Allele Distribution of rs7529229 in CHD Cases and Controls. Genotype distribution of rs7529229 in both CHD cases and controls met HWE (Table 2). Genotype

Table 1: Basic characteristics of the study population.

	CHD (n = 263)	Controls (n = 196)	P
Male, n (%)	182 (69.2%)	98 (50%)	<0.001
Smoking, n (%)	111 (43.5%)	53 (27.3%)	<0.001
Hypertension, n (%)	161 (63.1%)	100 (51.5%)	0.014
Diabetes, n (%)	53 (20.7%)	12 (6.1%)	<0.001
Mean age, years	61.04 ± 8.68	57.76 ± 7.97	<0.001
LDL-C (mmol/L)	2.54 ± 0.94	2.49 ± 0.89	0.600
Total cholesterol (mmol/L)	4.33 ± 1.10	4.28 ± 1.01	0.620
HDL-C (mmol/L)	1.06 ± 0.30	1.12 ± 0.29	0.048
Triglycerides (mmol/L)	1.64 ± 1.06	1.50 ± 0.86	0.125

Table 2: Genotype and allele distribution of rs7529229 in CHD cases and controls.

Group	Genotype (TT/TC/CC)	χ^2	P (df = 2)	HWE	Allele (T/C)	χ^2	P (df = 1)	OR (95% CI)
All CHD cases (n = 263)	77/133/53			0.80	287/239			
All controls (n = 196)	63/98/35	0.61	0.73	0.88	224/168	0.60	0.43	0.90 (0.69–1.17)
Female CHD cases (n = 82)	30/35/17			0.26	95/69			
Female controls (n = 98)	34/47/17	0.58	0.74	1.00	115/81	0.02	0.88	0.97 (0.63–1.47)
Male CHD cases (n = 181)	47/98/36			0.29	192/170			
Male controls (n = 98)	29/51/18	0.43	0.80	0.68	109/87	0.33	0.56	0.90 (0.63–1.27)

analysis of rs7529229 did not reveal significant difference between CHD cases and non-CHD controls (χ^2 = 0.61, P = 0.73). The allelic distribution of rs7529229 did not differ between CHD cases and non-CHD controls (P = 0.43, OR = 0.90, 95% CI = 0.69–1.17, Table 2). We further examined the roles of rs7529229 in males and females separately. However, no significant differences between cases and controls were observed in male and female subgroups (Table 2). In addition, we also performed an age-stratified analysis to investigate whether age influenced the contribution of rs7529229 to the risk of CHD. Again, no significant differences between CHD cases and controls were observed in all age-stratified subgroups (Table 3).

3.3. Stratified Analyses of rs7529229 between Cases and Controls by Smoking History or Status of Hypertension or Diabetes. Since smoking history, hypertension, and diabetes are risk factors of CAD [21], we further performed stratified association tests by the above three variables. Our results showed that there were no significant differences in the distribution of genotype and allele of rs7529229 between cases and controls (Table 4).

3.4. Association of rs7529229 with the Severity of Coronary Lesions. A linear regression test of the means of Gensini scores with rs7529229 genotype did not show a statistically significant correlation (Table 5). And there was no significant association between Gensini scores and rs7529229 under the dominant (Z = −0.38, P = 0.69, Table 5) and the recessive models (Z = −0.50, P = 0.61, Table 5).

3.5. Meta-Analysis of rs7529229 with CHD in Different Populations. A total of 41 studies were selected initially. After reading the full text of these articles, 9 eligible studies were harvested for the current meta-analysis of the association of rs7529229 with CHD [7, 22, 23]. Details of articles in the meta-analysis are shown in Figure 2. Our meta-analysis comprised 11,678 CHD cases and 12,861 controls from two ethnic populations (Europeans and Asians). No significant heterogeneity was found in this meta-analysis (P = 0.441, I^2 = 12.5%). Our result suggested that rs7529229-C allele was associated with CHD risk. A future subgroup meta-analysis showed that rs7529229 of *IL6R* gene was a protective factor of CHD, especially in Europeans (P < 0.0001, OR = 0.93, 95% CI = 0.89–0.96). Sensitivity analyses were repeatedly conducted when each particular study was omitted. As shown in Figure 3, the results were not altered with pooled ORs ranging from 0.92 to 0.94 for the meta-analysis in Europeans and Asians. There was no visual publication bias in Begg's funnel (P = 0.25) and Egger's regression plots (P = 0.251, Figure 4).

4. Discussion

In the present study, we aim to replicate previous significant association between *IL6R* rs7529229 polymorphism and the risk of CHD in Han Chinese. Our study analyzed the association of rs7529229 with both CHD susceptibility and its severity. We also explored the stratified association of rs7529229 with CHD, though we failed to observe significant associations between *IL6R* rs7529229 and the risk of CHD. The results of our study were inconsistent with the recent findings from a large study of European samples [7]. We speculated that the discrepancies might be due to ethnic difference in the prevalence of this SNP. In addition, a power calculation showed that our case-control study only had a 12.2% power to detect a relative risk of rs7529229 at a

Table 3: Post hoc analysis of rs7529229 with the risk of CHD in different age subgroups.

Age group	Genotype (TT/TC/CC)	χ^2	P (df = 2)	HWE	Allele (T/C)	χ^2	P (df = 1)	OR (95% CI)
≤55 CHD cases (n = 70)	19/37/14			0.80	75/65			
≤55 controls (n = 71)	25/36/10	1.49	0.47	0.80	86/56	1.40	0.23	0.75 (0.46–1.20)
55–65 CHD cases (n = 95)	23/50/22			0.68	96/94			
55–65 controls (n = 82)	25/39/18	0.89	0.64	0.82	89/75	0.49	0.48	0.86 (0.56–1.30)
≥65 CHD cases (n = 98)	35/46/17			0.83	116/80			
≥65 controls (n = 43)	13/23/7	0.54	0.76	0.75	49/37	0.12	0.72	1.09 (0.65–1.82)

Table 4: The stratified association analysis of rs7529229.

Group	Risk factor of CHD	Genotype (TT/TC/CC)	χ^2	P	T/C	χ^2	P
CHD (n = 115)	Smoking	31/61/23			123/107		
Control (n = 55)	Smoking	15/31/9	0.339	0.844	61/49	0.117	0.732
CHD (n = 148)	No smoking	46/72/30			164/132		
Control (n = 141)	No smoking	48/67/26	0.339	0.844	163/119	0.337	0.561
CHD (n = 165)	Hypertension	48/86/31			182/148		
Control (n = 102)	Hypertension	36/44/22	2.061	0.375	116/88	0.15	0.699
CHD (n = 98)	No hypertension	29/47/22			105/91		
Control (n = 94)	No hypertension	27/54/13	2.789	0.248	108/80	0.583	0.445
CHD (n = 57)	Diabetes	16/27/14			59/55		
Control (n = 14)	Diabetes	6/5/3	1.178	0.555	17/11	0.725	0.394
CHD (n = 206)	No diabetes	61/106/39			228/184		
Control (n = 182)	No diabetes	57/93/32	0.191	0.909	207/157	0.183	0.669

Table 5: Association tests of Gensini scores and CHD.

Genotype	Gensini score (mean/SD/median)	F/Z	P (df = 1)
TT (n = 77)	56.12/56.08/35.5		
TC (n = 133)	48.67/43.67/33.0		
CC (n = 53)	46.60/43.64/36.0	0.30	0.85
Recessive model			
TT + TC (n = 210)	51.40/48.59/35.2		
CC (n = 53)	46.60/43.64/36.0	−0.50	0.61
Dominant model			
TC + CC (n = 186)	48.08/43.56/34.5		
TT (n = 77)	56.12/56.08/35.5	−0.38	0.69

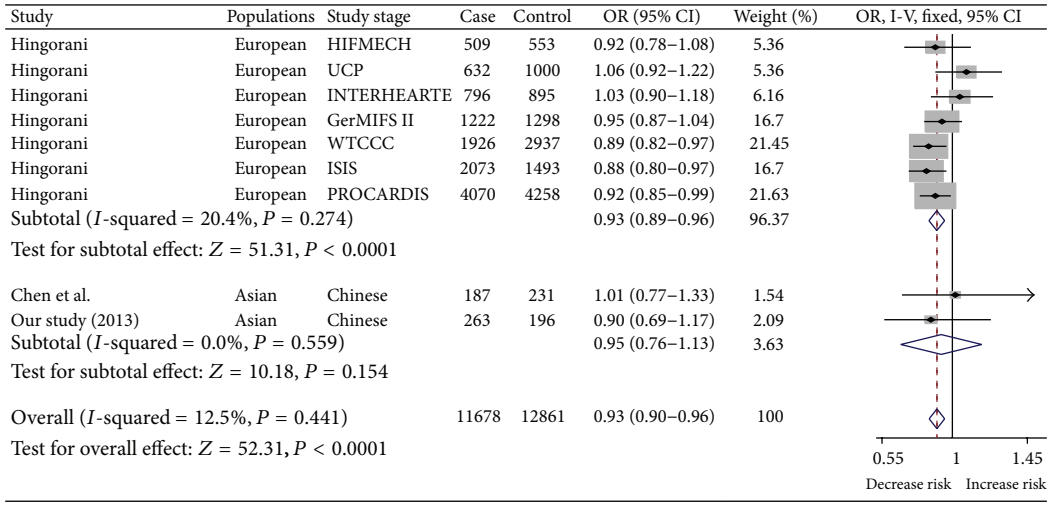


Figure 2: Meta-analysis of ten association studies of rs7529229 with CAD.

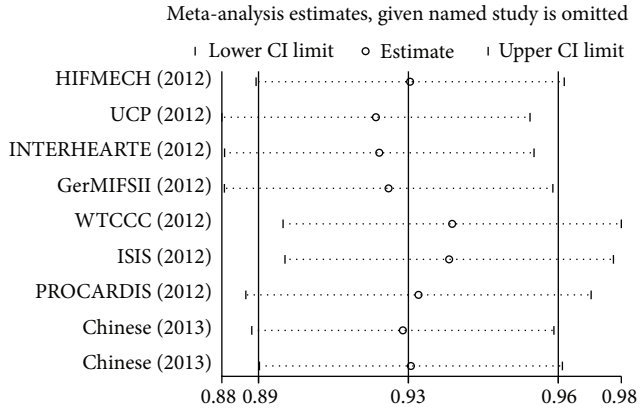


Figure 3: Sensitivity analysis for the association between *IL6R* rs7529229 polymorphism and CHD risk.

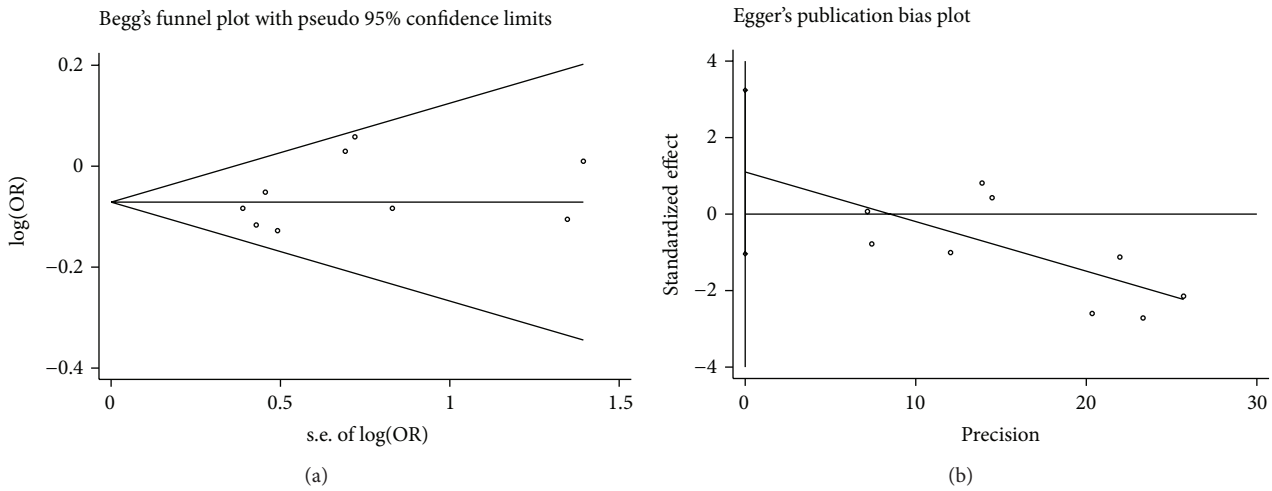


Figure 4: Begg's funnel plot and Egger's regression plot of 11 association tests between rs7529229 and CHD.

significant level of 0.05, suggesting that a lack of power was likely to explain our failure to find a significant association.

Our meta-analysis, including a total of 11,678 cases and 12,861 controls, examined the association between the rs7529229 polymorphism and CHD risk. We found that rs7529229 of *IL6R* gene was associated with the risk of CHD. A further subgroup analysis by race showed that rs7529229 of *IL6R* gene was a protective factor of CHD, especially in Europeans. Our study is the first association test between rs7629229 and CHD in the Chinese population. We carried out sensitivity analysis to assess the stability of this meta-analysis. Removal of each study did not alter the conclusion of the CHD risk, suggesting the reliability of these results. Meta-analysis can dramatically increase the power of association test through the combination of the data from various studies. For example, the power of some studies in the current meta-analysis is moderate (HIFMECH: 60.1%; UCP: 55.3%; INTERHEARTE: 32.2%; GerMIFSII: 65.2%; Chen et al.: 11.8%).

There are several limitations in our study. Firstly, the power of our case-control study only reached 12.2% at alpha level of 0.05, so we could not exclude the possibility of lack of power in our study mainly due to the relatively small sample size. Secondly, only one polymorphism of *IL6R* was investigated in the present study. According to the report in dbSNP, there were at least 256 SNPs on the *IL6R* gene locus. Therefore, the results of *IL6R* rs7529229 might not stand for the rest of the *IL6R* SNPs. Thirdly, we only searched the literatures in Chinese for the eligible research included in the meta-analysis. Meanwhile, case-control studies with negative results were more likely to be unpublished. Potential language and publication bias might exist in the meta-analysis.

In conclusion, our meta-analysis has established a strong contribution of rs7529229-C allele to reduced risk of CHD, especially in Europeans, although our case-control study is unable to find association of the *IL6R* with the risk of CHD. Further investigation on other SNPs on the gene is warranted to validate our findings in the Chinese population.

Conflict of Interests

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

Acknowledgments

The research was supported by Grants from the National Natural Science Foundation of China (31100919 and 81370207), Advanced Key Scientific and Technological Programs of Ningbo (2011C51001), Fund of Ningbo Science and Technology Innovation team (2011B82015), Natural Science Foundation of the Zhejiang Province (LY13H020008 and LR13H020003), and The Project of Ningbo Medicine and Science (2009A01), K. C. Wong Magna Fund in Ningbo University, and Ningbo Social Development Research Projects (2012C50032). Jianqing Zhou and Xiaoliang Chen are co-first authors of this work.

References

[1] J. He, D. Gu, X. Wu et al., "Major causes of death among men and women in China," *The New England Journal of Medicine*, vol. 353, no. 11, pp. 1124–1134, 2005.

[2] R. Ross, "Atherosclerosis—an inflammatory disease," *The New England Journal of Medicine*, vol. 340, no. 2, pp. 115–126, 1999.

[3] G. K. Hansson, "Inflammation, atherosclerosis, and coronary artery disease," *The New England Journal of Medicine*, vol. 352, no. 16, pp. 1626–1695, 2005.

[4] O. Harismendy, D. Notani, X. Song et al., "9p21 DNA variants associated with coronary artery disease impair interferon-γ 3 signalling response," *Nature*, vol. 470, no. 7333, pp. 264–268, 2011.

[5] T. Naka, N. Nishimoto, and T. Kishimoto, "The paradigm of IL-6: from basic science to medicine," *Arthritis Research*, vol. 4, no. 3, pp. 233–242, 2002.

[6] M. J. Boulanger, D. Chow, E. E. Brevnova, and K. C. Garcia, "Hexameric structure and assembly of the interieukin-6/IL-6 α-receptor/gp130 complex," *Science*, vol. 300, no. 5628, pp. 2101–2104, 2003.

[7] A. D. Hingorani and J. P. Casas, "The interleukin-6 receptor as a target for prevention of coronary heart disease: a mendelian randomisation analysis," *The Lancet*, vol. 379, no. 9822, pp. 1214–1224, 2012.

[8] H. S. Satti, S. Hussain, and Q. Javed, "Association of interleukin-6 gene promoter polymorphism with coronary artery disease in pakistani families," *The Scientific World Journal*, vol. 2013, Article ID 538365, 6 pages, 2013.

[9] Z. C. J. Higgs, D. A. L. Macafee, B. D. Braithwaite, and C. A. Maxwell-Armstrong, "The Seldinger technique: 50 years on," *The Lancet*, vol. 366, no. 9494, pp. 1407–1409, 2005.

[10] E. Rapaport, R. Bernard, and E. Corday, "Nomenclature and criteria for diagnosis of ischemic heart disease. Report of the Joint International Society and Federation of Cardiology/World Health Organization Task Force on standardization of clinical nomenclature," *Circulation*, vol. 59, no. 3, pp. 607–609, 1979.

[11] G. G. Gensini, "A more meaningful scoring system for determining the severity of coronary heart disease," *The American Journal of Cardiology*, vol. 51, no. 3, p. 606, 1983.

[12] J. Wang, K. Chuang, M. Ahluwalia et al., "High-throughput SNP genotyping by single-tube PCR with Tm-shift primers," *BioTechniques*, vol. 36, no. 6, pp. 885–893, 2005.

[13] S. Germer and R. Higuchi, "Single-tube genotyping without oligonucleotide probes," *Genome Research*, vol. 9, no. 1, pp. 72–78, 1999.

[14] S. Derzelle, C. Mendy, S. Laroche, and N. Madani, "Use of high-resolution melting and melting temperature-shift assays for specific detection and identification of *Bacillus anthracis* based on single nucleotide discrimination," *Journal of Microbiological Methods*, vol. 87, no. 2, pp. 195–201, 2011.

[15] S. H. Zhou, M. Liu, W. X. An et al., "Genotyping of human platelet antigen-15 by single closed-tube Tm-shift method," *International Journal of Laboratory Hematology*, vol. 34, no. 1, pp. 41–46, 2012.

[16] L. Excoffier and H. E. L. Lischer, "Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows," *Molecular Ecology Resources*, vol. 10, no. 3, pp. 564–567, 2010.

[17] W. D. Dupont and W. D. Plummer Jr., "Power and sample size calculations. A review and computer program," *Controlled Clinical Trials*, vol. 11, no. 2, pp. 116–128, 1990.

[18] D. F. Stroup, J. A. Berlin, S. C. Morton et al., "Meta-analysis of observational studies in epidemiology: a proposal for reporting," *Journal of the American Medical Association*, vol. 283, no. 15, pp. 2008–2012, 2000.

[19] R. DerSimonian and N. Laird, "Meta-analysis in clinical trials," *Controlled Clinical Trials*, vol. 7, no. 3, pp. 177–188, 1986.

[20] M. Egger, G. D. Smith, M. Schneider, and C. Minder, "Bias in meta-analysis detected by a simple graphical test," *British Medical Journal*, vol. 315, no. 7109, pp. 629–634, 1997.

[21] J. Frohlich and A. Al-Sarraf, "Cardiovascular risk and atherosclerosis prevention," *Cardiovascular Pathology*, vol. 22, no. 1, pp. 16–18, 2013.

[22] P. Elliott, J. C. Chambers, W. Zhang et al., "Genetic loci associated with C-reactive protein levels and risk of coronary heart disease," *Journal of the American Medical Association*, vol. 302, no. 1, pp. 37–48, 2009.

[23] Z. Chen, Q. Qian, C. Tang, J. Ding, Y. Feng, and G. Ma, "Association of two variants in the interleukin-6 receptor gene and premature coronary heart disease in a Chinese Han population," *Molecular Biology Reports*, vol. 40, no. 2, pp. 1021–1026, 2013.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 564940, 6 pages
http://dx.doi.org/10.1155/2014/564940

Research Article

Meta-Analysis of Low Density Lipoprotein Receptor (*LDLR*) rs2228671 Polymorphism and Coronary Heart Disease

Huadan Ye,[1,2] Qianlei Zhao,[3] Yi Huang,[1] Lingyan Wang,[4] Haibo Liu,[3] Chunming Wang,[3] Dongjun Dai,[1] Leiting Xu,[1] Meng Ye,[2] and Shiwei Duan[1]

[1] *Zhejiang Provincial Key Laboratory of Pathophysiology, School of Medicine, Ningbo University, Ningbo, Zhejiang 315211, China*
[2] *The Affiliated Hospital, School of Medicine, Ningbo University, Ningbo, Zhejiang 315211, China*
[3] *Yinzhou People's Hospital, School of Medicine, Ningbo University, Ningbo, Zhejiang 315040, China*
[4] *Bank of Blood Products, Ningbo No. 2 Hospital, Ningbo, Zhejiang 315010, China*

Correspondence should be addressed to Meng Ye; dryemeng@yahoo.com.cn and Shiwei Duan; duanshiwei@nbu.edu.cn

Received 29 January 2014; Revised 3 April 2014; Accepted 22 April 2014; Published 12 May 2014

Academic Editor: Hongwei Wang

Low density lipoprotein receptor (LDLR) can regulate cholesterol metabolism by removing the excess low density lipoprotein cholesterol (LDL-C) in blood. Since cholesterol metabolism is often disrupted in coronary heart disease (CHD), *LDLR* as a candidate gene of CHD has been intensively studied. The goal of our study is to evaluate the overall contribution of *LDLR* rs2228671 polymorphism to the risk of CHD by combining the genotyping data from multiple case-control studies. Our meta-analysis is involved with 8 case-control studies among 7588 cases and 9711 controls to test the association between *LDLR* rs2228671 polymorphism and CHD. In addition, we performed a case-control study of *LDLR* rs2228671 polymorphism with the risk of CHD in Chinese population. Our meta-analysis showed that rs2228671-T allele was significantly associated with a reduced risk of CHD ($P = 0.0005$, odds ratio (OR) = 0.83, and 95% confidence interval (95% CI) = 0.75–0.92). However, rs2228671-T allele frequency was rare (1%) and was not associated with CHD in Han Chinese ($P = 0.49$), suggesting an ethnic difference of *LDLR* rs2228671 polymorphism. Meta-analysis has established rs2228671 as a protective factor of CHD in Europeans. The lack of association in Chinese reflects an ethnic difference of this genetic variant between Chinese and European populations.

1. Introduction

Coronary heart disease (CHD) is a complex disease caused by an insufficient blood flow inside the coronary vessels [1]. The blockage of the arteries is often caused by the plaque accumulated in the wall of arteries. The plaque is formed by excess low density lipoprotein cholesterol (LDL-C) in blood that dramatically increases the risk of CHD [2]. Low density lipoprotein receptor (LDLR) plays a key role in the regulation of cholesterol metabolism by removing excess LDL-C in blood [3, 4].

CHD is caused by both environmental and genetic factors [5]. Variations of genes involved in lipoprotein and lipid metabolism are playing an important role in the susceptibility of CHD [6]. *LDLR* gene mutations can lead to deficiency or abnormality of LDLR in the cell membrane surface and thus disrupt lipid metabolism [4]. *LDLR* gene mutations are known to cause familial hypercholesterolemia (FH) [2] that is an important risk factor of CHD and other atherosclerotic diseases [7]. Recent genome-wide association studies (GWASs) showed that *LDLR* gene mutations were significantly associated with the abnormal blood lipid levels and CHD [8, 9]. Among the *LDLR* polymorphisms, rs2228671 was associated with LDL-C levels and CHD in German and British populations [10–14]. However, discrepancies were also shown in the association of *LDLR* rs2228671 with CHD in Italians and Germans [15, 16].

Meta-analysis is able to combine and review the results from previous studies [17, 18]. Meta-analysis improves the power of comprehensive statistics by pooling the data from different studies. In the present study, we performed a meta-analysis of *LDLR* rs2228671 polymorphism with CHD among 17299 individuals in 8 studies.

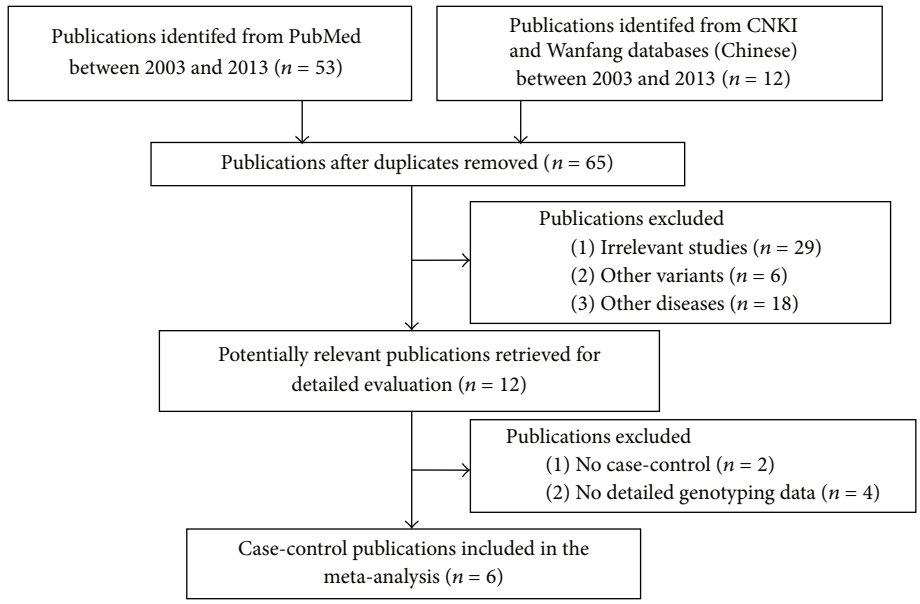


Figure 1: Flowing chart of selection publications in the current meta-analysis.

2. Material and Methods

2.1. Retrieval of Studies and Selection Criteria. We systematically search available studies from 2003 to 2013 in PubMed (English), CNKI, and Wanfang (Chinese). Keywords were "coronary heart disease" or "coronary artery disease" or "myocardial infarction" combined with "*LDLR*" or "low density lipoprotein receptor" or "rs2228671" and "polymorphism" or "genetic association." The inclusion criteria for the studies involved in this meta-analysis met the following criteria: (1) case-control study about *LDLR* rs2228671 polymorphism; (2) case-control study with genotyping or allelic information, or odd ratio (OR) with 95% confidential interval (CI).

2.2. Data Extraction. Data included in this meta-analysis was extracted independently from all studies using the same standard protocol by two reviewers (HY and YH). The inclusion criteria of our meta-analysis were as follows: first author's name, publication year, ethnicity, numbers of cases and controls, genotype distribution, and OR with 95% CI.

2.3. Patients and Controls. The study protocol was approved by the ethical committee of School of Medicine, Ningbo University. A total of 162 cases and 113 controls were recruited in this study from the Affiliated Hospital of Ningbo University. All the participants in this study have signed the informed consent forms. All the 275 individuals underwent coronary angiography and were categorized into CHD patients and non-CHD controls according to our previous descriptions [5, 19]. All the participants enrolled in this study were Han Chinese residing in or near Ningbo city. None of individuals in this study had congenital heart disease, cardiomyopathy and severe liver, or kidney disease.

2.4. SNP Genotyping. Genomic DNA was isolated from peripheral blood lymphocytes using standard phenol-chloroform method and then was stored in TE buffer. All DNA samples were amplified by polymerase chain reaction (PCR). PCR was denatured at 94°C for 15 s, followed by 45 cycles of denaturation at 94°C for 20 s, annealing for 30 s at 56°C, extension at 72°C for 1 min, and a final extension at 72°C for 3 min. DNA amplification and genotyping was performed on the SEQUENOM Mass-ARRAY iPLEX platform according to the manufacturer's instructions [5].

2.5. Statistical Analyses. Hardy-Weinberg equilibrium (HWE) was examined by the Arlequin program (version 3.5) [20]. The differences in the genotype and allele frequencies between cases and controls were analyzed by the CLUMP22 software with 10,000 Monte Carlo simulations [21]. Power analysis was performed by Power and Sample Size Calculation software [22]. Meta-analysis was made by REVMAN 5.0 (Cochrane Collaboration, Oxford, United Kingdom) and Strata 11.0 software (Strata Corporation, College Station, TX) [23, 24]. Publication bias was evaluated by Begg and Egger regression tests [25]. The combined ORs with 95% CI values were calculated by either fixed-effect or random-effect method [26]. A two-tailed P value of 0.05 or lower was defined to be statistically significant.

3. Results

We systematically searched in PubMed, CNKI, and Wanfang from 2003 to 2013, and selected a total of 57 literatures after removing the duplicated publications (Figure 1). According to the descriptions in the titles and abstracts, we excluded 26 irrelevant literatures, 6 literatures on other variants, and 12 literatures on other diseases. In addition, 1 literature without sufficient case-control genotyping data and 5 literatures

Table 1: Characteristics of the association studies between rs2228671 and CHD.

Author and year	Ethnic group	Genotype (CC/CT/TT)		P-allele
		Cases	Controls	
Ortlepp et al. (2003) [11]	German	937/216/10	972/255/22	0.0453
Krawczak et al. (2006) [12]	German	1755/379/19	1840/474/25	0.0184
Samani et al. (2007) [13]	German	781/93/1	1417/224/3	0.0281
Samani et al. (2007) [13]	British	1578/322/13	2332/569/34	0.0051
Schunkert et al. (2008) [14]	German	236/43/2	224/61/5	0.0343
Erdmann et al. (2009) [15]	German	282/64/3	671/164/15	0.3333
Martinelli et al. (2010) [16]	Italian	549/134/9	227/61/3	0.73
Our study (2013)	Chinese	157/4/1	111/2/0	0.485

Table 2: Genotype and allele frequency distributionof *LDLR* gene rs2228671 polymorphism in cases and controls*.

Gender	Group	CC/CT/TT	χ^2	P (df = 2)	C/T	χ^2	P (df = 1)	OR (95% CI)
All	Cases	157/4/1	0.86	1	318/6	0.87	0.49	0.47 (0.09–2.36)
	Controls	111/2/0			224/2			
Male	Cases	113/2/1	0.51	0.77	228/4	NA	NA	0.49 (0.05–4.41)
	Controls	58/1/0			117/1			
Female	Cases	44/2/0	0.53	0.76	90/2	NA	NA	0.42 (0.04–4.71)
	Controls	53/1/0			107/1			

*NA represents not analyzed; rs2228671 meets HWE in all groups ($P > 0.05$).

Table 3: Genotype and allele frequency distributionof *LDLR* gene rs2228671 polymorphism in European population.

Gender	Group	CC/CT/TT	χ^2	P (df = 2)	C/T	χ^2	P (df = 1)	OR (95% CI)
European population	Cases	6218/1251/57	20.59	<.0001	13687/1365	20.26	<.0001	1.180 (1.098–1.269)
	Controls	7685/1808/107			17178/2022			

without detailed SNP information were also removed. At last, 6 literatures [11–16] on 7 case-control studies were harvested in our meta-analysis (Table 1). Furthermore, we performed a case-control study in Han Chinese population, and it was later included in our meta-analysis.

Genotype distribution of rs2228671 in our case-control study met HWE for both CHD cases and non-CHD controls ($P > 0.05$), indicating that our case-control study had a well-characterized random sampling. Our case-control study suggested that *LDLR* rs2228671-T allele was rare in Chinese population (cases: 2%; controls: 1%), and this agrees with the frequency in HapMap Chinese Han in Beijing (CHB) population (0–2%). No significant difference in the genotype distribution between CHD cases and non-CHD controls are revealed in all samples ($P > 0.05$; Table 2) and in the subgroup analysis by gender ($P > 0.05$; Table 2). In summary, our case-control study showed that there was no association between *LDLR* rs2228671 and CHD in Chinese. However, significant association was found between *LDLR* rs2228671 and CHD in European population ($\chi^2 = 20.59$, $P < 0.0001$ by genotype; $\chi^2 = 20.26$; OR = 1.180, 95% CI = 1.098–1.269, $P < 0.0001$ by allele; Table 3). Using the fixed-effect method, our meta-analysis contained 7,588 CHD patients and 9,711 controls from German, British, Italian, and Chinese populations. As shown in Figure 2, significant association was observed between rs2228671 and CHD ($P = 0.0005$, OR = 0.83, and 95% CI = 0.75–0.92). In addition, no heterogeneity among the studies was included in this meta-analysis ($I^2 = 0\%$; Figure 2). Furthermore, no obvious visual evidence of publication bias in the meta-analysis was shown by funnel plot ($P > 0.05$; Figure 3).

4. Discussion

Aberrant LDLR level in blood can cause abnormal cholesterol metabolism [2]. As the main pathogenic gene of FH, *LDLR* gene is associated with multiple vascular diseases [15, 16, 27]. Polymorphisms of *LDLR* gene were associated with type 2 diabetes [28] and hypertension [29] that also related to CHD. Recently, a handful of *LDLR* polymorphisms have been studied in CHD, including those (rs14158, rs3826810, rs1433099, rs2738464, rs2738465, and rs2738466) in the 3′-untranslated region (3′-UTR) and rs2228671 in second exon [30–32]. SNPs in first intron (rs6511720) and 5′ flanking region (rs17248720) of *LDLR* gene were closely related to both LDL-C and CHD [33, 34]. Rs1433099 and rs2738466 in the 3′-UTR of *LDLR* were reported to be associated with baseline lipids in American population [32]. The T allele of rs2228671 polymorphism was associated with higher FVIII:c levels. In addition, LDLR rs2228671 may be regulated FVIII:c levels

Study or subgroup	Case		Control			Odds ratio	Year	Odds ratio
	Events	Total	Events	Total	Weight	M-H, fxed, 95% CI		M-H, fxed, 95% CI
Ortlepp et al. (2003)	118	1163	150	1249	16.5%	0.83 [0.64, 1.07]	2003	
Krawczak et al. (2006)	209	2153	262	2339	28.8%	0.85 [0.70, 1.03]	2006	
Samani et al. (2007)	48	875	115	1644	9.6%	0.77 [0.55, 1.09]	2007	
Samani et al. (2007)	174	1913	319	2935	29.0%	0.82 [0.68, 1.00]	2007	
Schunkert et al. (2008)	24	281	36	290	4.1%	0.66 [0.38, 1.14]	2008	
Erdmann et al. (2009)	35	349	97	850	6.4%	0.87 [0.58, 1.30]	2009	
Martinelli et al. (2010)	76	692	34	291	5.4%	0.93 [0.61, 1.43]	2010	
Our study (2013)	3	162	1	113	0.1%	2.11 [0.22, 20.58]	2013	
Total (95% CI)		**7588**		**9711**	**100.0%**	**0.83 [0.75, 0.92]**		
Total events	687		1014					
Heterogeneity: $\chi^2 = 1.92$, df = 7 ($P = 0.96$); $I^2 = 0\%$								
Test for overall effect: $Z = 3.50$ ($P = 0.0005$)								

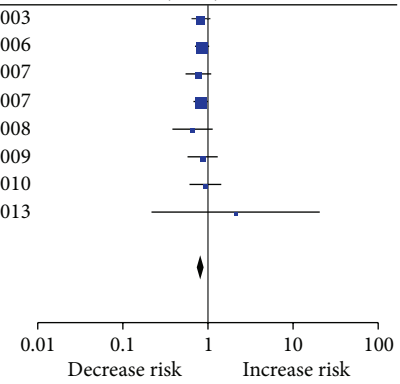


Figure 2: Meta-analysis of rs2228671 with CHD.

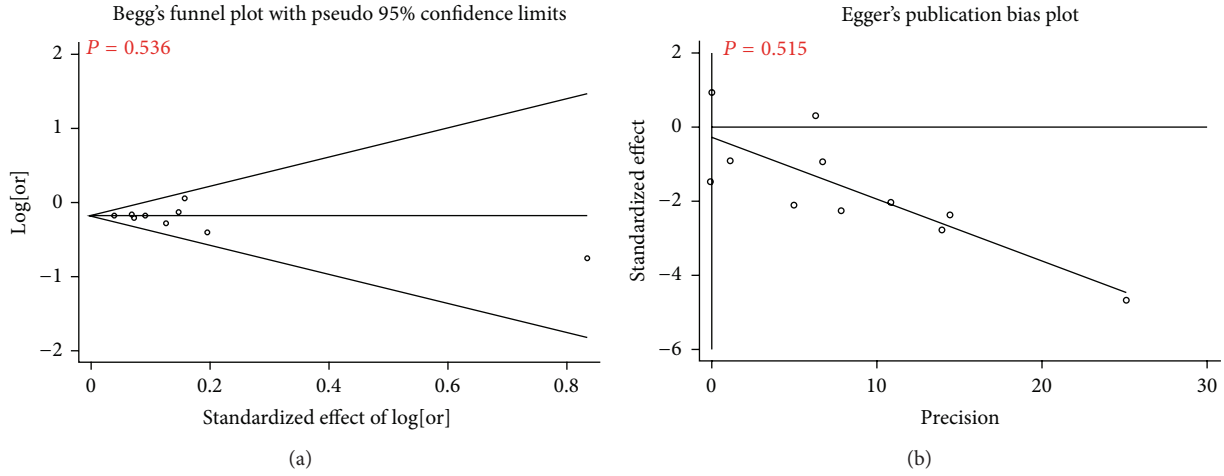


Figure 3: Publication bias analysis of 8 studies in the meta-analysis. The Begg's funnel plot and the Egger's publication bias plot test also indicated little evidence of publication bias among studies of rs2228671 and CHD risk (Begg, $P = 0.536$; Egger, $P = 0.515$).

and associated with the independence risk factor (plasma lipids) of CHD [16].

Our meta-analysis among 17299 individuals showed that rs2228671-T allele reduced the risk of coronary heart disease in the combined samples from German, British, Italian, and Chinese populations (OR = 0.83, $P = 0.0005$). Furthermore, rs2228671-T allele frequencies in the meta-analysis among German, British, and Italian populations were 7–12.2% that is similar to 10% in HapMap CEU population. However, rs2228671-T allele frequency is 0% in HapMap CHB population and 0.9% in the controls of our study. Due to the rare allele of *LDLR* rs2228671 in our samples, the power of our case-control study was only 5.1%, in contrast of 100% in the present meta-analysis. This suggests that a lack of association in our case-control study may largely be explained by the insufficient power for this rare polymorphism and the small sample size. Future investigation on other common *LDLR* polymorphisms is worth being performed in a large Chinese cohort.

Human *LDLR* is about 43 kb in length and has 1367 active polymorphism. As shown in our study, the allele frequency of rs1122608-T is much lower than those in the European studies; suggesting a cross-population comparison of this polymorphism may help one understand the role of *LDLR* in different ethnic population. Meanwhile, the previous tested *LDLR* rs1122608 polymorphism did not yield a significant result ($P = 0.148$) [35], in contrast to a significant result of rs2228671 in the current study ($P = 0.0005$). This suggests rs2228671 and rs1122608 might exert different contributions to the risk of CHD.

There were several limitations in our study as follows. Firstly, most of the involved individuals in our meta-analysis were Europeans; thus our result might not be applied to other populations such as Chinese. Secondly, although we had no evidence of the publication bias in our meta-analysis, we cannot exclude the possibility of existing potential bias upon reporting the studies without significant association results. Last but not least, the power of our case-control study in

Chinese is only 5.1%. The negative result of rs2228671 might not represent for other variants of *LDLR* gene in Chinese population.

In conclusion, the meta-analysis demonstrated that the *LDLR* rs2228671-T allele is a protective factor of CHD in Europeans. However, the case-control study showed no significant association of *LDLR* rs2228671 with CHD in Han Chinese population.

Conflict of Interests

The authors declare no conflict of interests.

Authors' Contribution

Huadan Ye, Qianlei Zhao and Yi Huang are co-first authors of this work.

Acknowledgments

The research was supported by the grants from National Natural Science Foundation of China (31100919 and 81371469), Natural Science Foundation of Zhejiang Province (LR13H020003), K. C. Wong Magna Fund in Ningbo University, Natural Science Foundation of Ningbo City (2011A610037), and Ningbo Social Development Research Projects (2012C50032).

References

[1] A. A. Phillips, A. T. Cote, S. S. Bredin, and D. E. Warburton, "Heart disease and left ventricular rotation—a systematic review and quantitative summary," *BMC Cardiovascular Disorders*, vol. 12, article 46, 2012.

[2] M. S. Brown and J. L. Goldstein, "Expression of the familial hypercholesterolemia gene in heterozygotes: mechanism for a dominant disorder in man," *Science*, vol. 185, no. 4145, pp. 61–63, 1974.

[3] I. Ejarque, J. T. Real, S. Martinez-Hervas et al., "Evaluation of clinical diagnosis criteria of familial ligand defective apoB 100 and lipoprotein phenotype comparison between LDL receptor gene mutations affecting ligand-binding domain and the R3500Q mutation of the apoB gene in patients from a South European population," *Translational Research*, vol. 151, no. 3, pp. 162–167, 2008.

[4] E. W. Lee, M. Michalkiewicz, J. Kitlinska et al., "Neuropeptide Y induces ischemic angiogenesis and restores function of ischemic skeletal muscles," *Journal of Clinical Investigation*, vol. 111, no. 12, pp. 1853–1862, 2003.

[5] J. Zhou, Y. Huang, R. S. Huang et al., "A case-control study provides evidence of association for a common SNP rs974819 in PDGFD to coronary heart disease and suggests a sex-dependent effect," *Thrombosis Research*, vol. 130, pp. 602–606, 2012.

[6] R. Bulbulia and J. Armitage, "LDL cholesterol targets—how low to go?" *Current Opinion in Lipidology*, vol. 23, pp. 265–270, 2012.

[7] R. Alonso, N. Mata, S. Castillo et al., "Cardiovascular disease in familial hypercholesterolaemia: Influence of low-density lipoprotein receptor mutation type and classic risk factors," *Atherosclerosis*, vol. 200, no. 2, pp. 315–321, 2008.

[8] S. Kathiresan, C. J. Willer, G. M. Peloso et al., "Common variants at 30 loci contribute to polygenic dyslipidemia," *Nature Genetics*, vol. 41, no. 1, pp. 56–65, 2009.

[9] S. Kathiresan, B. F. Voight, S. Purcell et al., "Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants," *Nature Genetics*, vol. 41, no. 3, pp. 334–341, 2009.

[10] P. Linsel-Nitschke, A. Götz, J. Erdmann et al., "Lifelong reduction of LDL-cholesterol related to a common varriant in the LDL-receptor gene decreases the risk of coronary artery disease—a Mendelian randomisation study," *PLoS ONE*, vol. 3, no. 8, Article ID e2986, 2008.

[11] J. R. Ortlepp, A. Von Korff, P. Hanrath, K. Zerres, and R. Hoffmann, "Vitamin D receptor gene polymorphism BsmI is not associated with the prevalence and severity of CAD in a large-scale angiographic cohort of 3441 patients," *European Journal of Clinical Investigation*, vol. 33, no. 2, pp. 106–109, 2003.

[12] M. Krawczak, S. Nikolaus, H. Von Eberstein, P. J. Croucher, N. E. El Mokhtari, and S. Schreiber, "PopGen: Population-based recruitment of patients and controls for the analysis of complex genotype-phenotype relationships," *Community Genetics*, vol. 9, no. 1, pp. 55–61, 2006.

[13] N. J. Samani, J. Erdmann, A. S. Hall et al., "Genomewide association analysis of coronary artery disease," *New England Journal of Medicine*, vol. 357, no. 5, pp. 443–453, 2007.

[14] H. Schunkert, A. Götz, P. Braund et al., "Repeated replication and a prospective meta-analysis of the association between chromosome 9p21.3 and coronary artery disease," *Circulation*, vol. 117, no. 13, pp. 1675–1684, 2008.

[15] J. Erdmann, A. Großhennig, P. S. Braund et al., "New susceptibility locus for coronary artery disease on chromosome 3q22.3," *Nature Genetics*, vol. 41, no. 3, pp. 280–282, 2009.

[16] N. Martinelli, D. Girelli, B. Lunghi et al., "Polymorphisms at LDLR locus may be associated with coronary artery disease through modulation of coagulation factor VIII activity and independently from lipid profile," *Blood*, vol. 116, no. 25, pp. 5688–5697, 2010.

[17] S. Cassese, A. de Waha, G. Ndrepepa et al., "Intra-aortic balloon counterpulsation in patients with acute myocardial infarction without cardiogenic shock. A meta-analysis of randomized trials," *American Heart Journal*, vol. 164, article e51, pp. 58–65, 2012.

[18] C. Wang, T. Sun, H. Li, J. Bai, and Y. Li, "Lipoprotein lipase Ser447Ter polymorphism associated with the risk of ischemic stroke: a meta-analysis," *Thrombosis Research*, vol. 128, no. 5, pp. e107–e112, 2011.

[19] E. Rapaport, R. Bernard, and E. Corday, "Nomenclature and criteria for diagnosis of ischemic heart disease. Report of the Joint International Society and Federation of Cardiology/World Health Organization Task Force on standardization of clinical nomenclature," *Circulation*, vol. 59, no. 3, pp. 607–609, 1979.

[20] L. Excoffier and H. E. L. Lischer, "Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows," *Molecular Ecology Resources*, vol. 10, no. 3, pp. 564–567, 2010.

[21] P. C. Sham and D. Curtis, "Monte Carlo tests for associations between disease and alleles at highly polymorphic loci," *Annals of Human Genetics*, vol. 59, no. 1, pp. 97–105, 1995.

[22] W. D. Dupont and W. D. Plummer Jr., "Power and sample size calculations. A review and computer program," *Controlled Clinical Trials*, vol. 11, no. 2, pp. 116–128, 1990.

[23] D. F. Stroup, J. A. Berlin, S. C. Morton et al., "Meta-analysis of observational studies in epidemiology: a proposal for reporting," *Journal of the American Medical Association*, vol. 283, no. 15, pp. 2008–2012, 2000.

[24] J. W. Bisson and V. J. Cabelli, "Membrane filter enumeration method for Clostridium perfringens," *Applied and Environmental Microbiology*, vol. 37, no. 1, pp. 55–66, 1979.

[25] M. Egger, G. Davey Smith, M. Schneider, and C. Minder, "Bias in meta-analysis detected by a simple, graphical test," *British Medical Journal*, vol. 315, pp. 629–634, 1997.

[26] R. DerSimonian and N. Laird, "Meta-analysis in clinical trials," *Controlled Clinical Trials*, vol. 7, no. 3, pp. 177–188, 1986.

[27] S. Kathiresan, O. Melander, C. Guiducci et al., "Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans," *Nature Genetics*, vol. 40, no. 11, pp. 189–197, 2008.

[28] S. H. Wu, Y. Q. Wang, and D. Q. Sun, "The association of HincII/low density lipoprotein receptor (LDLR) restriction fragment length polymorphism (RFLP) with diabetes mellitus and its lipid phenotype with PCR gene amplification," *Zhonghua Yi Xue Za Zhi*, vol. 73, no. 1, pp. 10–60, 1993.

[29] Y. Yamada, K. Kato, T. Yoshida et al., "Association of polymorphisms of ABCA1 and ROS1 with hypertension in Japanese individuals," *International Journal of Molecular Medicine*, vol. 21, no. 1, pp. 83–89, 2008.

[30] J. M. Murabito, C. C. White, M. Kavousi et al., "Association between chromosome 9p21 variants and the ankle-brachial index identified by a meta-analysis of 21 genome-wide association studies," *Circulation Cardiovascular Genetics*, vol. 5, pp. 100–112, 2012.

[31] W. Chen, S. Wang, Y. Ma et al., "Analysis of polymorphisms in the 3′ untranslated region of the LDL receptor gene and their effect on plasma cholesterol levels and drug response," *International Journal of Molecular Medicine*, vol. 21, no. 3, pp. 345–353, 2008.

[32] E. Polisecki, H. Muallem, N. Maeda et al., "Genetic variation at the LDL receptor and HMG-CoA reductase gene loci, lipid levels, statin response, and cardiovascular disease incidence in PROSPER," *Atherosclerosis*, vol. 200, no. 1, pp. 109–114, 2008.

[33] F. Takeuchi, M. Isono, T. Katsuya et al., "Association of genetic variants influencing lipid levels with coronary artery disease in Japanese individuals," *PLoS ONE*, vol. 7, Article ID e46385, 2012.

[34] P. Jeemon, K. Pettigrew, C. Sainsbury, D. Prabhakaran, and S. Padmanabhan, "Implications of discoveries from genome-wide association studies in current cardiovascular practice," *World Journal of Cardiology*, vol. 3, pp. 230–247, 2011.

[35] L. Zhang, F. Yuan, P. Liu et al., "Association between PCSK9 and LDLR gene polymorphisms with coronary heart disease: case-control study and meta-analysis," *Clinical Biochemistry*, vol. 46, pp. 727–732, 2013.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 745298, 11 pages
http://dx.doi.org/10.1155/2014/745298

Research Article

Using the Sadakane Compressed Suffix Tree to Solve the All-Pairs Suffix-Prefix Problem

Maan Haj Rachid,[1] Qutaibah Malluhi,[1] and Mohamed Abouelhoda[2,3]

[1] *KINDI Lab for Computing Research, Qatar University P.O. Box 2713, Doha, Qatar*
[2] *Faculty of Engineering, Cairo University, Giza, Egypt*
[3] *Center for Informatics Sciences, Nile University, Giza, Egypt*

Correspondence should be addressed to Qutaibah Malluhi; qmalluhi@qu.edu.qa

Received 16 January 2014; Accepted 10 March 2014; Published 16 April 2014

Academic Editor: Ryuji Hamamoto

The all-pairs suffix-prefix matching problem is a basic problem in string processing. It has an application in the de novo genome assembly task, which is one of the major bioinformatics problems. Due to the large size of the input data, it is crucial to use fast and space efficient solutions. In this paper, we present a space-economical solution to this problem using the generalized Sadakane compressed suffix tree. Furthermore, we present a parallel algorithm to provide more speed for shared memory computers. Our sequential and parallel algorithms are optimized by exploiting features of the Sadakane compressed index data structure. Experimental results show that our solution based on the Sadakane's compressed index consumes significantly less space than the ones based on noncompressed data structures like the suffix tree and the enhanced suffix array. Our experimental results show that our parallel algorithm is efficient and scales well with increasing number of processors.

1. Introduction

Given a set S of strings $S_1, S_2, \ldots, S_k$, the all-pairs suffix-prefix problem (APSP) is to find the longest suffix-prefix match for each ordered pair of the set S. Solving this problem is a basic step in the de novo genome assembly task, where the input is a set of strings representing random fragments coming from multiple copies of the input genome. These fragments can be ordered based on suffix-prefix matching and after some postprocessing, the input genome can be reconstructed.

With the recent advances in high throughput genome sequencing technologies, the input size became very huge in terms of the number of sequences and length of fragments. This calls for both faster and memory efficient solutions for the APSP problem.

The APSP is a well-studied problem in the field of string processing. The first nonquadratic solution was introduced by Gusfield et al. [1]. Their algorithm was based on the generalized suffix tree and it takes $O(n + k^2)$ time and linear space, where n is the total length of all k strings. Although the theoretical bounds of this algorithm are optimal, the cache performance and space consumption of the suffix tree are major bottlenecks to solve large size problems (note that the best implementation of a suffix tree consumes 20 bytes per input character [2]).

Ohlebusch and Gog [3] introduced a solution to APSP using the enhanced suffix array [4], which is an index data structure that uses only 8 bytes per input character. Their algorithm has the same complexity as that of [1]. Their algorithm has exploited interesting features of the enhanced suffix array, which has not only reduced the space consumption but also improved the cache performance and accordingly the running time. Experimental results have shown that their solution is 1.5 to 2 times faster in practice and can indeed handle large problem sizes.

In an effort to reduce the space consumption of solving the problem, Simpson and Durbin [5] used the FM index [6] to solve the problem in an indirect way as follows. The index is constructed for all strings after concatenating them in one string. The index is then queried by the reads, one by one, to find prefix-suffix matches. The time complexity of this algorithm is not as optimal as the one of [1, 3], because one

examines more suffixes than the output size. (This limitation stems also from the fact that the FM index lacks structural information to run the algorithms of [1] or [3] on it.) But its space consumption is much less than that of the previous algorithms.

In this paper, we present new methods based on the compressed suffix tree of Sadakane [7] and variations of the algorithms of [1] and [3]. The compressed suffix tree is considered as a self-index data structure, because the original text is already encoded in the index in a compressed fashion and can be extracted from it; that is, there is no need to keep the original text in the memory. It is also fully functional like the uncompressed suffix tree, as it offers the typical suffix tree operations such as checking if a node is a leaf, moving to the next sibling, using a suffix link, or even performing lowest common ancestor queries. Such compressed suffix tree consumes much less space than the suffix tree and the enhanced suffix array but more space than the FM index [6], as it includes additional structural information.

To further speed up the solution of APSP, we introduce different parallelization strategies to the sequential algorithm that can be used on multiprocessor shared memory computers. Our parallelization methods exploit important features and available operations of the Sadakane's compressed suffix tree. Experimental results show that our method is efficient and scales well with the number of processors.

This paper is organized as follows. In Section 2, the data structures and the functions used in our solutions are explained. In Section 4, the two different approaches for solving APSP using Sadakane index are demonstrated. Section 5 describes how our solutions can be parallelized, and finally we show our experimental results and conclusions in Sections 6 and 7, respectively.

2. Overview

2.1. Basic Notions. We write $S = S_1, S_2, \ldots, S_k$ to denote a set S of k strings. Each string S_i is defined over an ordered alphabet Σ. For strings representing genomic sequences, $\Sigma = \{A, C, G, T\}$, which is the standard alphabet for DNA sequence data. We write $|S_i|$ to denote the length of the string S_i and $S_i[j]$ to denote the jth character, where $1 \leq j \leq |S_i|$. The jth suffix of a string S_i is the substring $S_i[j..|S_j|]$ and it is denoted by $S_i(j)$. A prefix of length j of a string S_i is the substring $S_i[1..j]$. For two strings S_i and S_j, the longest suffix-prefix match of the pair (S_i, S_j) is the match with the greatest r such that $S_i[r..|S_i|] = S_j[1..r]$.

The suffix tree of a string S is an index structure in which each suffix of S is stored as a path from the root to a leaf. Obviously many suffixes will share partial path before they end in different leaves. Accordingly, the suffix tree of a string S has n leaves and at most $n-1$ internal nodes, where $n = |S|$. Suffix tree can be constructed and stored in linear time and space ([8, 9]).

2.2. Compressed Suffix Tree. Sadakane's compressed suffix tree [7] is composed of three major components.

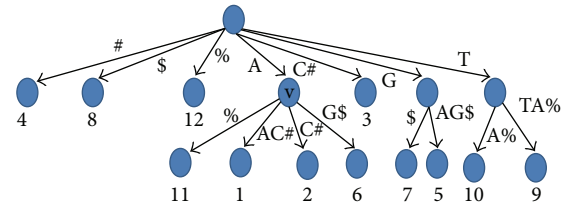


Figure 1: A generalized suffix tree for the string AAC#GAG$TTA%. The numbers below the leaves are the text positions for the string paths which are represented by these leaves.

(i) *Compressed suffix array* (CSA): the suffix array SA of a string S is an array of integers including the positions of the lexicographically sorted suffixes of S; that is, for any two integers $0 \leq i' < i'' < n$, $S(\mathrm{SA}[i'])$ is lexicographically less than $S(\mathrm{SA}[i''])$. The CSA is a compressed version of SA with reduced space requirements, which can perform the traditional suffix array operations with a slight slowdown. The implementation of Sadakane suffix tree that is used in our experiments utilizes the CSA presented in [10]. It is based on wavelet tree [11] built on Burrows-Wheeler transform [12]. The space consumption of this CSA is $O(n \log \Sigma)$.

(ii) The *largest common prefix array* (LCP): this is an array of integers in the range 1 to n such that LCP [1] = 0 and LCP[i] is the largest common prefix between SA[i] and SA[$i-1$], where $1 \leq i \leq n$. The LCP is also compressed. In the implementation which we use, LCP is encoded using a technique described by [7] which can store LCP in only $2n + O(n)$ bits.

(iii) The *balanced parenthesis representation* (BP): BP of a tree is generated by traversing the tree in preorder manner to produce open and closed parentheses. Initially, BP is empty. Whenever a node is visited, an open parenthesis, (, is added to BP. Whenever a node is left, a close parenthesis,), is added to BP [13]. Accordingly, each node can be encoded using 2 bits. Since suffix tree has at most $n-1$ internal nodes and n leaves, BP takes at most $4n$ bits. For example, the BP representation of the tree in Figure 1 is (()()()(()()()())()(()())(()())).

The following BP functions are used in this paper.

(i) $\mathrm{Rank}_{()}(\mathrm{BP}, i)$: returns the number of occurrences of () in BP up to position i.

(ii) $\mathrm{Select}_{()}(\mathrm{BP}, i)$: returns the position of the ith () in BP.

(iii) IsLeaf (i): returns true if the position i in BP belongs to a leaf.

(iv) Parent (i): returns the position in BP for the parent node of node v, where i is the BP position of v.

(v) IsOpen (i): returns true if i is a position for an open parenthesis in BP.

(vi) Edge (i, d): returns the dth character of the edge label of an edge pointing to node v, where i is the BP position of v.

(vii) Child (v, c): returns the position in BP for the node i which is the child of a node p, where v is p's position in BP, and there is an edge directed from p towards i labeled by a string that starts with character c.

2.3. Constructing Generalized Suffix Tree. To solve the all-pairs suffix-prefix problem for a set of k strings $S_1, S_2, S_3, \ldots, S_k$, we build a compressed suffix tree for the string resulting by concatenating all strings together in one string. Each two concatenated strings are separated by a distinct separator. These separators do not occur in any of these k strings. For example, if the strings are $S_1 = AAC$, $S_2 = GAG$, $S_3 = TTA$, then we build a compressed suffix tree for the text AAC#GAG$TTA%, where #, $, and % are the distinct separators. These separators should be lexicographically smaller than any character in all strings (i.e., in the alphabet Σ). Since, in practice, there is a limitation for the number of available distinct separators, our implementation uses more than one character to encode a separator. Assuming that there are c distinct characters that can be used for constructing separators, $\log_c k + 1$ characters are needed to encode a separator in our work.

We use an array *StartPos* of size k to store the starting positions of the k strings. The size of such array is $\Omega(k \log n)$ bits, where n is the size of the whole text. To map each position to the appropriate string, another array of size n is needed. This array requires space of $\Omega(n \log k)$ bits. To avoid the expensive cost of this array, a binary search in the array *StartPos* can be done to retrieve the number of the string to which a specific position belongs. However, that will increase the time cost of retrieving the string identifier to $O(\log k)$ instead of $O(1)$ time. It is easy to notice that both arrays are not necessary if the k strings are equal in size. In this case, we can get the string number to which a position p belongs by simply calculating p/l, where l is the length of each string.

3. Review of the Basic APSP Algorithm

The algorithm of [1] works as follows. First, the suffix tree is constructed for the string $\widehat{S} = S_1 \#_1, S_2, \ldots, \#_{k-1}, S_k \#_k$. The characters $\#_1, \ldots, \#_k$ are distinct and do not exist in any of the given strings. These distinct characters are further referred to as *terminal characters* in this paper. Second, the suffix tree is traversed to create for each internal node v a list L_v. The list L_v contains the children of v such that each child c is a leaf, and the label of the edge connecting v to c starts with a terminal character. Third, the suffix tree is traversed in a preorder fashion once again to report matches according to the following observation. Consider a leaf such that the string annotating the edges from the root to it is a complete given string S_j. We call such leaf a *prefix leaf*. For each node v_r on the path from the root to the prefix leaf, the prefix-suffix matches of length $|v_r|$ are those between each element in L_{v_r} and S_j. Accordingly, in the preorder traversal, we use k stacks representing the given strings and push v_r in stack i if S_i is in L_{v_r}. When reaching a prefix leaf S_j, the candidates from all parent nodes would already be in the stacks and the maximal matches are those between S_j and the top of each stack.

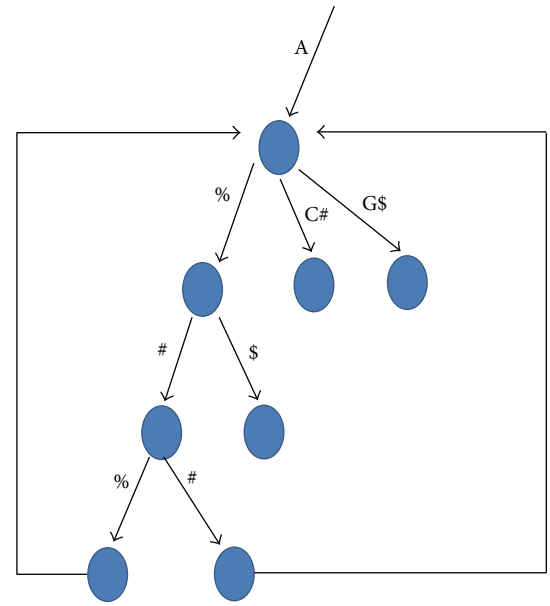


Figure 2: The text position for each leaf which has a terminal edge will be added to the L list of the closest ancestor which does not have a terminal edge pointing to it.

4. Solutions Based on the Compressed Suffix Tree

4.1. First Method. The compressed suffix tree supports all necessary informations to run the original Algorithm of [1] as it is. However, we observe some interesting properties that could significantly improve the performance of the algorithm with no additional time or space costs.

For filling the L_v lists, we will not simulate traversal of the whole tree over the compressed suffix tree. Rather, we will make use of the BP vector to move from a leaf to another using the *Select* function in constant time. Specifically, the $Select_{()}(BP, i)$ function will return the position of the ith () which represents a leaf. For each leaf and only if it has a terminal edge pointing to it (which can be checked using edge function), we add the text position of that leaf to the L_v list of the parent of that leaf node. In Figure 1, we give an example, where the L list for node v, which is the fourth child of the root, has one value 11 that belongs to string 3. Note that we can safely ignore the first k-leaves as these correspond to the terminal suffixes, where the length of each of these suffixes is a single character (one of the terminal characters).

In the case of using more than one character to encode a distinct separator, it is possible to have an internal node to which a terminal edge is pointing (usually only leaves have this possibility). Accordingly, the text position of a terminal leaf should be added to the L list of its closest ancestor to which no terminal edge is pointing (see Figure 2). Let i be a BP position of leaf v and j is the BP position of v's parent, node z. The pseudocode for the bottom-up traversal:

While z is not the root and the edge pointing to z is terminal,

$$j = \text{parent}(j). \quad (1)$$

```
(1) i = The position in BP of the first child of the root (ignoring children which belong to the distinct separators)
(2) g ← rank_()(i)
(3) g ← g + 1
(4) while i < The position in BP of the rightmostleaf in the tree do
(5)     i ← select_()(g)
(6)     if The node with the position i in BP has a terminal edge then
(7)         Add the text position of i to L_v where v is the parent of the node which has a position i in BP
(8)     end if
(9)     g ← g + 1
(10) end while
(11) firstchild = The position of the first child of the root in BP
(12) for i = firstchild To rightmostleaf of the tree do
(13)     if isLeaf(i) and (i is a Starting Position in a string f) then
(14)         for j = 1 to k do
(15)             if j ≠ f then
(16)                 Sol[j, f] = top(stack(j))
(17)             end if
(18)         end for
(19)         Increment i by 1 to avoid the closing parenthesis
(20)     else if i is an opening parenthesis of an internal node v then
(21)         for j = 1 to size of L_v do
(22)             Push L_v(j) to the corresponding stack
(23)         end for
(24)     else if i is a closing parenthesis of an internal node v then
(25)         for j = 1 to size of L_v do
(26)             Pop L_v(j) from the corresponding stack
(27)         end for
(28)     end if
(29) end for
```

Algorithm 1: First method.

In the second stage, we make another scan for the BP representation from left to right, but this time we move one by one (parenthesis by parenthesis) instead of jumping from leaf to leaf. We distinguish 3 cases.

(i) Case 1: if the scanned node is a leaf and it is representing a starting position of a string i, then the top of each stack j, where $j \neq i$ and $1 \leq j \leq k$, is the longest suffix prefix match between string i and string j (for a proof, see [1]). We can move one step ahead since the next parenthesis is the closing parenthesis of this leaf node (lines 13–19, Algorithm 1).

(ii) Case 2: if we scan an opening parenthesis for an internal node v, we push each value in the list L of that internal node v to the appropriate stack (which can be found in $\log k$ time). We can determine which stack we should push the value to since this value is a text position. In Figure 1, the value 11 in L_v will be pushed to stack 3 (lines 20–23, Algorithm 1).

(iii) Case 3: if we scan a closing parenthesis for an internal node v, we pop all values that belong to v from the stacks. We can easily determine which stacks to pop using L_v (lines 24–28, Algorithm 1).

Algorithm 1 specifies our method based on the compressed suffix tree. Lines 4–10 in Algorithm 1 compute the L_v lists as described above. We use k stacks to keep track of the leaves. The second loop (lines 11–28 in Algorithm 1) mimics a preorder traversal. All ancestors of any leaf will be visited before the leaf itself, which will guarantee that all stacks for the k strings will be filled before checking any leaf with a starting position. When a leaf with a starting position of a string S_j is reached, the top of each stack i will represent the longest suffix prefix match between string i and string j. Finally the closing parenthesis for any internal node will be reached after reaching all leaves in all subtrees of that internal node which guarantees the appropriate update (pop up) to all stacks. The two-dimensional array, Sol, will carry the solution at the end of the second loop.

4.2. Complexity Analysis. The correctness of the algorithm follows from the proof in [1]. However, in our implementation, we start the first loop with the gth leaf. Since g is incremented, we are moving from leaf to leaf until we reach the rightmost leaf. It is clear that all L_v lists for all internal nodes will be filled at the end of the loop.

The construction of the generalized suffix tree consumes $O(n \log n)$ time [14]. We have n leaves so we need $O(n)$ time in the first loop. The second loop requires $3n$ steps since we have 2 parentheses for each leaf and 2 parentheses for each internal node, but we are avoiding the closing parenthesis of any leaf node by incrementing the counter by 1. In the second loop,

we will have at most one push and one pop for each leaf so we have $O(n)$ time complexity since all index operations which we are using (like isLeaf, Child, and Parent) have constant time [7]. The string to which the value on the top of a stack belongs is known since it is equal to the number of the stack, accordingly the time for outputting the results is k^2.

As a result, the solution requires $O(n \log n + k^2)$ time. The complexity stands even without the usage of an array to map a position to a string (this can be done by using binary search in *StartPos* array), since $n \log k$ is less than $n \log n$.

In term of space, we need $|CSA| + 6n + O(n)$ bits to construct the tree, where $|CSA|$ is the size of the compressed suffix array [7]. Since the total number of all values in all L lists is $O(n)$, we need $O(n \log n)$ bits for these lists and for the stacks. The two arrays which are mentioned in the end of Section 2 require $O(k \log n)$ and $O(n \log k)$ which are both less than $O(n \log n)$. Accordingly, the solution consumes $O(n \log n)$ space.

4.3. Further Space Optimization

4.3.1. Space Optimization 1. It is clear the L_v lists which are used in this method are very expensive in terms of space. One way to avoid using them is to scan the leaves once. For each leaf e and only if it represents a complete string S, we check every ancestor using the parent function until the root is reached. For each ancestor, we check every terminal edge which is coming from it. A terminal edge indicates a match between a prefix of S and a suffix in another string. Accordingly L_v lists are avoided and so are the k stacks. Let ℓ be the maximum length of all paths from the root to all leaves (which is usually less than 1500, the maximum length for a sequence). There are at most k terminal edge for each internal node, thereby the time consumption will be $O(n \log n + \ell k^2)$.

4.3.2. Space Optimization 2. Another variation for the first method is to keep the first stage as is, but in the second scan we check only the leaves. In this variation we will use the L_v lists but we will not use the stacks. If a leaf represents a complete string S, we check every ancestor of this leaf. Since the L_v lists are filled from the first stage, the values inside the L_v lists of the current internal node are suffix-prefix matches between S and suffixes from other strings. The time complexity will be the same which is $O(n \log n + k^2)$.

4.4. Second Method. The running time of the previous method can be improved based on the following two observations of [3].

(i) All the distinct characters $\{\#_1, \#_2, \ldots, \#_k\}$ exist in the first k slots in the (compressed) suffix array, because they are lexicographically less than any other character in the given strings.

(ii) The terminal leaves (suffixes) sharing a prefix of length ω exist in the (compressed) suffix array before the other suffixes sharing also a prefix of length ω with them.

In this method, we scan the BP vector and move from leaf to leaf using the *Select* function. When a leaf is visited, we check if this leaf represents a suffix that is a prefix of the next leaf in BP. If it is, then it is pushed to the stack of the string which it belongs to. This continuous pushing is similar to creating the L_v lists and copying their values to the appropriate stacks. When a prefix leaf (i.e., corresponding to a whole given string) is scanned, then all pairwise prefix-suffix matches are already in the stack. An additional stack is used to keep track of the match length. Algorithm 2 specifies how this algorithm works.

As in the first method, we ignore parentheses which belong to separators using the Child, Rank, and Select functions. We move from leaf to leaf using the Select function (lines 1–3, Algorithm 2).

To check if a leaf i is a prefix of the next leaf q, we check if i is a terminal leaf, and it has the same parent as the next leaf j in BP. If this is the case, we push the text position of i to the stack of S_1, where S_1 is the string to which the text position of i belongs (lines 32–42, Algorithm 2).

If the text position of i is a starting position of a string S (which can be verified using a binary search in *StartPos* array), then the top of each stack j, where $j \neq S$ and $1 \leq j \leq k$, is the longest suffix prefix match between string S and string j (lines 7–11, Algorithm 2).

There is one exception for that; if there is a suffix in j which matches the string S and follows lexicographically the current suffix. This condition can be checked by investigating if the i and j both are terminal leaves, and they have the same parent. (lines 12–20, Algorithm 2).

The definition of the same parent depends on the number of characters used to encode the separators; if more than one character is used, then the parent of a leaf is the closest ancestor which does not have a terminal edge (Figure 2).

The second method has the same time complexity as the first method, since the construction of the tree requires $O(n \log n)$ time. For space complexity, let ℓ denote the maximum length of a sequence. We need at most ℓ of L lists to hold at most n values. Accordingly, $n \log k$ bits are needed for all L lists. We also need $n \log \ell$ bits to store at most n values in the k stacks. Since ℓ is less than k, the space complexity for the second method is $O(n \log k)$, regardless of the usage of the array to map a position to a string.

5. Parallelizing the Algorithm

In this section, we introduce parallel versions of the above-described methods for solving the APSP problem. These versions are for shared memory multiprocessor computers. We will handle two parallelization strategies: The first, which we will call *top-down decomposition* is based on a straightforward top-down tree decomposition. The second, which we call *leaf-decomposition* is based on bottom-up decomposition.

5.1. Strategy 1: Top-Down Decomposition. The generalized suffix tree is divided into P' subtrees occupying the highest levels of the tree. These subtrees can be processed independently in parallel. For P processors, we choose $P' = \gamma P$, where

```
(1) i = The position of the first child of the root in BP (ignoring children which belong to the distinct separators)
(2) g ← rank_()(i)
(3) g ← g + 1
(4) leafnum ← k
(5) for i = (BP position of the leaf with rank g) To (BP position of the rightmost leaf) do
(6)    if the text position of i is a starting position of the string f then
(7)       for j = 1 to k do
(8)          if j ≠ f then
(9)             Sol[j, f] = top(stack(j))
(10)         end if
(11)      end for
(12)      if i is less than BP position of the rightmost leaf then
(13)         q is BP position of the node which is next to the one indicated by i in BP
(14)         while i and q have a terminal edge, and the same parent do
(15)            S_1 is the string to which the text position of i belongs
(16)            S_2 is the string to which the text position of q belongs
(17)            Sol[S_1][S_2] = The ending position of S_2 − the text position of q
(18)            q ← q + 2
(19)         end while
(20)      end if
(21)   end if
(22)   if i is less than BP position of the rightmost leaf then
(23)      if LCP(leafnum + 1) < LCP(leafnum) then
(24)         while top(lcp_stack) > LCP(leafnum + 1) do
(25)            TopLcpStack is the top of lcp_stack
(26)            for each element j in the list l[TopLcpStack] do
(27)               Pop the stack[j])
(28)               Pop l[TopLcpStack]
(29)            end for
(30)            Pop(lcp_stack)
(31)         end while
(32)      else if i has a terminal edge then
(33)         q is BP position of the node which is next to the one indicated by i in BP
(34)         if q and i have the same parent then
(35)            S_1 is the string to which the text position of i belongs
(36)            Push(Stack[S_1], Ending position of S_1 − Text position of i)
(37)            Push(l[LCP(leafnum + 1)], S_1)
(38)            if LCP(leafnum + 1) ≠ top(lcp_stack) then
(39)               Push(lcp_stack, LCP(leafnum + 1))
(40)            end if
(41)         end if
(42)      end if
(43)   end if
(44)   leafnum ← leafnum + 1
(45)   g ← g + 1
(46) end for
```

Algorithm 2: Second method.

γ is a user defined parameter (We usually set it to 1.5). The roots of these subtrees are maintained in a queue. Whenever a processor is free, then one subtree is assigned to it. Each processor executes either Algorithm 1 or Algorithm 2. The P' subtrees are selected by breadth-first traversal of the tree. Over the BP representation, these are selected using the child function.

For Algorithm 1, we should consider the following. Let ω_r, where $r \in [1..P']$, denote the string annotating the edges from the root of the generalized suffix tree to the root of rth subtree. Let $\ell' = \max|\omega_r|$ is the length of the longest ω_r strings. Here we distinguish between two cases: (1) the minimum match length ℓ is larger than ℓ' or (2) ℓ is less than ℓ'.

For the first case, the subtrees can be easily processed independently in parallel. The L_v lists on the nodes from the root of the generalized tree to the roots of the subtrees need not to be created as the respective nodes will not be processed. A processor can start executing on a subtree without filling the stacks with the values related to its ancestors.

For the second case, we will have some L_v lists that can be shared among two processors. For reporting the matches, there is no problem as the L_v lists are read only. For creating

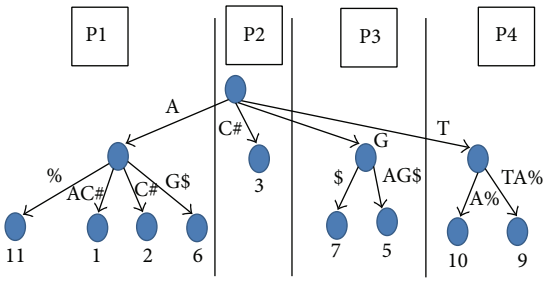


Figure 3: Each processor is working on one branch of the generalized suffix tree for the string AAC#GAG$TTA%. The numbers below the leaves are the text positions for the string paths which are represented by these leaves.

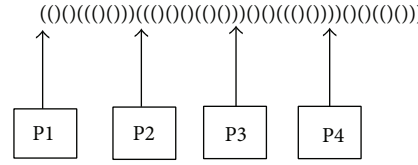


Figure 4: Each processor is working on its share of BP or the leaves. The stacks should be filled first for each processor before continuing with the algorithm.

them, however, we need to use only $P'' < P'$ processers, where P'' is the number of L_v lists to be created. The stacks should be filled first with the values related to the subtree's ancestors before executing the algorithm. In our second algorithm based on [3], the L_v lists are not created and accordingly the above-two cases can be ignored.

In Figure 3, we give a simple example where only subtrees from the top level are pushed to the queue. Assuming that 4 processors are utilized for the problem, processor 1 will work on the first child (we ignored the children which belong to #, $, and %). Processor 1 will find the answers for string 1 which is starting with an "A," while Processor 3 which is working on the third child of the root will find the answers for string 2 which is starting with a "G." Processor 4 which is working on the fourth child of the root will find the answers for string 3 which is starting with a "T." Processor 2 is not going to find any answer since none of the k strings start with a "C." No communication is required between processors for execution.

5.2. Strategy 2: Bottom-Up Decomposition. In the previous algorithm, we cannot guarantee that the subtrees are of equal sizes. Therefore, we use two tricks. First, we select γP subtrees, in hope of having trees of almost equal size. Second, we used a queue to keep all processors as busy as possible, which is a kind of dynamic load balancing.

Interestingly, the structure of CSA allows more robust strategy which can lead to better performance. The idea is to distribute the load equally between processors either by dividing the leaves or by dividing BP between them. Each processor starts working from the starting point of its share. It is clear that the situation is not simple; therefore, let us analyze the content of the stack for an internal node in the sequential case when the algorithm reaches that node. It can be observed that the content of each stack is whatever was pushed when visiting the node's ancestors. All other pushing work is irrelevant since it is followed by an equivalent popping before reaching the node.

Therefore, each processor can start from a specific point if its stacks are filled with the values which would be in the stacks if we reach this point while running the sequential algorithm.

To apply this concept on the first Algorithm, let us analyze the two stages for this algorithm. The first stage is relatively trivial; each leaf, if it has a terminal edge, should push its text position to the L list of its parent (or to the L list of the closest ancestor which does not have a terminal edge pointing to it). Accordingly, if leaves are distributed between processors, we will have a relatively fair deal between processors.

In the second stage, BP vector will be divided equally between processors. Let i be the starting parenthesis for the processor p's share in BP (if the starting parenthesis is closing parenthesis, i is the first open parenthesis which comes after the starting parenthesis). The stacks of the processor p should be filled with whatever values that would be pushed when passing through the ancestors of i if we were working with the sequential algorithm. The parent function is recursively called for i until the root is reached. For each ancestor of i, we scan the children leaves which belongs to separators and push them in first-in-first-out way into the stacks. Each processor can then execute the algorithm on its share as if the case is sequential until the ending point of the processor's share is reached. Figure 4 demonstrates the concept of this technique.

In the second algorithm, the n leaves are divided between processors using Rank and Select. Let e be the starting leaf for the processor p's share. Again, the Parent function is recursively called until the root is reached. For each ancestor of i, we scan the children leaves which belong to separators and push them in first-in-first-out way into the stacks. The algorithm then can be executed exactly as the sequential case.

5.3. Managing the Space Overhead. It is clear that both techniques use k stacks for each processer, which may appear as a problem when a large number of processors are utilized. The space issue can be solved by implementing the k stacks using an efficient data structure such as balanced binary search tree instead of using an array of k stacks. Another solution is to use the technique presented in Section 4.3, which avoids using the k stacks.

6. Experimental Results

A summary for the discussed algorithms is shown in Table 1. Experiments have been conducted to show the gain in space by comparing the space consumed by Sadakane compressed suffix tree with the space consumed by a standard pointer-based suffix tree and enhanced suffix array. We also investigated the space and time consumed in the overlap stage of a recent string graph-based sequence assembler called

Table 1: Comparison between the two methods in term of time and space complexity. Time and space used for output are ignored.

Algorithm	Used data structures	Time complexity	Space complexity
First method	BP and CSA	$O(n \log n)$	$O(n \log n)$
Second method	BP, LCP and CSA	$O(n \log n)$	$O(n \log k)$

SGA [15]. SGA is a software pipeline for the de novo assembly of sequencing readsets. The experiments also evaluate the scalability of the proposed parallel technique and compare it with the traditional ways to parallelize a suffix tree.

To compare our work with previously presented solutions, we downloaded a solution for all-pairs suffix-prefix problem using Kurtz implementation for a standard suffix tree and the implementation presented by Ohlebusch and Gog for an enhanced suffix array from http://www.uni-ulm.de/fileadmin/website_uni_ulm/iui.inst.190/Forschung/Projekte/seqana/all_pairs_suffix_prefix_problem.tar.gz.

SGA can be downloaded from http://github.com/jts/sga/zipball/master.

In our experiments, the implementation of Sadakane compressed tree presented by Välimäki et al. ([14, 16]) is used. This implementation is tested in the work of Gog [17]. It is available at http://www.cs.helsinki.fi/group/suds/cst/cst_v_1_0.tar.gz. We used it to write two C++ solutions for the APSP problem, compiled with openMP flag to support multithreading. Our implementation is available for download at http://confluence.qu.edu.qa/download/attachments/9240580/SADAApsp.zip.

6.1. Experimental Setup. In our solutions, the user can specify the parallel technique from the command line. For each algorithm, we implement both bottom-up and top-down parallelizing techniques. The number of threads can also be given as a parameter. If the top-down technique is used, the number of threads should be 4^b, where $b \geq 0$. Another parameter is the minimal length to be accepted as a suffix-prefix match between two strings. Accordingly if the length of the longest suffix-prefix match between any two strings is less than the minimal length, then 0 is reported.

In our solution, all strings are concatenated together in one text to build a generalized suffix tree. To overcome the limitation of the number of separators, we used 3 characters to encode enough separators for $k \leq 200^3 = 8000000$ strings (assuming that a character can encode around 200 separators). Our experiments for the sequential test were run on machines having Linux Ubuntu version 11.10, 32-bit with 3 GB RAM, Intel 2.67 GHZ CPU, and 250 GB hard disk.

Our results are obtained by running against randomly generated as well as real data. The random data were generated by a program that outputs random k strings with random lengths, but with a total length of n, where n and k are specified by the user. The random numbers were drawn from a uniform distribution. The real data, which are the complete EST database of *C. elegans*, are downloaded from: http://www.uni-ulm.de/in/theo/research/seqana. The size of the total length for the real data is 167,369,577 bytes with $k = 334,465$ strings. We use the average of 5 readings for each data point. Table 2 describes our data sets.

Table 2: Data sets used in experiments. Sizes in megabytes.

Data Set	Type	Size	Number of strings
Generated by a program	Random data	10–300	100,000
EST of *C. elegans*	Real data	167	334,465

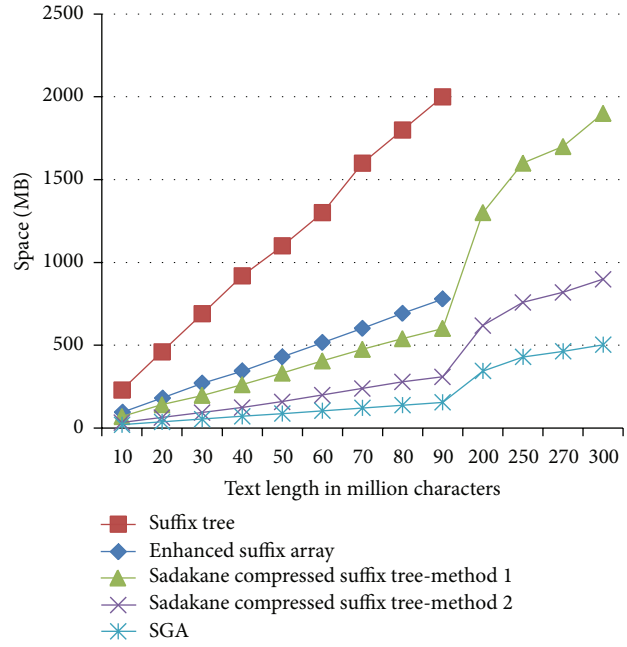


Figure 5: Comparison of space requirements for the three structures (standard suffix tree, enhanced suffix array, and Sadakane compressed suffix tree methods 1 and 2). In addition, the space consumed by the overlap stage in SGA is also shown. The used minimal length is 15.

To test our parallel technique, we used Amazon Web services (AWS) to obtain an instance with 16 cores. Our parallel implementation uses the OpenMP library.

6.2. Experimental Evaluation. Experimental results demonstrate that the first method uses around one-third of the space used by a standard pointer-based suffix tree to solve the same problem, while the second method uses less than one-fifth of the space consumed by a standard suffix tree (see Figure 5). We interpret the difference in space consumption between the two methods as a consequence of the difference in space complexity and the difference in number of the L lists that are used in the two methods.

However, this gain in space has some consequences. Figure 6 demonstrates an obvious slowdown of our solution, which is an expected price to pay as a result of using a compressed data structure. Nevertheless, we were able to run

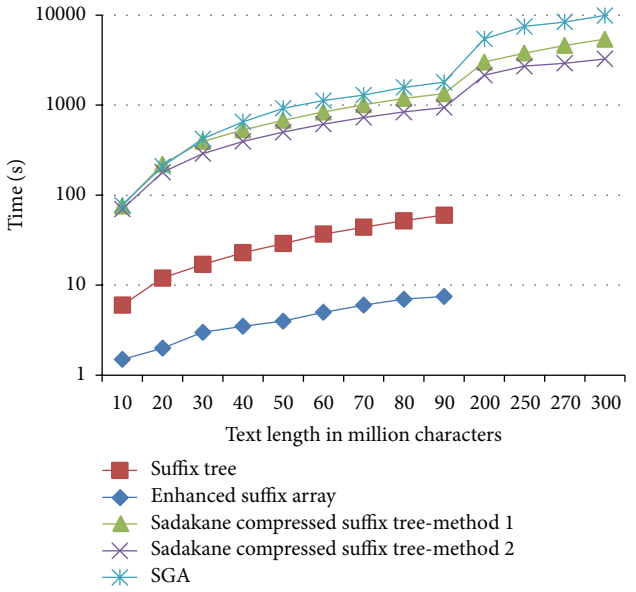


Figure 6: Comparison of time requirements for the three structures (standard suffix tree, enhanced suffix array, and Sadakane compressed suffix tree methods 1 and 2): we could not run the code to build a standard suffix tree for a text with a size which is bigger than 80 MB or an enhanced suffix tree for a text with a size more than 90 MB. The time consumed by SGA in overlap stage is also shown. The used minimal length is 15.

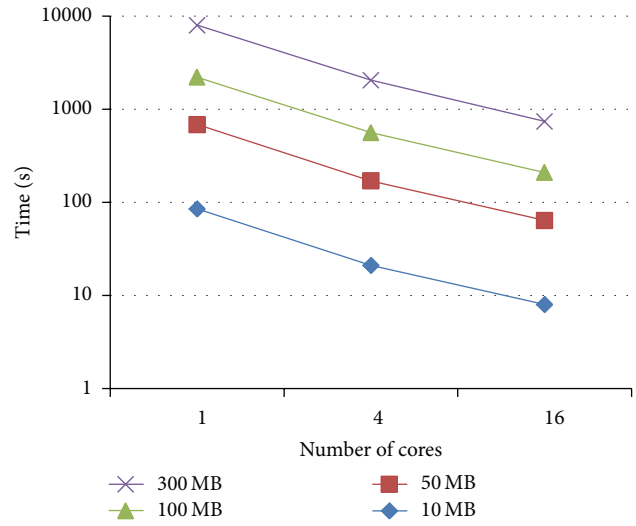


Figure 7: Time requirements for solving APSP for random data with four different text lengths (10 MB, 50 MB, 100 MB and 300 MB), using top-down technique with 1, 4, and 16 cores.

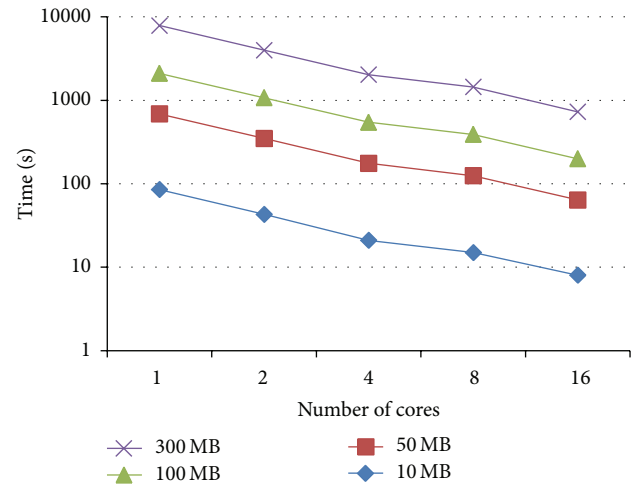


Figure 8: Time requirements for solving APSP for random data with four different text lengths (10 MB, 50 MB, 100 MB, and 300 MB), using bottom-up technique and various number of cores.

our tests using Sadakane compressed suffix tree with a text of a length that is larger than 300 MB, while the maximum size of text which we could run our test on, using a standard suffix tree or an enhanced suffix array, was 90 MB. Therefore, our solution offers better utilization of space resources and allows the user to run larger jobs without the need to upgrade hardware. In addition, our solutions overcome the practical total number of strings limitation (i.e., k is not limited to 200).

Despite the impressive space consumption of SGA, our solutions consume less time than SGA. In addition, the performance of SGA depends dramatically on two factors: the maximum length of a sequence and the minimal length of a match. Since the time complexity of our solution depends on n where n is the total length of all strings, both factors do not affect the performance in our solutions. Our results show that SGA fails to create its index for the overlap stage when $\ell \geq 4000$, where ℓ is the maximum length of a sequence.

The parallel tests show the following: with random data, all techniques take around 24–26% and 9–11% of the time required by the sequential test, with 4 and 16 cores, respectively. Figures 7 and 8 show that both techniques demonstrate good scalability. No significant difference in performance is observed between the two methods.

With real data, the bottom-up technique achieves a speedup of 11–13% compared with the performance of the top-down technique. It is also noticable that the second method [3] consumes with real data more time than the first method [1] (Figure 9). This is due to the fact that the real data has a considerable number of strings which are suffixes of others, which causes the special case (exception) in method 2 to occur frequently.

7. Conclusion

This paper provides two solutions for the all-pairs suffix-prefix problem using Sadakane compressed suffix tree, which reduce the expensive cost of suffix tree in term of space. In spite of significant slowdown in performance, it is clear that the proposed solutions may be preferred when dealing with huge sizes of data because of its modest space requirement. To reduce the performance overhead, the paper presented static

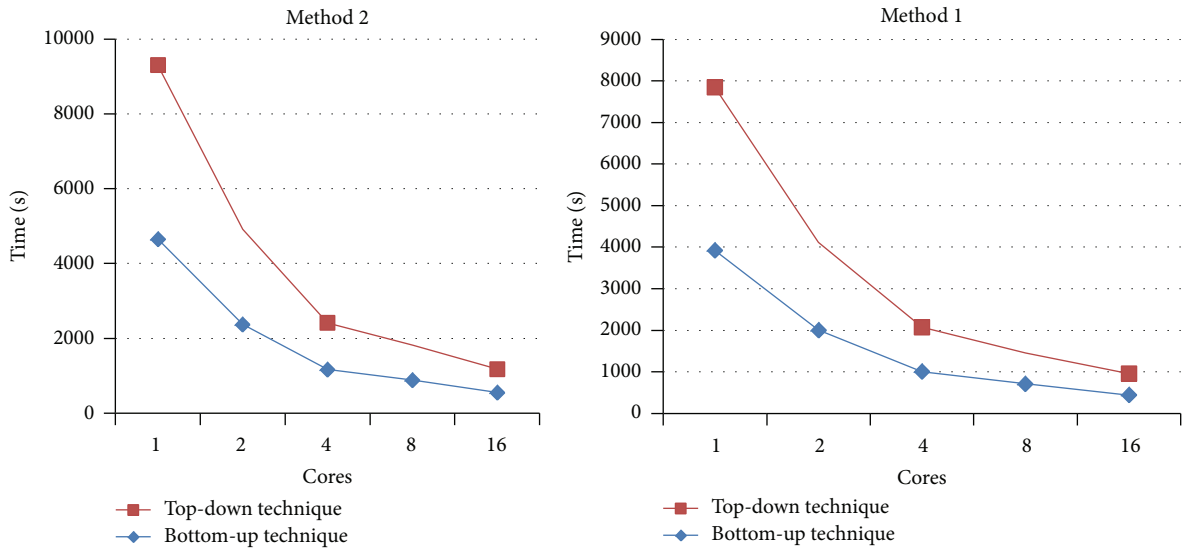


Figure 9: Time requirements for solving APSP for real data (167,369,577 bytes), for both methods using top-down and bottom-up techniques.

and new dynamic techniques to parallelize the proposed solutions. The bottom-up technique performs more efficiently when real data is used, while both techniques perform equally with random data. The presented solutions are not limited to cases with a small number of strings. SGA is superior in terms of space, but it consumes more time than the presented solutions and it does not handle sequences which have large lengths. The paper has demonstrated that it is beneficial to use an enhanced suffix array to solve APSP. It could be worthwhile to explore solving the problem using a compressed suffix array and a compressed largest common prefix (LCP) array by adapting the algorithm presented by Ohlebusch and Gog, which makes the topic a good subject for future study.

Conflict of Interests

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

Acknowledgment

This publication was made possible by NPRP Grant no. 4-1454-1-233 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

References

[1] D. Gusfield, G. M. Landau, and B. Schieber, "An efficient algorithm for the all pairs suffix-prefix problem," *Information Processing Letters*, vol. 41, no. 4, pp. 181–185, 1992.

[2] S. Kurtz, "Reducing the space requirement of suffix trees," *Software: Practice and Experience*, vol. 29, no. 13, pp. 1149–1171, 1999.

[3] E. Ohlebusch and S. Gog, "Efficient algorithms for the all-pairs suffix-prefix problem and the all-pairs substring-prefix problem," *Information Processing Letters*, vol. 110, no. 3, pp. 123–128, 2010.

[4] M. I. Abouelhoda, S. Kurtz, and E. Ohlebusch, "Replacing suffix trees with enhanced suffix arrays," *Journal of Discrete Algorithms*, vol. 2, no. 1, pp. 53–86, 2004.

[5] J. T. Simpson and R. Durbin, "Efficient de novo assembly of large genomes using compressed data structures," *Genome Research*, vol. 22, no. 3, pp. 549–556, 2012.

[6] P. Ferragina, G. Manzini, V. Mäkinen, and G. Navarro, "An alphabet-friendly fm-index," in *String Processing and Information Retrieval*, A. Apostolico and M. Melucci, Eds., vol. 3246 of *Lecture Notes in Computer Science*, pp. 150–160, 2004.

[7] K. Sadakane, "Compressed suffix trees with full functionality," *Theory of Computing Systems*, vol. 41, no. 4, pp. 589–607, 2007.

[8] P. Weiner, "Linear pattern matching algorithms," in *Proceedings of the 14th Annual Symposium on Switching and Automata Theory (SWAT '73)*, pp. 1–11, 1973.

[9] E. Ukkonen, "On-line construction of suffix trees," *Algorithmica*, vol. 14, no. 3, pp. 249–260, 1995.

[10] V. Mäkinen and G. Navarro, "Succinct suffix arrays based on run-length encoding," *Nordic Journal of Computing*, vol. 12, no. 1, pp. 40–66, 2005.

[11] R. Grossi, A. Gupta, and J. S. Vitter, "High-order entropy-compressed text indexes," in *Proceedings of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '03)*, pp. 841–850, November 1998.

[12] M. Burrows and D. J. Wheeler, "A block-sorting lossless data compression algorithm," Tech. Rep., Digital SRC Research Report, 1994.

[13] J. I. Munro, V. Raman, and S. S. Rao, "Space efficient suffix trees," *Journal of Algorithms*, vol. 39, no. 2, pp. 205–222, 2001.

[14] N. Välimäki, V. Mäkinen, W. Gerlach, and K. Dixit, "Engineering a compressed suffix tree implementation," *Journal of Experimental Algorithmics*, vol. 14, article 2, 2009.

[15] E. W. Myers, "The fragment assembly string graph," *Bioinformatics*, vol. 21, supplement 2, pp. 79–85, 2005.

[16] N. Välimäki, W. Gerlach, K. Dixit, and V. Mäkinen, "Compressed suffix tree—a basis for genome-scale sequence analysis," *Bioinformatics*, vol. 23, no. 5, pp. 629–630, 2007.

[17] S. Gog, *Compressed Suffix Trees: Design, Construction, and Applications*, 2011.

Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 236702, 6 pages
http://dx.doi.org/10.1155/2014/236702

Research Article

Association between ε2/3/4, Promoter Polymorphism (−491A/T, −427T/C, and −219T/G) at the Apolipoprotein E Gene, and Mental Retardation in Children from an Iodine Deficiency Area, China

Jun Li,[1,2] Fuchang Zhang,[3] Yunliang Wang,[4] Yan Wang,[5] Wei Qin,[1] Qinghe Xing,[1] Xueqing Qian,[1] Tingwei Guo,[6] Xiaocai Gao,[3] Lin He,[1,2,6] and Jianjun Gao[1,2,6]

[1] *Bio-X Institute, Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders (Ministry of Education), Shanghai Jiao Tong University, 1954 Huashan Road, Shanghai 200030, China*
[2] *Institute for Neuropsychiatric Science and Metabonomics, Changning Mental Health Center, Shanghai 200335, China*
[3] *Key Laboratory of Resource Biology and Biotechnology in Western China (Ministry of Education), College of Life Science, Institute of Population and Health, Northwest University, Xi'an 710069, China*
[4] *Department of Neurology, the 148th Hospital, Zibo, Shandong 255300, China*
[5] *Shanghai Institute of Planned Parenthood Research, 200013, Shanghai, China*
[6] *Institute for Nutritional Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Graduate School of Chinese Academy Sciences, Shanghai 200031, China*

Correspondence should be addressed to Lin He; helinhelin3@gmail.com and Jianjun Gao; jgao@health.bsd.uchicago.edu

Received 16 January 2014; Accepted 18 February 2014; Published 25 March 2014

Academic Editor: Shiwei Duan

Background. Several common single-nucleotide polymorphisms (SNPs) at apolipoprotein E (ApoE) have been linked with late onset sporadic Alzheimer's disease and declining normative cognitive ability in elder people, but we are unclear about their relationship with cognition in children. *Results.* We studied −491A/T, −427T/C, and −219G/T promoter polymorphisms and ε2/ε3/ε4 at ApoE among children with mental retardation (MR, $n = 130$), borderline MR ($n = 124$), and controls ($n = 334$) from an iodine deficiency area in China. The allelic and genotypic distribution of individual locus did not significantly differ among three groups with Mantel-Haenszel χ^2 test ($P > 0.05$). However, frequencies of haplotype of −491A/−427T/−219T/ε4 were distributed as MR > borderline MR > controls (P uncorrected = 0.004), indicating that the presence of this haplotype may increase the risk of disease. *Conclusions.* In this large population-based study in children, we did not find any significant association between single locus of the four common ApoE polymorphisms (−491A/T, −427T/C, −219T/G, and ε2/3/4) and MR or borderline MR. However, we found that the presence of ATTε4 haplotype was associated with an increased risk of MR and borderline MR. Our present work may help enlarge our knowledge of the cognitive role of ApoE across the lifespan and the mechanisms of human cognition.

1. Introduction

Mental retardation (MR) is a condition of neuropsychiatric dysfunction featured by an impairment of intellectual abilities and a deficit of adaption to the environment and the social milieu [1]. Among the complex potential causes of MR, iodine deficiency constitutes the world's greatest single cause of preventable MR [2]. Children born in iodine deficient areas are at great risk for loss of intelligence quotient (IQ) caused by the combined effects of maternal, fetal, and neonatal hypothyroxinemia [3].

Familial aggregation of MR is common in the Qin-Ba Mountain region (Qinba) in midwestern China. Our previous investigation demonstrated that nonspecific mental retardation (NSMR) in Qinba has a heritability with 58.2% (Sd = 19.3%) and may be determined by a lot of minor

effective genes [4]. Numeric associations between general intelligence and specific single-nucleotide polymorphisms (SNPs) in several candidate genes involved in brain function have been reported [5, 6]. Apolipoprotein E (ApoE) plays an important role in cholesterol transport and plasma lipoprotein metabolism [7]. Several overlapping functions have been attributed to ApoE potentially, such as lipid transport, neuronal repair, dendritic growth, maintenance of synaptic plasticity, and anti-inflammatory activities [8]. Its potential role in brain functioning is wide ranging. The common variations of ApoE-ε2/3/4, defined by two common SNPs (rs7412 for ApoE ε2 and rs429358 for ApoE ε4) [9], are the best established susceptibility gene for late-onset Alzheimer's disease (AD) [10, 11]. They were also linked with declining cognitive ability in older people without Alzheimer's disease [12]. The hypothesis that ApoE ε4 is also associated with cognitive decline in old age has been investigated in a few studies with inconsistent results. Despite the existing inconsistency, accumulating evidence or stronger effect on intelligence was found in the older people; by contrast, fewer pieces of evidence were found in younger people. Recently, a few studies have examined the association between cognition and the common ApoE polymorphisms in children or younger adults, but the results were inconsistent [13, 14].

In addition to the ApoE-ε2/ε3/ε4 polymorphisms in coding sequence, other common polymorphisms in promoter including −491A/T (numbered relatively to the transcription start site, rs449647), −427T/C (rs769446), and −219G/T (rs405509) were linked with the quantitative expression of ApoE [15, 16] and associated with AD [17–22]. However, to the best of our knowledge, only one study has looked into whether these polymorphisms are related to intelligence in children individually [13]. No study has investigated the haplotype constituting the polymorphisms in promoter and ApoE-ε2/3/4. Their relationship to intelligence at an early age is a thought of particular interest, because it would potentially aid in understanding the role of ApoE for cognitive functioning across the lifespan [14, 23] and the nature of human cognitive mechanisms.

2. Materials and Methods

2.1. Subjects. Between 1995 and 1998, we carried out an investigation on MR in two areas with high risk in Qinba Mountains, Zhashui county and Ankang county. Among the two counties, some typical villages with high risk of MR were selected as the target areas in our investigation. All children aged 0–14 years from these areas (n = 2974 in Zhashui and n = 2178 in Ankang) were recruited and then screened by different instruments/scales. Details about the study design can be found elsewhere [24]. Briefly, based on the *Chinese Classification of Mental Disorders 2nd Revision* and the classification of mental and behavioral disorders from the World Health Organization (WHO), the clinical psychiatric pediatricians diagnosed, identified, and classified the MR or borderline MR in the Qinba region of China. We also drew family pedigrees to investigate if there existed possible familial mental retardation. All subjects were Han Chinese in origin. We also collected peripheral blood samples from children aged 6–14 years because they were more collaborated. Blood specimens were saved under −70°C until the analysis. The protocol was reviewed and approved by the Ethical Committee of the National Human Genome Center. The guardians of all participated children provided written informed consent.

2.2. Assessment of Mental Retardation. For children aged 6–14 years, we first screened their intelligence using Chinese Standardization of Raven's Standard Progressive Matrices [25]. If children's IQ <85, we assessed their adaptive behavior using Chinese Revised Scale of Social Adaption Ability of Infant-Junior Middle School Student [26]. For children with scores of adaption ability ≤9, we reassessed their intelligence with Chinese-Wechsler Intelligence Scale for Children [27]. MR or borderline MR was finally diagnosed by professional assessment of intelligence and adaptive behavior according to the International Classification of Diseases-10 (1990, WHO).

2.3. Genotyping. A total of 588 blood samples were drawn from children aged 6–14 years with MR (n = 130), borderline MR (n = 124), and non-MR (controls, n = 334).

Leukocyte DNA was extracted using a standard phenol/chloroform method. ApoE promoter polymorphisms −491A/T, −427T/C, and −219G/T were genotyped via a nested PCR amplification. Firstly, the parent 1426 bp fragment was amplified using the primers 5′-CAAGGTCACACAGCTGGCAAC-3′ (forward) and 5′-TCCAATCGACGGCTAGCTACC-3′ (reverse) [17] under the following conditions: (1) 94°C for 5 min, 1 cycle, (2) 94°C for 50 s, 65°C for 50 s, and 72°C for 1 min, 35 cycles, and (3) 72°C for 10 min, 1 cycle and stored at 4°C. The PCR product above was then diluted and used as the templates for the 471 bp fragment amplification with the primers 5′-CACCACGCCTGGCTAACTT-3′ (forward) and 5′-TCACGAGGTGGGCTGTTCT-3′ (reverse) under the following conditions: (1) 95°C for 3 min, 1 cycle, (2) 95°C for 30 s and 68°C for 1 min, 37 cycles, and (3) 72°C for 10 min, 1 cycle and stored at 4°C. The PCR products were subsequently treated according to the standard sequencing procedure of BigDye Terminator v3.1 Cycle Sequencing Kit in the PE Applied Biosystem (PE Applied Biosystem) using either the forward primer or the reverse primer. Electrophoresis was conducted on the ABI PRISM 3100 Genetic Analyzer (PE Applied Biosystem). ApoE-ε2/3/4 genotypes were determined by sequencing and restriction fragment length polymorphism (RFLP) as described in our earlier work [28].

2.4. Statistical Analysis. Genotype and allele frequencies of ApoE and promoter polymorphisms were determined in study groups. Hardy-Weinberg equilibrium and putative haplotypes estimation analysis for −491A/T, −427T/C, −219G/T, and ApoE-ε2/ε3/ε4 were calculated using the software program ARLEQUIN (version 2.0; Genetics and Biometry Laboratory, University of Geneva, Switzerland) [29]. All comparisons for differences of allele, genotype, and

Table 1: Population characteristics according to case and control.

	County	N	Sex ratio (F : M)	Mean age ± SD
Control	Zhashui	250	117/133	9.3 ± 2.8
	Ankang	84	37/47	
Border[a]	Zhashui	80	42/38	10.2 ± 2.9
	Ankang	44	22/22	
MR[b]	Zhashui	73	38/35	10.1 ± 2.8
	Ankang	57	27/30	
Sum		588	283/305	9.7 ± 2.9

[a]Borderline mental retardation (border) group.
[b]Mental retardation (MR) group.

haplotype distributions among groups of MR, borderline MR, and controls were tested with Mantel-Haenszel χ^2 test using SAS 9.2 (SAS, Cary, NC). Finally, pairwise linkage disequilibrium coefficients (D' and r^2) were calculated using the EMLD program developed by Qiqing Huang (http://www.mybiosoftware.com/population-genetics/4717).

3. Results

The distributions of gender, age, and counties did not differ significantly among MR, borderline MR, and controls ($P > 0.05$, Table 1). The individual allelic and genotypic frequencies of −491A/T, −427T/C, −219G/T, and ApoE-ε2/ε3/ε4 by study groups were shown in Table 2. The distributions of genotypes of all the selected loci were in Hardy-Weinberg equilibrium ($P > 0.05$). Comparisons among MR, borderline MR, and control groups with Mantel-Haenszel χ^2 test did not find any significant difference in allele or genotype distributions ($P > 0.05$). However, the comparison of the haplotype frequencies among three study groups revealed that (Table 3) haplotype between the promoter polymorphisms −491A, −427T, −219T, and ε4 was significantly associated with the phenotypes showing the high frequency in MR, lower frequency in borderline MR, and the lowest frequency in controls (Mantel-Haenszel $\chi^2 = 8.09$, P uncorrected = 0.004). LD between all possible pairs of the four polymorphisms was calculated and the pairwise D' and r^2 were shown in Table 4.

4. Discussion

In this study, we investigated if ApoE-ε2/3/4 and promoter polymorphisms of −491A/T, −427T/C, and −219G/T are associated with the risk of MR in children from the iodine deficiency area with high prevalence of MR. We did not find any significant association between individual variation at four common polymorphisms (−491A/T, −427T/C, −219T/G, and ε2/3/4) and MR or borderline MR. Interestingly, haplotype analysis showed that ATTε4 is associated with increased risk of MR and borderline MR.

The physiological and pathological roles of ApoE in the central nervous systems are not entirely clear, but ApoE protein is produced in abundance in the brain by glia, macrophages, and neurons [30, 31]. It is well known that ApoE ε4 is a major genetic risk factor for late onset sporadic AD [10, 22] and has also been investigated for its association not only with dementia but also with normative cognitive development. Most studies have been conducted on adults, especially on older people, but much fewer studies have explored the relationship of ApoE and cognition in children. Several studies on ApoE genotype in relation to cognition in children showed inconsistent results. Most prior studies in school-aged children did not find any significant association with cognitive performance or school assessments, measured in different ways [5, 13, 14, 32], although one study detected a significant association with general cognitive factor [33] and ApoE ε4 predicted higher education in another study [34]. Cavani et al. investigated the association between ApoE genotype and MR in Down syndrome patients in 2000 and found no statistical differences between ApoE allele frequencies of Down syndrome, normal controls, and MR cases [35]. Our analyses based on individual locus found null associations with mental retardation, which is consistent with the prior studies with null associations. Interestingly, haplotype analysis suggested that ApoE may have relationship with intelligence in children. It may be very well helpful for future studies to include haplotype analysis.

To the best of our knowledge, our study is the first attempt to examine haplotype association of ApoE promoter and ε2/3/4 with intelligence in children. Therefore, our analyses should be considered as exploratory and hypothesis generating. The association should be interpreted with cautions although the significance can pass the Bonferroni correction (threshold = 0.005 = 0.05/10 tests). Due to moderate sample size, we could not exclude the possibility of chance finding. An alternative explanation for the positive association is that our subjects had some specialties compared to other studies. The samples in the study were recruited from the relatively isolated iodine deficiency area. The iodine deficient exposure may bring some new features to increase the feasibility of detection. The exon 3 of the ApoE gene possesses sequence homology with coding of the three major thyroid hormone plasma transport proteins (thyroid-binding globulin (TBG), transthyretin (TTR), and albumin) [36, 37]. If the ApoE genotypic variation affects the efficiency of transportation and metabolism of thyroid hormone and therefore influences neuronal cell growth during the first and second trimesters of fetal development [38, 39], the effect size of association between ApoE and intelligence can be modified by the exposure of iodine deficiency. This hypothesis needs confirmation from other studies. On the other hand, MR and borderline MR can be regarded as extreme outcomes of intelligence. Using the extreme phenotypes can increase the power to detect association.

The samples in the study were recruited from the relatively isolated Qinba mountainous area and we did not find a significant difference in allele frequencies among the two counties, which indicated lower risk of stratification bias. Due to the fact of poor education, less developed economy, and transportation, the area is almost isolated from other areas and has much less gene flow [28], so the subjects are helpful in controlling population stratification and have the advantage from the view of a genetic investigation [40, 41].

Table 2: Distribution of allele and genotype across ApoE ε2/ε3/ε4, −491A/T, −427T/C, and −219G/T polymorphisms in MR, borderline MR, and control.

Loci			Allele				Genotype			
	N[a]	ε2	ε3	ε4	P[b]	N[a]	ε2+[c]	ε4+[c]	ε3/ε3	P[b]
ε2/ε3/ε4										
MR	260	26 (23.4)	206 (21.4)	28 (26.9)		130	24 (23.4)	80 (20.3)	26 (28.3)	
Borderline MR	248	20 (18.0)	211 (22.0)	17 (16.4)	0.82	123[a]	18 (18.0)	89 (22.5)	16 (17.4)	0.46
Control	668	65 (58.6)	544 (56.6)	59 (56.7)		329[a]	53 (58.6)	226 (57.2)	50 (54.3)	
	N[a]	T	A		P[b]	N[a]	T/T	T/A	A/A	P[b]
−491A/T										
MR	260	3 (14.3)	257 (22.2)			130	0 (0)	3 (15.8)	127 (22.4)	
Borderline MR	248	3 (14.3)	245 (21.1)		0.21	124	0 (0)	3 (15.8)	121 (21.3)	0.45
Control	668	15 (71.4)	653 (56.7)			334	1 (100)	13 (68.4)	320 (56.3)	
	N[a]	C	T		P[b]	N[a]	C/C	C/T	T/T	P[b]
−427T/C										
MR	260	24 (21.8)	236 (22.1)			130	1 (14.3)	22 (22.9)	107 (22.1)	
Borderline MR	248	20 (18.2)	228 (21.4)		0.64	124	0 (0)	20 (20.8)	104 (21.4)	0.49
Control	668	66 (60.0)	602 (56.5)			334	6 (85.7)	54 (56.3)	274 (56.5)	
	N[a]	G	T		P[b]	N[a]	G/G	G/T	T/T	P[b]
−219G/T										
MR	260	81 (22.2)	179 (22.0)			130	16 (26.2)	49 (20.2)	65 (23.6)	
Borderline MR	248	72 (19.8)	176 (21.7)		0.77	124	10 (16.4)	52 (21.5)	62 (18.9)	0.78
Control	668	211 (58.0)	457 (56.3)			334	35 (57.4)	141 (58.3)	158 (57.4)	

[a]Counts of alleles or genotypes may not add up to total due to excluding the individuals genotyped as ε2/4.
[b]P values are from Mantel-Haenszel χ^2 test in which MR, borderline MR, and control are ordinal variables.
[c]ε2+: ε2/ε2 + ε2/ε3; ε4+: ε3/ε4 + ε4/ε4.

Table 3: Estimated haplotype frequencies for linkage disequilibrium among ApoE −491, −427, −219, and ε2/ε3/ε4.

Haplotype[a]	Frequency	MR (N = 260)	Borderline MR (N = 248)	Control (N = 668)	Mantel-Haenszel χ^2	P value[b]	
						Uncorrected	Corrected
ATTε2	0.06	0.07	0.02	0.06	0.0005	0.98	1.00
ATGε3	0.19	0.19	0.22	0.18	0.37	0.54	1.00
ATTε3	0.55	0.51	0.60	0.55	0.40	0.53	1.00
ATTε4	0.07	0.10	0.07	0.05	8.09	0.004	0.02
ACGε2	0.02	0.01	0.06	0.02	0.03	0.87	1.00
ACGε3	0.06	0.08	0.01	0.07	0.07	0.80	1.00

[a]Haplotypes are from alleles of 4 polymorphisms: ApoE −491, −427, −219, and ε2/3/4. Only the haplotypes with frequency 0.01 and higher are shown.
[b]Corrected P value according to Bonferroni correction (6 tests).

Table 4: Pairwise linkage disequilibrium (D'/r^2) of −491A/T, −427T/C, −219T/G, and ε2/ε3/ε4 polymorphisms.

SNPs	LD estimate (D' or r^2) for marker pair			
	−491A/T	−427T/C	−219T/G	ε2/ε3/ε4
−491A/T	—	0.921	0.241	0.216
−427T/C	0.002	—	0.895	0.589
−219T/G	0.001	0.171	—	0.286
ε2/ε3/ε4	0.001	0.075	0.049	—

The standardized D' values are shown above the diagonal, and the r^2 values are shown below the diagonal.

5. Conclusions

In summary, in this large population-based study, we did not find any significant association between single locus of the four common ApoE polymorphisms (−491A/T, −427T/C, −219T/G, and ε2/3/4) and MR or borderline MR in children. However, we found that the presence of ATTε4 haplotype was associated with an increased risk of MR and borderline MR. Our present work may help enlarge our knowledge of the role of ApoE in cognitive functioning across the lifespan and the mechanisms of human cognition.

Abbreviations

ApoE: Apolipoprotein E
MR: Mental retardation
IQ: Intelligence quotient
NSMR: Nonspecific mental retardation
SNP: Single-nucleotide polymorphism
AD: Alzheimer's disease
RFLP: Restriction fragment length polymorphism.

Conflict of Interests

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

Authors' Contribution

Jianjun Gao, Fuchang Zhang, and Lin He planned and supervised the study. Fuchang Zhang and Lin He gained funding for this study. Jianjun Gao, Tingwei Guo, and Fuchang Zhang collected the blood samples. Jianjun Gao and Jun Li conducted the experiments. Jianjun Gao, Jun Li, and Yan Wang analyzed data and wrote the final version. Yunliang Wang, Wei Qin, Qinghe Xing and Xueqing Qian provided intellectual inputs and revised the paper. All the authors read and approved the final paper and submission. Jun Li, Fuchang Zhang, and Yunliang Wang contributed equally to this work.

Acknowledgments

The authors sincerely thank all participants in this study, particularly Drs. Ruilin Li, Zhenlin Wang, and Hongxing Dai. This work was supported by Grants from the Ministry of Education, China, the National Natural Science Foundation of China, the Shanghai Municipal Commission for Science and Technology, the national 863 Project and 973 Project (no. 2002BA711A07-01), and "the Tenth Five-Year Plan" National Tackle Problem Item (no. 2001BA901A49).

References

[1] A. P. Association, *Diagnostic and Statistical Manual of Mental Disorders*, The American Psychiatric Association, Washington, DC, USA, 4th edition, 1994.

[2] F. Delange, "The role of iodine in brain development," *Proceedings of the Nutrition Society*, vol. 59, no. 1, pp. 75–79, 2000.

[3] D. Glinoer and F. Delange, "The potential repercussions of maternal, fetal, and neonatal hypothyroxinemia on the progeny," *Thyroid*, vol. 10, no. 10, pp. 871–887, 2000.

[4] K. J. Zhang, K. J. Z. F, Z. J. Zheng et al., "An analysis of inheriting type of non-causing mental retarded children in Zhashui experimental station," *Joumal of Northwest University*, vol. 35, pp. 597–600, 2005.

[5] C. F. Chabris, B. M. Hebert, D. J. Benjamin et al., "Most reported genetic associations with general intelligence are probably false positives," *Psychological Science*, vol. 23, no. 11, pp. 1314–1323, 2012.

[6] L. M. Houlihan, S. E. Harris, M. Luciano et al., "Replication study of candidate genes for cognitive abilities: The Lothian Birth Cohort 1936," *Genes, Brain and Behavior*, vol. 8, no. 2, pp. 238–247, 2009.

[7] K. H. Weisgraber and R. W. Mahley, "Human apolipoprotein E: the Alzheimer's disease connection," *The FASEB Journal*, vol. 10, no. 13, pp. 1485–1494, 1996.

[8] D. K. Lahiri, K. Sambamurti, and D. A. Bennett, "Apolipoprotein gene and its interaction with the environmentally driven risk factors: molecular, genetic and epidemiological studies of Alzheimer's disease," *Neurobiology of Aging*, vol. 25, no. 5, pp. 651–660, 2004.

[9] J. Gao, X. Huang, Y. Park et al., "Apolipoprotein E genotypes and the risk of Parkinson disease," *Neurobiology of Aging*, vol. 32, no. 11, pp. 2106.e1–2106.e6, 2011.

[10] L. A. Farrer, L. A. Cupples, J. L. Haines et al., "Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease: a meta-analysis," *The Journal of the American Medical Association*, vol. 278, no. 16, pp. 1349–1356, 1997.

[11] K. I. Morley and G. W. Montgomery, "The genetics of cognitive processes: candidate genes in humans and animals," *Behavior Genetics*, vol. 31, no. 6, pp. 511–531, 2001.

[12] O. J. G. Schiepers, S. E. Harris, A. J. Gow et al., "APOE E4 status predicts age-related cognitive decline in the ninth decade: longitudinal follow-up of the Lothian Birth Cohort 1921," *Molecular Psychiatry*, vol. 17, no. 3, pp. 315–324, 2012.

[13] D. Turic, P. J. Fisher, R. Plomin, and M. J. Owen, "No association between apolipoprotein E polymorphisms and general cognitive ability in children," *Neuroscience Letters*, vol. 299, no. 1-2, pp. 97–100, 2001.

[14] A. E. Taylor, P. A. I. Guthrie, G. D. Smith et al., "IQ, educational attainment, memory and plasma lipids: associations with apolipoprotein e genotype in 5995 children," *Biological Psychiatry*, vol. 70, no. 2, pp. 152–158, 2011.

[15] J.-C. Lambert, C. Berr, F. Pasquier et al., "Pronounced impact of Th1/E47cs mutation compared with -491 AT mutation on neural APOE gene expression and risk of developing Alzheimer's disease," *Human Molecular Genetics*, vol. 7, no. 9, pp. 1511–1516, 1998.

[16] J. D. Smith, A. Melian, T. Leff, and J. L. Breslow, "Expression of the human apolipoprotein E gene is regulated by multiple positive and negative elements," *The Journal of Biological Chemistry*, vol. 263, no. 17, pp. 8300–8308, 1988.

[17] T. Town, D. Paris, D. Fallin et al., "The -491A/T apolipoprotein E promoter polymorphism association with Alzheimer's disease: independent risk and linkage disequilibrium with the known APOE polymorphism," *Neuroscience Letters*, vol. 252, no. 2, pp. 95–98, 1998.

[18] S. Helisalmi, M. Hiltunen, P. Valonen et al., "Promoter polymorphism (-491A/T) in the APOE gene of Finnish Alzheimer's

disease patients and control individuals," *Journal of Neurology*, vol. 246, no. 9, pp. 821–824, 1999.

[19] A. Juhász, A. Palotás, Z. Janka et al., "ApoE -491A/T promoter polymorphism is not an independent risk factor, but associated with the ε4 allele in Hungarian Alzheimer's dementia population," *Neurochemical Research*, vol. 30, no. 5, pp. 591–596, 2005.

[20] G. Parra-Bonilla, G. Arboleda, J. Yunis et al., "Haplogroup analysis of the risk associated with APOE promoter polymorphisms (-219T/G, -491A/T and -427T/C) in Colombian Alzheimer's disease patients," *Neuroscience Letters*, vol. 349, no. 3, pp. 159–162, 2003.

[21] B. T. Heijmans, P. E. Slagboom, J. Gussekloo et al., "Association of APOE epsilon2/epsilon3/epsilon4 and promoter gene variants with dementia but not cardiovascular mortality in old age," *American Journal of Medical Genetics*, vol. 107, no. 3, pp. 201–208, 2002.

[22] X.-Y. Xin, J.-Q. Ding, and S.-D. Chen, "Apolipoprotein e promoter polymorphisms and risk of Alzheimer's disease: evidence from meta-analysis," *Journal of Alzheimer's Disease*, vol. 19, no. 4, pp. 1283–1294, 2010.

[23] O. Sternång and Å. Wahlin, "What is the role of apolipoprotein e for cognitive functioning across the lifespan?" *Biological Psychiatry*, vol. 70, no. 2, pp. 109–110, 2011.

[24] J. Gao, X. Gao, W. Qin et al., "No observable relationship between the ACE gene insertion/deletion polymorphism and psychometric IQ and psychomotor ability in Chinese children," *Neuropsychobiology*, vol. 53, no. 4, pp. 196–202, 2006.

[25] H. Zhang and X. P. Wang, "Chinese standardisation of Raven's Standard Progressive Matrices," *Psychological Test Bulletin*, vol. 2, no. 11, pp. 36–39, 1989.

[26] Q. Zuo, Z. Zhang, and W. Liang, *Social Adaption Ability of Infant-Junior Middle School Student*, Beijing Medical University, Beijing, China, 1988.

[27] Y. Gong and X. Y. Dai, "China-Wechsler younger children scale of intelligence (CWYCSI)," *Acta Psychologica Sinica*, vol. 20, pp. 364–375, 1988.

[28] J. Gao, F. Zhang, T. Guo et al., "Distribution of apolipoprotein E allele frequencies of the Han Chinese in an iodine-deficient mountainous area," *Annals of Human Biology*, vol. 31, no. 5, pp. 578–585, 2004.

[29] L. Excoffier, G. Laval, and S. Schneider, "Arlequin (version 3. 0): an integrated software package for population genetics data analysis," *Evolutionary Bioinformatics Online*, vol. 1, pp. 47–50, 2005.

[30] R. E. Pitas, J. K. Boyles, and S. H. Lee, "Astrocytes synthesize apolipoprotein E and metabolize apolipoprotein E-containing lipoproteins," *Biochimica et Biophysica Acta*, vol. 917, no. 1, pp. 148–161, 1987.

[31] D. A. Elliott, W. S. Kim, D. A. Jans, and B. Garner, "Apoptosis induces neuronal apolipoprotein-E synthesis and localization in apoptotic bodies," *Neuroscience Letters*, vol. 416, no. 2, pp. 206–210, 2007.

[32] I. J. Deary, M. C. Whiteman, A. Pattie et al., "Ageing: cognitive change and the *APOE ε4* allele," *Nature*, vol. 418, no. 6901, p. 932, 2002.

[33] M. Luciano, A. J. Gow, S. E. Harris et al., "Cognitive ability at age 11 and 70 years, information processing speed, and APOE variation: the Lothian Birth Cohort 1936 Study," *Psychology and Aging*, vol. 24, no. 1, pp. 129–138, 2009.

[34] J. A. Hubacek, J. Pitha, Z. Škodová, V. Adámková, V. Lánská, and R. Poledne, "A possible role of apolipoprotein E polymorphism in predisposition to higher education," *Neuropsychobiology*, vol. 43, no. 3, pp. 200–203, 2001.

[35] S. Cavani, A. Tamaoka, A. Moretti et al., "Plasma levels of amyloid beta 40 and 42 are independent from ApoE genotype and mental retardation in Down syndrome," *American Journal of Medical Genetics*, vol. 95, no. 3, pp. 224–228, 2000.

[36] S. Benvenga, "A thyroid hormone binding motif is evolutionarily conserved in apolipoproteins," *Thyroid*, vol. 7, no. 4, pp. 605–611, 1997.

[37] S. Benvenga, H. J. Cahnmann, D. Rader, M. Kindt, and J. Robbins, "Thyroxine binding to the apolipoproteins of high density lipoproteins HDL2 and HDL3," *Endocrinology*, vol. 131, no. 6, pp. 2805–2811, 1992.

[38] S. Benvenga, H. J. Cahnmann, and J. Robbins, "Characterization of thyroid hormone binding to apolipoprotein-E: localization of the binding site in the exon 3-coded domain," *Endocrinology*, vol. 133, no. 3, pp. 1300–1305, 1993.

[39] C. Xue-Yi, J. Xin-Min, D. Zhi-Hong et al., "Timing of vulnerability of the brain to iodine deficiency in endemic cretinism," *The New England Journal of Medicine*, vol. 331, no. 26, pp. 1739–1744, 1994.

[40] S. Shifman and A. Darvasi, "The value of isolated populations," *Nature Genetics*, vol. 28, no. 4, pp. 309–310, 2001.

[41] T. Laitinen, "The value of isolated populations in genetic studies of allergic diseases," *Current Opinion in Allergy and Clinical Immunology*, vol. 2, no. 5, pp. 379–382, 2002.