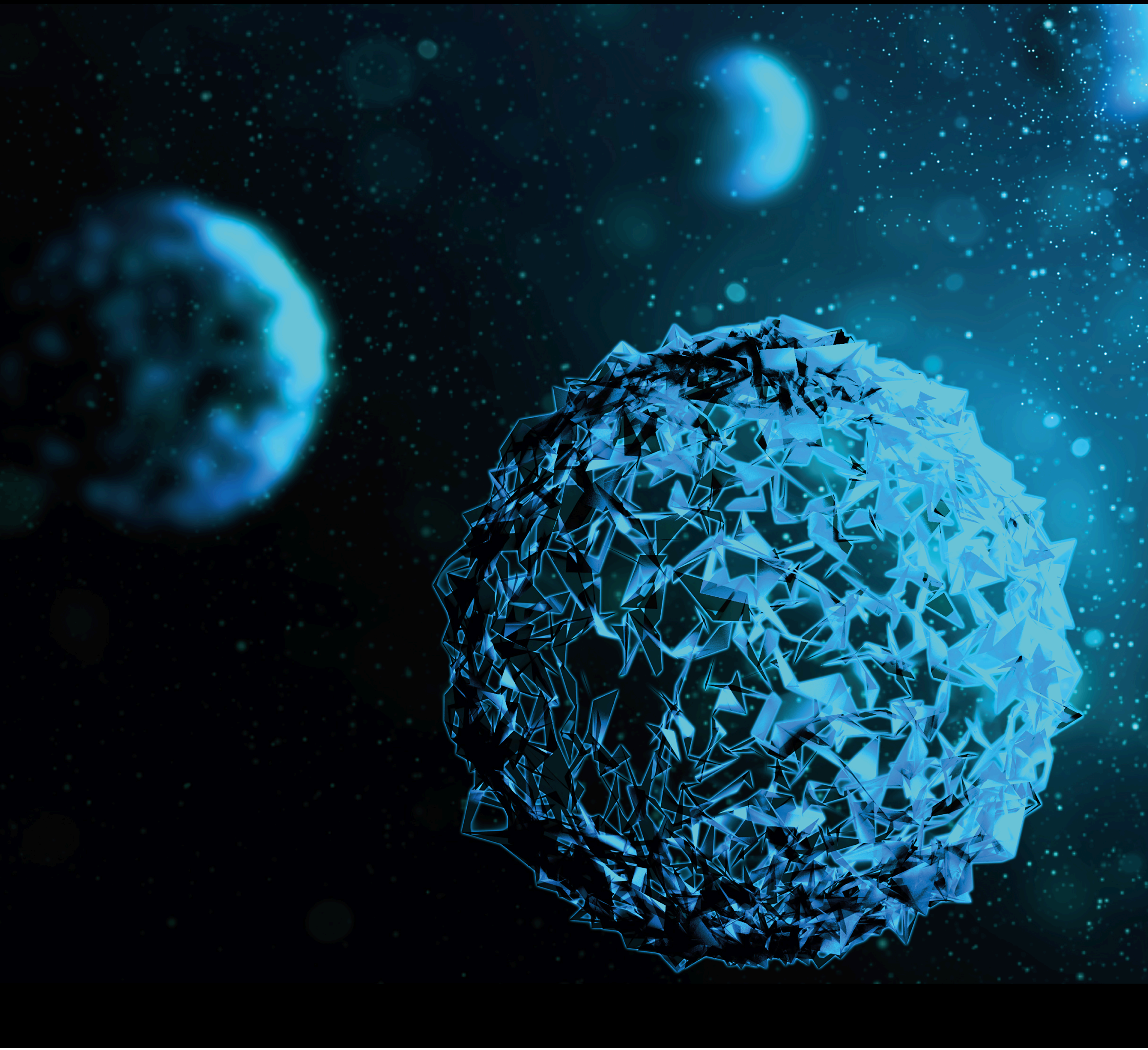
BioMed Research International

Cloud Analytic Based Applications for Bioinformatics

Lead Guest Editor: Parameshachari B. D.
Guest Editors: Siddesh G. M. and Hanifa Abdullah



Cloud Analytic Based Applications for Bioinformatics

BioMed Research International

Cloud Analytic Based Applications for Bioinformatics

Lead Guest Editor: Parameshachari B. D.
Guest Editors: Siddesh G. M. and Hanifa Abdullah

Section Editors

Academic Editors

Bioinformatics

Contents

Contents

Hindawi
BioMed Research International
Volume 2024, Article ID 9898717, 1 page
https://doi.org/10.1155/2024/9898717

Retraction

Retracted: Multiomics Analysis of Transcriptome, Epigenome, and Genome Uncovers Putative Mechanisms for Dilated Cardiomyopathy

BioMed Research International

Received 12 March 2024; Accepted 12 March 2024; Published 20 March 2024

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope

(2) Discrepancies in the description of the research reported

(3) Discrepancies between the availability of data and the research described

(4) Inappropriate citations

(5) Incoherent, meaningless and/or irrelevant content included in the article

(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] L. Liu, J. Huang, Y. Liu et al., "Multiomics Analysis of Transcriptome, Epigenome, and Genome Uncovers Putative Mechanisms for Dilated Cardiomyopathy," *BioMed Research International*, vol. 2021, Article ID 6653802, 29 pages, 2021.

Hindawi
BioMed Research International
Volume 2024, Article ID 9895849, 1 page
https://doi.org/10.1155/2024/9895849

Retraction

Retracted: Dysregulated Circulating Apoptosis- and Autophagy-Related lncRNAs as Diagnostic Markers in Coronary Artery Disease

BioMed Research International

Received 12 March 2024; Accepted 12 March 2024; Published 20 March 2024

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope

(2) Discrepancies in the description of the research reported

(3) Discrepancies between the availability of data and the research described

(4) Inappropriate citations

(5) Incoherent, meaningless and/or irrelevant content included in the article

(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] L. Zhang, D. Lou, D. He et al., "Dysregulated Circulating Apoptosis- and Autophagy-Related lncRNAs as Diagnostic Markers in Coronary Artery Disease," *BioMed Research International*, vol. 2021, Article ID 5517786, 19 pages, 2021.

Hindawi
BioMed Research International
Volume 2024, Article ID 9879061, 1 page
https://doi.org/10.1155/2024/9879061

Retraction

Retracted: A Novel Smart City-Based Framework on Perspectives for Application of Machine Learning in Combating COVID-19

BioMed Research International

Received 12 March 2024; Accepted 12 March 2024; Published 20 March 2024

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope

(2) Discrepancies in the description of the research reported

(3) Discrepancies between the availability of data and the research described

(4) Inappropriate citations

(5) Incoherent, meaningless and/or irrelevant content included in the article

(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] A. E. Ezugwu, I. A. T. Hashem, O. N. Oyelade et al., "A Novel Smart City-Based Framework on Perspectives for Application of Machine Learning in Combating COVID-19," *BioMed Research International*, vol. 2021, Article ID 5546790, 15 pages, 2021.

Hindawi
BioMed Research International
Volume 2024, Article ID 9858439, 1 page
https://doi.org/10.1155/2024/9858439

Retraction

Retracted: Optimum Feature Selection with Particle Swarm Optimization to Face Recognition System Using Gabor Wavelet Transform and Deep Learning

BioMed Research International

Received 12 March 2024; Accepted 12 March 2024; Published 20 March 2024

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope

(2) Discrepancies in the description of the research reported

(3) Discrepancies between the availability of data and the research described

(4) Inappropriate citations

(5) Incoherent, meaningless and/or irrelevant content included in the article

(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] S. Ahmed, M. Frikha, T. D. H. Hussein, and J. Rahebi, "Optimum Feature Selection with Particle Swarm Optimization to Face Recognition System Using Gabor Wavelet Transform and Deep Learning," *BioMed Research International*, vol. 2021, Article ID 6621540, 13 pages, 2021.

Hindawi
BioMed Research International
Volume 2024, Article ID 9854370, 1 page
https://doi.org/10.1155/2024/9854370

Retraction

Retracted: A Prognostic Model for Brain Glioma Patients Based on 9 Signature Glycolytic Genes

BioMed Research International

Received 12 March 2024; Accepted 12 March 2024; Published 20 March 2024

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope

(2) Discrepancies in the description of the research reported

(3) Discrepancies between the availability of data and the research described

(4) Inappropriate citations

(5) Incoherent, meaningless and/or irrelevant content included in the article

(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] X. Bingxiang, W. Panxing, F. Lu, Y. Xiuyou, and D. Chao, "A Prognostic Model for Brain Glioma Patients Based on 9 Signature Glycolytic Genes," *BioMed Research International*, vol. 2021, Article ID 6680066, 15 pages, 2021.

Hindawi
BioMed Research International
Volume 2024, Article ID 9848271, 1 page
https://doi.org/10.1155/2024/9848271

Retraction

Retracted: Identification and Validation of Potential Biomarkers and Pathways for Idiopathic Pulmonary Fibrosis by Comprehensive Bioinformatics Analysis

BioMed Research International

Received 12 March 2024; Accepted 12 March 2024; Published 20 March 2024

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope

(2) Discrepancies in the description of the research reported

(3) Discrepancies between the availability of data and the research described

(4) Inappropriate citations

(5) Incoherent, meaningless and/or irrelevant content included in the article

(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] W. Qian, X. Cai, Q. Qian, and X. Zhang, "Identification and Validation of Potential Biomarkers and Pathways for Idiopathic Pulmonary Fibrosis by Comprehensive Bioinformatics Analysis," *BioMed Research International*, vol. 2021, Article ID 5545312, 15 pages, 2021.

Hindawi
BioMed Research International
Volume 2024, Article ID 9846925, 1 page
https://doi.org/10.1155/2024/9846925

Retraction

Retracted: Distinct Molecular Subtypes of Diffuse Large B Cell Lymphoma Patients Treated with Rituximab-CHOP Are Associated with Different Clinical Outcomes and Molecular Mechanisms

BioMed Research International

Received 12 March 2024; Accepted 12 March 2024; Published 20 March 2024

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope

(2) Discrepancies in the description of the research reported

(3) Discrepancies between the availability of data and the research described

(4) Inappropriate citations

(5) Incoherent, meaningless and/or irrelevant content included in the article

(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] H. Yu, S. Peng, S. Han, X. Chen, Q. Lyu, and T. Lei, "Distinct Molecular Subtypes of Diffuse Large B Cell Lymphoma Patients Treated with Rituximab-CHOP Are Associated with Different Clinical Outcomes and Molecular Mechanisms," *BioMed Research International*, vol. 2021, Article ID 5514726, 13 pages, 2021.

Hindawi
BioMed Research International
Volume 2024, Article ID 9838747, 1 page
https://doi.org/10.1155/2024/9838747

Retraction

Retracted: Multiomics Analysis of Genetics and Epigenetics Reveals Pathogenesis and Therapeutic Targets for Atrial Fibrillation

BioMed Research International

Received 12 March 2024; Accepted 12 March 2024; Published 20 March 2024

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope

(2) Discrepancies in the description of the research reported

(3) Discrepancies between the availability of data and the research described

(4) Inappropriate citations

(5) Incoherent, meaningless and/or irrelevant content included in the article

(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] L. Liu, J. Huang, B. Wei et al., “Multiomics Analysis of Genetics and Epigenetics Reveals Pathogenesis and Therapeutic Targets for Atrial Fibrillation,” *BioMed Research International*, vol. 2021, Article ID 6644827, 36 pages, 2021.

Hindawi
BioMed Research International
Volume 2024, Article ID 9837415, 1 page
https://doi.org/10.1155/2024/9837415

Retraction

Retracted: Transcriptional Profiling Uncovers Biologically Significant RNAs and Regulatory Networks in Nucleus Pulposus from Intervertebral Disc Degeneration Patients

BioMed Research International

Received 12 March 2024; Accepted 12 March 2024; Published 20 March 2024

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope

(2) Discrepancies in the description of the research reported

(3) Discrepancies between the availability of data and the research described

(4) Inappropriate citations

(5) Incoherent, meaningless and/or irrelevant content included in the article

(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] Y. Chen, B. Cai, X. Lian, J. Xu, and T. Zhang, "Transcriptional Profiling Uncovers Biologically Significant RNAs and Regulatory Networks in Nucleus Pulposus from Intervertebral Disc Degeneration Patients," *BioMed Research International*, vol. 2021, Article ID 6696335, 33 pages, 2021.

Hindawi
BioMed Research International
Volume 2024, Article ID 9762690, 1 page
https://doi.org/10.1155/2024/9762690

Retraction

Retracted: Weighted Gene Coexpression Network Analysis Uncovers Critical Genes and Pathways for Multiple Brain Regions in Parkinson's Disease

BioMed Research International

Received 12 March 2024; Accepted 12 March 2024; Published 20 March 2024

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope

(2) Discrepancies in the description of the research reported

(3) Discrepancies between the availability of data and the research described

(4) Inappropriate citations

(5) Incoherent, meaningless and/or irrelevant content included in the article

(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] J. Huang, L. Liu, L. Qin, H. Huang, and X. Li, "Weighted Gene Coexpression Network Analysis Uncovers Critical Genes and Pathways for Multiple Brain Regions in Parkinson's Disease," *BioMed Research International*, vol. 2021, Article ID 6616434, 46 pages, 2021.

Hindawi
BioMed Research International
Volume 2024, Article ID 9753878, 1 page
https://doi.org/10.1155/2024/9753878

Retraction

Retracted: A Permissioned Blockchain-Based Clinical Trial Service Platform to Improve Trial Data Transparency

BioMed Research International

Received 12 March 2024; Accepted 12 March 2024; Published 20 March 2024

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope

(2) Discrepancies in the description of the research reported

(3) Discrepancies between the availability of data and the research described

(4) Inappropriate citations

(5) Incoherent, meaningless and/or irrelevant content included in the article

(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] L. Hang, B. Kim, K. Kim, and D. Kim, "A Permissioned Blockchain-Based Clinical Trial Service Platform to Improve Trial Data Transparency," *BioMed Research International*, vol. 2021, Article ID 5554487, 22 pages, 2021.

Hindawi
BioMed Research International
Volume 2024, Article ID 9879061, 1 page
https://doi.org/10.1155/2024/9879061

Retraction

Retracted: A Novel Smart City-Based Framework on Perspectives for Application of Machine Learning in Combating COVID-19

BioMed Research International

Received 12 March 2024; Accepted 12 March 2024; Published 20 March 2024

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope

(2) Discrepancies in the description of the research reported

(3) Discrepancies between the availability of data and the research described

(4) Inappropriate citations

(5) Incoherent, meaningless and/or irrelevant content included in the article

(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] A. E. Ezugwu, I. A. T. Hashem, O. N. Oyelade et al., "A Novel Smart City-Based Framework on Perspectives for Application of Machine Learning in Combating COVID-19," *BioMed Research International*, vol. 2021, Article ID 5546790, 15 pages, 2021.

Hindawi
BioMed Research International
Volume 2021, Article ID 5546790, 15 pages
https://doi.org/10.1155/2021/5546790

Review Article

A Novel Smart City-Based Framework on Perspectives for Application of Machine Learning in Combating COVID-19

Absalom E. Ezugwu,[1] Ibrahim Abaker Targio Hashem,[2] Olaide N. Oyelade,[1] Mubarak Almutari,[3] Mohammed A. Al-Garadi,[4] Idris Nasir Abdullahi,[5] Olumuyiwa Otegbeye,[6] Amit K. Shukla,[7] and Haruna Chiroma[8]

[1] *School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, King Edward Road, Pietermaritzburg Campus, Pietermaritzburg, KwaZulu-Natal 3201, South Africa*
[2] *College of Computing and Informatics, Department of Computer Science, University of Sharjah, 27272 Sharjah, UAE*
[3] *College of Computer Science, University of Hafr Al Batin, Saudi Arabia*
[4] *Department of Biomedical Informatics, Emory University, Atlanta, USA*
[5] *Department of Medical Laboratory Science, College of Medical Sciences, Ahmadu Bello University, Zaria, Nigeria*
[6] *School of Computer Science and Applied Mathematics, University of the Witwatersrand, South Africa*
[7] *IRISA Laboratory, ENSSAT, University of Rennes 1, France*
[8] *Future Technology Research Center, National Yunlin University of Science and Technology, Taiwan*

Correspondence should be addressed to Absalom E. Ezugwu; ezugwua@ukzn.ac.za
and Haruna Chiroma; chiromaharun@fcetgombe.edu.ng

Received 23 February 2021; Revised 17 May 2021; Accepted 17 August 2021; Published 11 September 2021

Academic Editor: Parameshachari B. D.

The spread of COVID-19 worldwide continues despite multidimensional efforts to curtail its spread and provide treatment. Efforts to contain the COVID-19 pandemic have triggered partial or full lockdowns across the globe. This paper presents a novel framework that intelligently combines machine learning models and the Internet of Things (IoT) technology specifically to combat COVID-19 in smart cities. The purpose of the study is to promote the interoperability of machine learning algorithms with IoT technology by interacting with a population and its environment to curtail the COVID-19 pandemic. Furthermore, the study also investigates and discusses some solution frameworks, which can generate, capture, store, and analyze data using machine learning algorithms. These algorithms can detect, prevent, and trace the spread of COVID-19 and provide a better understanding of the disease in smart cities. Similarly, the study outlined case studies on the application of machine learning to help fight against COVID-19 in hospitals worldwide. The framework proposed in the study is a comprehensive presentation on the major components needed to integrate the machine learning approach with other AI-based solutions. Finally, the machine learning framework presented in this study has the potential to help national healthcare systems in curtailing the COVID-19 pandemic in smart cities. In addition, the proposed framework is poised as a pointer for generating research interests that would yield outcomes capable of been integrated to form an improved framework.

1. Introduction

The novel coronavirus named severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has caused unprecedented numbers of deaths from coronavirus disease 2019 (COVID-19) worldwide. For instance, the United States of America recorded over 3,000 deaths within one period of 24 hours in December 2020, and the highest in the world for a single day and as of October 2020 has recorded a total of 270,642 deaths. The first human-to-human transmission of COVID-19 was reported to the World Health Organization (WHO) on 30th December 2019. Thereafter, several retrospective studies revealed that many COVID-19 patients started showing pneumonia symptoms in early December

([1–3]. Even though there are scientific controversies and theories over the date and origin of the SARS-CoV-2, it is widely accepted that the novel coronavirus originated from Wuhan, China [4]. Genomic sequences of the early isolates of SARS-CoV-2 from infected patients in Wuhan showed over 88% nucleotide homology with two bat-like SARS coronaviruses, which pointed strongly towards the zoonotic source with bats serving as reservoir hosts of the SARS-CoV-2 [4]. There are ongoing searches for possible intermediate hosts, which might have aided the transmission of the virus to humans. SARS-CoV-2 is a droplet borne pathogen that spreads by contact with humans when they are exposed to oral or nasal secretions of clinically symptomatically or asymptomatically infected persons [5].

SARS-CoV-2 has tropism in cells and tissues that express the angiotensin-converting enzyme 2 as a receptor. These receptors are mainly found in the respiratory tract and to a limited extent in the kidney, heart, and gastrointestinal tract. The virus docks onto the receptor through the receptor-binding domain (RBD) of its spike glycoprotein. This represents the first step of viral replication and pathogenesis [6]. As the virus replicates in the respiratory tract, it provokes respiratory symptoms, mainly dry cough, difficulty in breathing, and sore throat. Then, it disseminates through the blood to other tissues and organs, causing viremia and high fever. Hence, these symptoms, together with body weakness and pains, represent the primary clinical symptoms of COVID-19 [7].

The majority of SARS-CoV-2-infected persons remain asymptomatic, with the infection being self-limiting. However, some 5% of infected persons suffer severe COVID-19 [6]. The major determining factors for severe and possibly fatal COVID-19 include advanced age (>60 years) and underlying cardiovascular, immunological, metabolic, or respiratory comorbidities. Based on available scientific reports, the transmission of SARS-CoV-2 depends on human-human, animal-human, and environment-human transmission [8]. For now, preserving human life and health security is the major concern of most countries and territories. Hence, it prompted legislation to implement and enforce adequate infection prevention and control measures for these high priority pathogens [9]. Therefore, combatting the COVID-19 needs better understanding.

To better understand COVID-19 in terms of its pattern of spread, identifying the most susceptible people according to their distinct genetic and physiological characteristics for COVID-19, to improve the accuracy and speed of its diagnosis, and develop new therapies, machine learning algorithms are required to analyze the large scale COVID-19 datasets [10]. As a result, scholars have attempted to apply machine learning algorithms to combating COVID-19 from various perspectives. For example, drug discovery targeting COVID-19 is proposed in Ge et al. [11]; machine learning approaches are applied in CRISPR-based COVID-19 surveillance using genomic data as proposed by Metsky, Freije, Kosoko-Thoroddsen, Sabeti, and Myhrvold [12]. Similalry, Pandey, Gautam, Bhagat, and Sethi [13] develop a system for creating awareness about the importance of hand washing to contain the spread of COVID-19, while Yan et al. [14], in their study, employed machine learning techniques to predict the survival of severely affected COVID-19 patients. More so, the classification of novel pathogens for the COVID-19 is presented in Randhawa et al. [15]; a deep learning-based system for quantifying the volume of lung infections is reported in Yan et al. [14], and automated deep learning COVID-19 patient detection and monitoring systems are reported in Gozes et al. [16], Oyelade and Ezugwu [17], and Oyelade et al. [18]; lastly, a generative network is used for the design of COVID-19 3C-like protease inhibitors [19].

However, each of these previous studies focuses on only a single aspect of combatting the COVID-19 pandemic, whereas a multifaceted approach considering all critical aspects required to fight the COVID-19 pandemic is needed. As an example, [20] reported that hospital emergency rooms across the globe experience unprecedented floods of people infected with COVID-19 needing urgent treatment. As a result, doctors have to grapple with the problem of patient's triage as they struggle to decide which of the COVID-19 patients require intensive care. For this, the condition of the patient's lungs must be assessed by doctors and nurses. However, doctors and nurses without pulmonary training cannot assess the patient lungs. At the peak of the COVID-19 crisis in Italy, doctors were faced with a serious problem of making decisions on the patient that should be given much needed assistance. Given the difficulties of making triage decisions in COVID-19 cases, a machine learning system could support doctors and nurses making clinical decisions and plays a critical role in the COVID-19 crisis by assisting hospitals in functioning better in keeping COVID-19 patients alive.

A multidimensional framework for automated machine learning solutions to combat COVID-19 on different fronts could provide better means of combating the COVID-19 pandemic, for example, predicting COVID-19 vaccine immunogenicity, COVID-19 contact tracing, monitoring social distancing, and mask wearing, optimizing COVID-19 resource allocation, detecting COVID-19 severity and triage of COVID-19 patients, predicting COVID-19 patients who require a ventilator or predicting those who are beyond medical intervention, predicting COVID-19 mortality, and discovering COVID-19 drugs. All these aspects could be integrated into a single framework to work automatically within a city. Therefore, a smart city could be repurposed to combat the COVID-19 by applying a framework of multiple measures to fight the COVID-19 pandemic. Many technologies are connected to provide smart applications in smart cities, including wireless sensor networks, broadband communications services, sensor devices through the internet, and cloud services. Sullivan [21] predicted that smart cities in 2020 would be embedded with smart structures such as smart healthcare, smart security, smart mobility, smart buildings, smart governance, smart citizens, smart infrastructure, smart technology, smart energy, and smart education.

To the best of our knowledge, our proposed framework stands out from other similar existing frameworks reported in the literature; in that, this study presents the most comprehensive integration of components that are perceived to integrate well with machine learning in order to fight

COVID-19 automatically across multiple dimensions in smart cities. The framework could address the unique challenges in fighting COVID-19, thereby easing the work of healthcare workers in saving lives and providing a guide for real-world execution of a program in smart cities. Yu, Wang, Liu, & Zomaya [22] pointed out that exploring such a theory is critical for providing a guide for effective applications.

The gap in literature motivating the need for this study leads to the following research questions:

(i) Considering the high volume and raw data generated from internet-connected sensory devices, how can machine learning models be adapted and adopted for inferring higher-level information?

(ii) What are the challenges associated with the interoperability of internet of things (IoT) technologies with machine learning algorithms in collecting and analyzing COVID-19 data?

(iii) Could the higher-level information be repurposed in supporting applications (such as detection, prevention, contact tracing, and alert-level dissemination systems) designed to combat the COVID-19 pandemic?

(iv) How can a computational solution based on a robust framework be applied to evaluate the environmental and social weaknesses of a city for containing new COVID-19 outbreaks?

This paper proposes a solution framework integrated with machine learning to combat COVID-19 in smart cities from multiple dimensions. The novelty of the study lies in the essential elements of the framework and how its elements such as physical structures, institutions, policymakers, medical personnel, ICT, IoT, big data, and machine learning algorithms seamlessly integrate to manage all kinds of applications and devices.

The remainder of the paper is organized as follows: Section 2 presents the background information about fundamental concepts considered in the study, Section 3 presents the proposed solution framework for combatting COVID-19 in a smart city from multiple fronts, Section 4 presents the applications of machine learning in fighting COVID-19 pandemic in smart cities, and Section 5 discusses the outcome of the study by outlining case studies on combatting COVID-19 via machine learning in smart cities, before the concluding remarks in Section 6.

2. Preliminaries: Overview of the Novel Coronavirus Diseases, Smart Cities, and Machine Learning

This section presents an overview of the fundamental concepts, components, and related works on which the study is focused. This general overview allows conceptualization of the proposed framework in the subsequent subsections. Our review provides details of the COVID-19 disease and its clinical manifestation and also of the concept of smart cities as they relate to IoT. In addition, the background to machine learning and its associated algorithms is discussed.

2.1. Novel Coronavirus Disease: Background and Primary Clinical Features. In this subsection, an attempt is made to present fundamental knowledge about the disease COVID-19 caused by the novel coronavirus SARS-CoV-2, with background information and clinical features which are relevant for the implementation of the proposed framework.

2.1.1. Background of Novel Coronavirus SARS-CoV-2 and COVID-19 Disease. Out of the seven known coronaviruses, SARS-CoV-2 is the third-most highly pathogenic coronaviruses to have afflicted the human race. As the SARS-CoV-2, the etiological agent of COVID-19 spreads across more than 210 countries and territories, infections, and subsequent fatality rates continue to rise. Accordingly, several preventive and control measures have been adopted to halt the spread of the SARS-CoV-2 and minimize COVID-19-associated death. As of 7 : 30 AM GMT+1, 27th April 2020, there were over 3 million confirmed cases of SARS-CoV-2 infection globally with a case fatality rate (CFR) of around 7.0% (Worldometer, 2020). At that time, European and American countries, with only a few Asian countries, appeared to have the worst CFRs associated with COVID-19, with the least being in Africa, without any categorical explanation for this variation. Several observers have attributed the low incidence rate of COVID-19 in sub-Saharan Africa to underdiagnosis, probably due to inadequate molecular diagnostic capacity, but variation in the genetics, strains, viral protein mutations, and host immune response could have contributed to SARS-CoV-2 virulence and pathogenesis [23].

Although there have been controversies about the origin of SARS-CoV-2, several studies have traced the zoonotic source of this virus to the first patients exposed in a live animal market in Wuhan, China [4]. Subsequently, efforts have been made to search for a reservoir host and intermediate hosts of SARS-CoV-2, from which the infection might have spread to humans. Initially, two snake species were identified in this regard. However, the only consistently identified SARS-CoV-2 reservoirs have been mammals such as bats [4, 24]. In particular, early isolates of SARS-CoV-2 from infected patients in Wuhan showed genomic sequencing of some SARS-CoV-2 isolates to have 88% nucleotide homology with two bat-derived SARS-like coronaviruses [25], thus indicating bats as the most likely reservoir hosts for SARS-CoV-2 [25].

The first principal transmission mode of the SARS-CoV-2 is through droplets emanating from the respiratory system of infected people. These droplets are transferred by coughing and sneezing so that when they come into contact with the respiratory system of uninfected persons, it may cause a COVID-19 infection. These, therefore, form the second principal transmission mode, which is by contact [26]. Although younger persons have exhibited some resistance to the disease due to their strong immune systems, studies have shown that people of all ages are at risk of contracting it. The aged who have had contact with infected persons or surfaces carrying the virus often progress quickly to acute respiratory distress syndrome (ARDS) and multiple organ failures. Infected younger persons may succumb to mild syndromes like fever, fatigue, and dry cough, with only a small percentage of cases degenerating quickly, as seen in the elderly. The disease's propagation level in a city or population is often evaluated using fatality rates and reproduction number (popularly referred to as R_0 value, see Section 2.1.2) and, recently, the index c value [27]. The effect of this propagation has spilled over into social economic problems, which are directly a result of contracting the GDP growth of countries due to lockdown enforced in several countries [28].

2.1.2. Clinical Features of Novel SARS-CoV-2 and COVID-19 Disease. Virologically, SARS-CoV-2 is a single-stranded RNA virus with positive polarity and variable open reading frames (ORFs) [29]. It has been shown that two thirds of the SARS-CoV-2 genome is located within the first ORF, which translates the pp1a and pp1ab polyproteins. These polyproteins encode 16 nonstructural proteins [29], which are the remaining ORFs code viral structural and accessory proteins of SARS-CoV-2. The remaining one third of the genome codes the nucleocapsid (N) protein, spike (S) glycoprotein, matrix (M) protein, and small envelope (E) protein of SARS-CoV-2. Of these four proteins, the S glycoprotein is key because it plays a role in attachment to host cells and the pathogenesis of COVID-19. This protein, alongside the viral RNA-dependent RNA polymerase (RdRP), has largely been utilized in the synthesis of primers and antigens for, respectively, molecular and serological tests of SARS-CoV-2 infection [30].

RNA viruses, including SAR-CoV-2, have high mutation rates, which is significantly correlated with enhanced virulence and evolvability [31]. At the proteomic level, amino acid substitutions have been reported in the NSP2, NSP3, and S proteins [32]. Another study of interest has suggested that NSP2 and NSP3 mutations play a significant role in the virulence and differentiation mechanism of SARS-CoV-2 [33]. Of interest is the mutation in S-protein. This has made scientists explore the possible differences between the host tropism and the transmission rate of SARS-CoV-2. It is worth noting that the NSP2 and NSP3 mutations in SARS-CoV-2 were isolated from many COVID-19 patients in China [33]. These have mutations sparked scientific interest in genomic surveillance of SARS-CoV-2 to determine the correlation between these mutations and virulence diversity, with its implications for reinfection, immunity, and vaccine development [34].

A measure of the transmissibility or infection rate of SARS-CoV-2 can be measured by R_0, which predicts the number of people to whom an infected person could transmit SARS-CoV-2 in a population with no prior immunity to the pathogen. Generally, the higher the R_0, the more contagious the pathogen. An R_0 of <1 means that the outbreak would die out, while $R_0 > 1$ means the infection will continue to spread [35]. Based on available genetic analysis, SARS-CoV-2 is related to SARS-CoV-1, which along with MERS-CoV, is endemic in certain countries; so, R_0 is not very high. However, an early report on mathematical modelling for SARS-CoV-2 revealed an R_0 of 2 to 3 [14], which could explain why SARS-CoV-2 is more contagious than either SARS-CoV-1 or MERS-CoV. This model highlights that a single SARS-CoV-2-infected individual has the ability to infect two to three uninfected persons [36].

Infection by SARS-CoV-1 occurs through contact with respiratory droplets, which are the size of nanoparticles and can contaminate surfaces and hands, and where they remain stable for hours [36]. Hands, therefore, become a mechanical vector and are, thus, a potential site to eliminate the virus and prevent it from invading the body. However, suppose the virus is not eliminated at this stage, in that case, it can move towards its predilection site (i.e., cells of the lungs), where it attaches using its spikes and uses the angiotensin-converting enzyme-2 (ACE-2) as receptors to gain access to epithelial cells of the respiratory tract. At this stage, SARS-CoV-2 compromises innate lung immunity [36]. It then takes advantage of these cells as a replication site. The virus regenerates and sheds by disassembling itself and utilizing the machinery of the alveoli cells, to be precise, the Golgi apparatus, to reproduce, and repackage itself [36].

The SARS-CoV-2 exists so that it can replicate and disrupt the protective function of the ACE-2 receptor, which induces the process of fibrosis (scarring). It has been shown that patients with fatalities associated with SARS-CoV-2 present a characteristic ground glass effect in their lungs, and this sequela impedes efficient oxygenation. As the body tries to compensate for this deficiency, the result is a severe acute respiratory syndrome (SARS), in which it becomes impossible for the respiratory system to make oxygen available to the rest of the body (hypoxia) [36]. This ultimately results in multiple organ failures. Based on available clinical data, those susceptible to developing a severe form of SARS-CoV-2 infection include the elderly (>60 years) and persons with underlying disease conditions (e.g., cardiovascular, metabolic, respiratory, and immunological disorders) [37].

When susceptible individuals get infected by SARS-CoV-2, that person may either remain asymptomatic (no apparent illness) or be symptomatic. If symptomatic, the disease passes through three stages of severity. Patients present with mild clinical symptoms in the early infection stage (Stage I), including dry cough, diarrhea, fever, and headache. This could last for 3 to 5 days. This stage is usually accompanied by lymphopenia (low white blood cell counts), elevated prothrombin time, D-dimer, and a mild increase in lactose dehydrogenase (LDH). Almost all (98%) of SARS-CoV-2-infected patients remain at this stage and eventually recover. However, those with underlying medical disorders may proceed to stage II (Pulmonary Phase), predominantly characterized by shortness of breath and hypoxia (inadequate oxygen supply to the body), which could last from 5 days to 3 weeks. At this stage, patients would display an abnormal chest radiograph, transaminitis, and declined procalcitonin levels. Very few patients (2%) proceed to this severe stage of COVID-19. Stage II (hyper-inflammation phase) is largely characterized by acute respiratory distress syndrome (ARDS), severe inflammatory response syndrome (SIRS), shock, and cardiac failure. The majority of patients who reach this stage eventually die. At this stage, COVID-19 patients experience significantly high blood inflammatory markers such as elevated C-reactive protein (CRP), interleukin-6 (IL-6), D-dimer, and ferritin. In addition, affected patients present with the increased blood level of cardiac markers, especially troponin and N-terminal (NT)-prohormone B-type natriuretic peptide (NT-proBNP) [38].

Diagnostically, the use of viral culture for establishing acute COVID-19 diagnosis is not practicable due to the long turnaround time (3 days) for SARS-CoV-2 to cause obvious cytopathic effects (CPE) on Vero E6 cells. In addition, isolation of SARS-CoV-2 is laborious and requires biosafety level-3 (BSL-3) facilities, which are unavailable in most healthcare centers, especially in developing countries. So far, all available serum antigens (such as the S-glycoprotein) and antibody (IgA, IgM, and IgG) detection tests have not been validated by the WHO. However, it has been suggested that serological assays could assist in analyzing an ongoing SARS-CoV-2 outbreak and retrospective evaluation of the incidence rate of an outbreak [9]. In some instances, where epidemiological data of suspected cases correlates to SARS-CoV-2 infection, the demonstration of fourfold rising antibody titer between acute and convalescent-phase sera could support the diagnosis of COVID-19 when RT-PCR results are negative [9]. In addition, it has been revealed that a significant proportion of COVID-19 patients have tested RT-PCR negative despite having suitable clinical features and radiologic findings that are highly indicative of SARS-CoV-2 infection [39]. In most cases, these are termed false negatives, which could have been due to wrong sampling if SARS-CoV-2 had been present in the lower respiratory tracts rather than in the upper respiratory samples usually collected for laboratory diagnosis. Hence, this difficulty in diagnosis poses a challenge in the proper evaluation of SARS-CoV-2 symptomatic patients [40].

2.2. Rudiments of Smart Cities and Machine Learning. In this section, the discussion of the concept of a smart city, including case studies and the brief explanation of machine learning, is intended to help readers new to these domains comprehend the concepts of smart city and machine learning.

2.2.1. Smart Cities. There is, as yet, no universally accepted standard definition of a "smart city." However, a smart city may be understood as a city that encourages the prudent utilization of quality resource management and the provision of services within a limited time. Information and communication technology (ICT) is one of the major components and an integral element in innovative city projects. The operations in a smart city cannot be achieved with ICT in isolation. This state-of-the-art view of city development has resulted in the new smart city model. The quality and scale of cities have grown significantly as a result of urbanization since the industrial revolution. The expansion in urbanization has prompted many challenges, including [41]

(i) Large scale consumption of resources

(ii) The degradation of the environment

(iii) The unfair widening of the gap between the rich and the poor

The smart city model can help cities achieve sustainability goals, such as high-level efficiency, high economy, an improved standard of living for people, and a beautiful city environment. Many criteria to assess the smartness of a city have been proposed, which include all or some of the following: smart energy production and conservation, smart mobility, smart economy, smart living, ICT economics, smart environment, smart governance, and the connection between the standard of living and smart society [41]. To put a smart city in its proper position, a smart city is the combination of a wide range of services that are required by a city and the need to offer the services in a way that complies with the current administration requirement through the use of state-of-the-art technology.

2.2.2. Case Studies of Smart Cities. To buttress our contention about the potential for smart cities in providing critical support to their population, we briefly review three case studies of the application of IoT technology to a city, namely, Kuala Lumpur, Copenhagen, and Stockholm.

(1) Case Study 1: Kuala Lumpur, Malaysia. In Kuala Lumpur, many projects in relation to a smart city are designed to use the city's resources optimally. For instance, many innovations have been implemented in the transport sector to reduce traffic congestion. The innovations include using smartphones for flight check-in, train and bus schedules being monitored through electronic boards in the city train and bus stations, and smart cards, referred to as "touch and go," and are commonly used to avoid long queues in purchasing bus or train tickets; an application referred to as "GrabCar" is used to book a cap and track its position and the estimated time of arrival to the pick-up position. The weather can be monitored through smartphones, such as showing the daily temperature in different city locations. Nonsmoking areas are also embedded with sensors to trigger an alarm in case of smoking cigarettes in the nonsmoking zone. In addition, many of the vehicles in Kuala Lumpur are hybrid, which enables switching the engine from electric to petrol and back to electric, as needed.

(2) Case Study 2: Copenhagen, Denmark. The Boyd Cohen list of smart cities in Europe ranked Copenhagen in eighth place [42]. For Copenhagen, numerous smart city projects were analyzed from the perspectives of success factors and economics. Copenhagen has the vision of becoming the world pioneer carbon-neutral capital by the year 2025. As such, Copenhagen is presently implementing innovations in the field of transportation, waste management, water supply, heating, and sources of alternative energy to support the 2025 target vision for the city and enhance sustainability. Copenhagen has currently expanded its network of cycle lanes to be embedded in the broad transportation concept to improve traffic flow in the city (Catriona [43]).

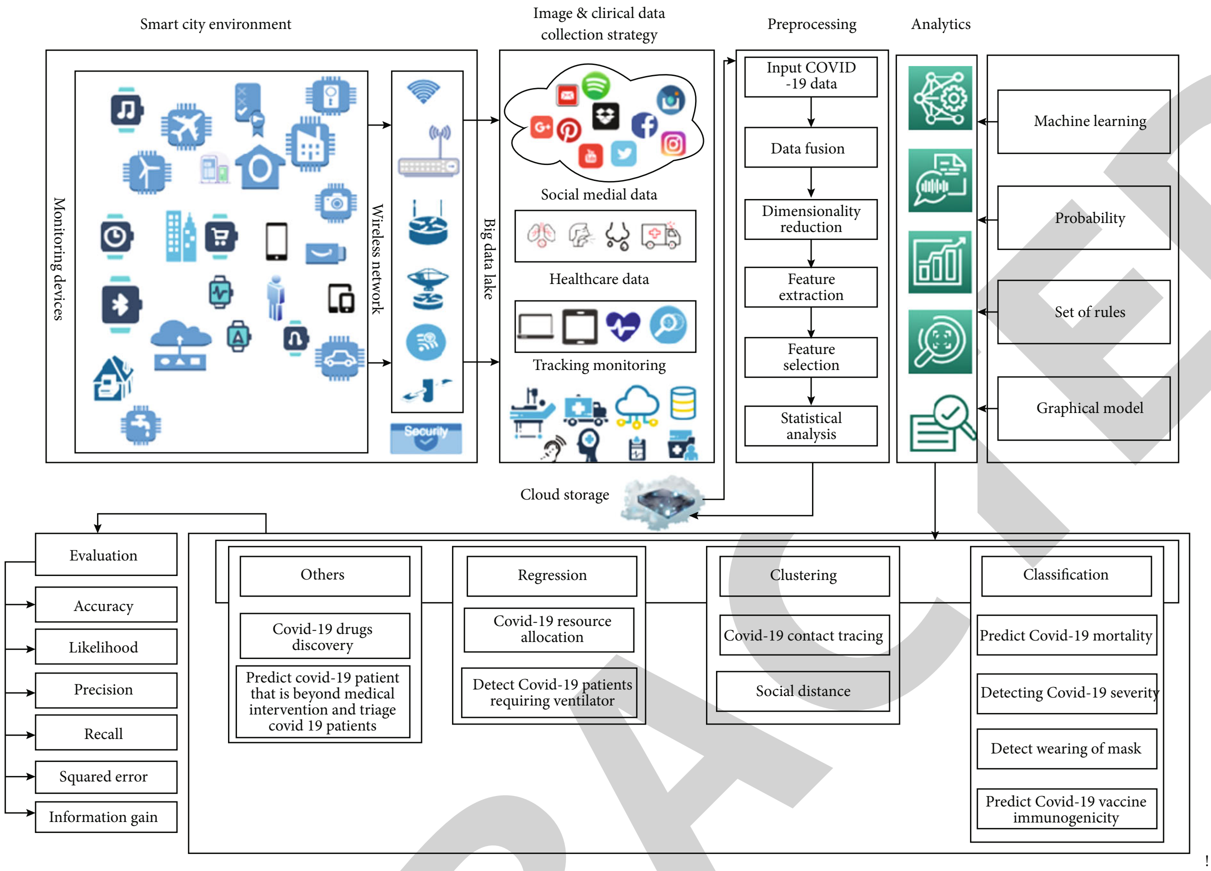


Figure 1: Proposed smart city framework integrated with machine learning for fighting COVID-19.

2.3. Machine Learning. Machine learning is a field of science that centers on how a computer learns from data [44]. According to Portugal, Alencar, & Cowan [45], machine learning is an algorithm that "uses computers to simulate human learning and allows computers to identify and acquire knowledge from the real world, and improve the performance of some tasks based on this new knowledge." Machine learning is a subdiscipline in artificial intelligence and cuts across many fields of studies that correlate with data mining, pattern recognition, computer science (theoretical), and statistics [46]. In statistics, it seeks to determine the relationship that exists in data, whereas in computer science, it emphasizes the effectiveness of the computational algorithm. Machine learning research in computer science examines the algorithm used for the learning to make a prediction based on data. To achieve that, the input data is employed to construct a model so that a data-driven decision can be made with various static program instructions [47]. Machine learning algorithms can be broadly categorized into supervised learning, unsupervised learning, and semisupervised learning.

Supervised learning (input observation mapped with output observation) is learning where the input observation consists of features, and the output observation consists of labels [48]. Thus, it constructs a model by utilizing a labeled dataset as input [49] and produces labeled output data. The primary purpose of supervised learning is to drive a functional correlation from the training data with well-generalized testing data. Some examples of supervised learning algorithms are employed in classification and regression problems, including naïve Bayes, decision tree, and logistic regression. On the other hand, unsupervised learning is a learning algorithm that is employed when there are difficulties in finding the labeled sample, since it does not rely on previous training for mining the data. The primary purpose of unsupervised learning is to find a correlation between the samples behind the observation. One of the notable examples of unsupervised learning is a clustering system. Semisupervised learning is a combination of supervised and unsupervised learning, which uses a small amount of labeled data and a huge amount of unlabeled data [50] during the training process. Information recommendation systems and semisupervised classification are examples of a semisupervised learning algorithm.

The machine (and deep) learning algorithm can be applied in many research fields, including natural language processing [17], medical diagnosis, financial data analysis, bioinformatics, and video surveillance. The following section presents our approach to harmonizing machine learning algorithms and IoT technologies using a novel framework. Furthermore, a detailed discussion on the applicability of the proposed framework in combating the COVID-19 pandemic is presented in Section 4.

3. Proposed Method

The ubiquitous nature of internet-connected sensory devices, which are often capable of generating relevant data for analytics purposes, has motivated the approach promoted in this study. These devices can capture a high volume of structured and unstructured data based on the time and location of the physical world. We argue that intelligently processing these volumes of data requires learnable algorithms that, with minimal human intervention, can derive a pattern sufficient to present higher-level information to support combating COVID-19. Hence, this section presents a novel framework that intelligently allows for the interoperability of IoT concepts and machine learning models.

3.1. The Smart City-Based Machine Learning Framework for Combating COVID-19. To address the multiple dimensional challenges posed by the COVID-19 pandemic, as outlined in Section 1, we propose that a framework is required within the smart city context to allow decision makers to make the crucial decisions on the best ways to combat COVID-19 from multiple dimensions. The framework consists of multiple components, as shown in Figure 1. Each component has a major impact on enhancing the quality of the analytics to combat COVID-19.

3.2. Modules of the Smart City Framework. The smart city-based framework has four core modules, namely, smart city environment, image and clinical collection strategy, image preprocessing and analytics, machine learning models, cloud-based storage, and evaluation strategies. The following subsections provide details of the modules.

3.2.1. Smart Environment. Smart city technologies have recently demonstrated their potential for enhancing citizens' quality of life. Many smart-based technologies have arisen from the adoption of the internet of things (IoT), which has led to the development of intelligent applications such as smart homes, smart grids, smart transportation, smart industry, and smart healthcare. Moreover, recently sensors and video cameras surveillance have become part of smart city monitoring; they can also be used for early detection of a pandemic. During the COVID-19 pandemic, smart technologies could help in tackling the major clinical, social, and economic problems due to the disease. Specifically, health agencies may utilize IoT platforms to access data for monitoring the COVID-19 pandemic. For example, "Worldometer" allows viewing of instant updates about the severity of COVID-19 for the entire world. These updates include daily new cases and deaths due to COVID-19, cumulative numbers of cases and deaths, and distribution of COVID-19 by country [51]. In Figure 1, we demonstrate a smart environment that could be used for healthcare purposes, in which the IoT and various sensors and monitoring devices interact within a limited area to generate data on clinical signs and symptoms. These devices are connected via next-generation wireless connectivity, which can efficiently transfer the collected data to be stored in a big data lake. Big data plays a critical role in smart cities because its ecosystem of data analytics can allow decision makers to decide critically on the best strategy to be developed to combat COVID-19. With big data and with adequate smart city framework implementation, users can be traced at all times with the potential to mitigate any health problems they might encounter during their movement. Therefore, improving the efficiency and effectiveness of a smart city framework would, in turn, improve the lives of the citizens in the smart cities. A literature review was conducted by Al-Turjman [52], which presents comprehensive background about 5G standards and their specific applications for the IoT and an overview of recent developments in use of smartphone sensors that could contribute to a scalable operation in smart social spaces.

3.2.2. Image and Clinical Data Collection Strategy. The image and clinical data generated in real-time for smart cities can be subjected to big data processing to understand healthcare trends, model risk associations, and predict outcomes. The government authorities can use the results of the big data lake with private/public healthcare providers to improve healthcare services to the citizens; this process would continue until the government and healthcare services providers satisfy the citizens living in the smart cities. The various means of collecting data on a large scale include social media platforms such as Facebook, Twitter, Google+, Instagram, healthcare services data collected during diagnosis and treatment, and tracking of monitoring devices such as GPS and vehicle tracking systems, smartwatches, and sensors. All collected data would be integrated and stored in a single location within the smart city to be accessed by the authorized entities. The technologies that make storage of such dad possible are Hadoop distributed file system (HDFS) and NoSQL, in which both structured and unstructured data can be stored and processed. Balduini et al. [53] proposed a new conceptual framework that uses a variation of big data sources. The unified approach in their framework uses spatial and temporal analysis on a heterogeneous stream of data. Their results show the proposed framework's generality, feasibility, and effectiveness across many cases and examples obtained from real-world requirements using data collected in many cities.

3.2.3. Preprocessing. In order to provide more accurate and better input to achieve more reliable results in the detection and prevention of COVID-19 cases, data preprocessing is considered an important first stage. The first step in preprocessing is to extract all the relevant COVID-19 data from storage. The second step is to perform data fusion, whereby the collected data are integrated to produce more consistent, accurate, and useful information. The third step during preprocessing of COVID-19 data is to reduce dimensionality, in which the number of variables is reduced by extracting a set of main variables. The fourth and fifth steps focus on feature extraction and selection. The last two steps are very important in filtering irrelevant or redundant features from the selected datasets. The last step involves a basic statistical analysis of COVID-19 data in order to interpret the data before intelligence-based algorithms are applied.

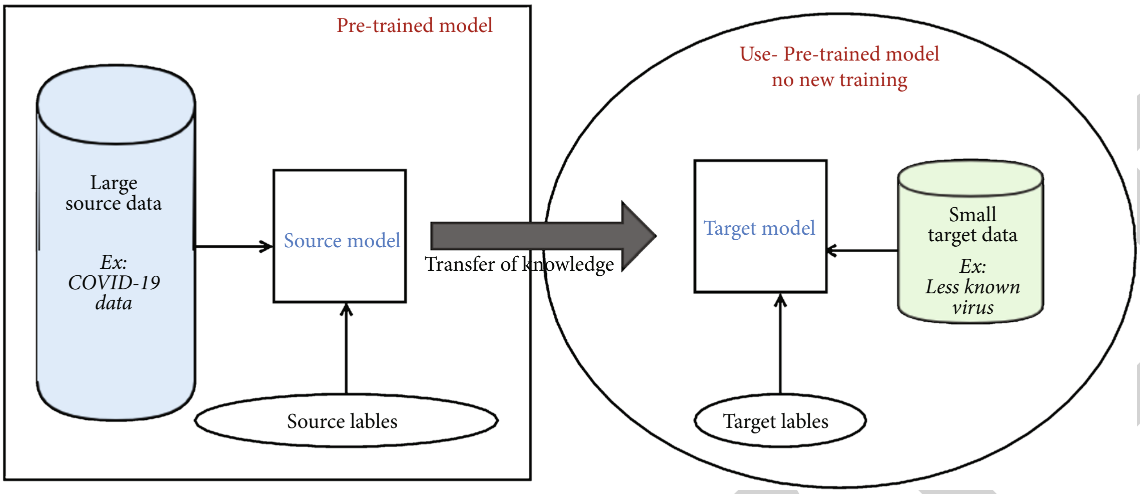


Figure 2: Typical transfer learning model for predicting future pandemics.

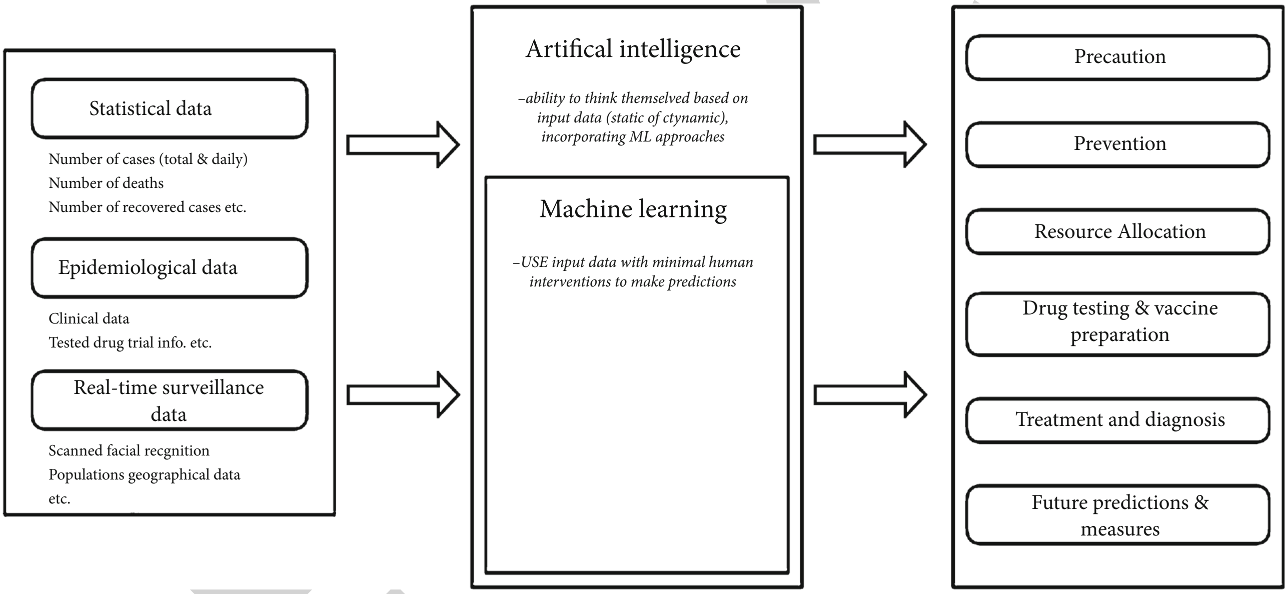

Statistical data
Number of cases (total & daily)
Number of deaths
Number of recovered cases etc.
Epidemiological data
Clinical data
Tested drug trial info. etc.
Real-time surveillance data
Scanned facial recgnition
Populations geographical data
etc.
Artifical intelligence
–ability to think themselved based on input data (static of ctynamic), incorporating ML approaches
Machine learning
–USE input data with minimal human interventions to make predictions
Precaution
Prevention
Resource Allocation
Drug testing & vaccine preparation
Treatment and diagnosis
Future predictions & measures

Figure 3: Application and flow of machine learning and other subdomains of AI in combatting COVID-19.

3.2.4. Analytics. Recently, image processing in healthcare using convolutional neural networks has become a significant approach for handling large quantities of images generated from smart cities. Allam and Jones [54] discuss the universal data sharing standards that are coupled with AI to benefit urban health monitoring and management. Our proposed framework can incorporate various machine learning and deep learning algorithms to develop the analytical model. These algorithms range from traditional shallow approaches such as neural networks, decision tree, naïve Bayes, and K-nearest neighbor. These algorithms can be applied to run on COVID-19, a dataset in applications that help combat COVID-19 in hospitals, in smart cities, and across the world. These applications include detecting and preventing the spread of COVID-19, forecasting the next epidemic, diagnosis of cases, monitoring COVID-19 patients, tracking potential patients, suggesting methods for vaccine development, helping in COVID-19 drug discovery, and together providing a better understanding of the effect of the COVID-19 virus in smart cities.

(1) Social Media Information Verification. The COVID-19 pandemic has brought associated challenges of fake news, including conspiracy theories. Since the COVID-19 pandemic started, there has much fake news regarding its origin, cures, mode of spread, treatment, and many other myths. This is especially prevalent on social media platforms such as Facebook, Twitter, Instagram, and YouTube. In a smart city, citizens would voice their opinions on social media regarding COVID-19 to generate unstructured data. It is reported by Obeidat [55] that no systematic quantitative study has been conducted to ascertain the magnitude of the problem of myths perpetuated on social media around COVID-19, but certainly, the figures for misinformation

about COVID-19 are significant. The fake news regarding COVID-19 can come in the form of manipulated content, misleading content, satire, false context, malicious accounts, fabricated content, false connections, and imposter content. Therefore, machine learning or deep learning algorithms can be applied to detect fake news regarding COVID-19 on social media and alert citizens living in smart cities.

(2) Prediction of Future Pandemic. Although being the worst pandemic in recent times, this pandemic has, nevertheless, come in the time of the digital age. Therefore, every aspect of the analysis can now be captured, including at a macrolevel, logistically, and biologically, in terms of data; this will definitely be fruitful for predicting the behavior of this new pandemic or of unknown future pandemics. For example, with the help of the machine learning approach (random forest), Eng, Tong, & Tan [56] could predict possible zoonotic strains of influenza, i.e., some viruses that usually only affected animals but might also be dangerous to humans. This, therefore, implies that machine learning could help to predict future pandemics arising from any species. The only limitation is that the data could be from a different domain, e.g., the source of COVID-19 is possibly from “bats;”,so, pandemic sources may be different from those encountered in the past (for instance, having a different genome structure, etc.). Further, traditional machine learning requires the data distribution to be from the same domains in training and testing. However, transfer learning (TL)—a type of machine learning—can effectively handle situations where training and testing data might be from different data distributions. That is, the knowledge learned from the past pandemics could be used in future situations with a new domain, even with smaller amounts of data. Such a scenario is shown in Figure 2, where the pretrained model from the current COVID-19 pandemic (with large data and labels) could be used, with significantly less data and labels, to predict future pandemics, prepare the smart city for that situation, and quickly help address the spread of the disease.

The modules of the smart city-based framework for combating COVID-19 shown in Figure 2 present very promising applicability; this study further provides a perspective on how this is achievable. The following section is focused on detailing this aspect.

4. Perspective on the Applicability of the Proposed Framework

As an interconnected urban society, the smart city implies collecting data every moment from several embedded devices, which means that smart cities can work effectively with machine learning approaches during this COVID-19 pandemic. Machine learning techniques are dependent on data for better learning and predictive models, through which they can bring out some intrinsic and valuable insights to help the decision makers in smart cities take preventive measures during the COVID-19 pandemic. Different machine learning techniques operate with other fields of artificial intelligence (AI), which gives the model its ability to provide a rich self-learning platform. It is important to discuss the role of AI and machine learning in combatting COVID-19, because the data availability is limited, and we have to deal with real-time data streaming. Thus, the significance of self-learning systems becomes much more desirable in smart cities. Figure 3 demonstrates the overall flow of how AI and machine learning approaches can help in fighting the COVID-19 pandemic in smart cities.

As shown in Figure 3, several types of data are generated from the information and communication technology equipment embedded in smart cities. These are as follows:

(i) The statistical data that usually contains the cumulative daily number of identified cases, number of new positive cases, number of deaths, number of recovered cases, etc. would help predict future cases to prepare for emergencies

(ii) The epidemiological data primarily concerns all the clinical patient test data, data relating to tests on different medications, various drug trials, patients' medical histories, patients' responses to different medications, etc.

(iii) The real-time surveillance data generated from sensors and cameras in the smart cities would also be helpful to track and prevent the spread of COVID-19. For example, one of the initial identifiers of COVID-19 is based on symptoms of fever; so, body temperature from facial recognition and other personal information can be monitored

The data is processed and analyzed through machine learning approaches for extracting insights in various applications. The applications of machine learning in different aspects for combatting COVID-19 are discussed in the following subsections.

4.1. Prevention and Precaution. Based on the statistical data, the machine learning model can be used to predict the nature of the identified cases to take better preventive measures. During the situation of a pandemic, there may be a degree of chaos. The requirement for rapid and large-scale testing of individuals is very challenging. So, rather than going door to door to each patient, a faster approach would be more acceptable, even if less accurate. Machine learning may help in quickly diagnosing the patients in the smart cities as follows:

(i) Facial recognition with the help of sensors and cameras to scan the patients for body temperature and personal information so that if the particular patient is positive, then their nearby individuals can be tested and alerted to their status

(ii) Helping patients get information and create self-awareness with the AI-powered chatbots because the medical professional might find it impossible to address these queries during the COVID-19 pandemic because of the exceptionally high number of patients they must help

(iii) Using the data from smartphones and wearable smartwatches to monitor the citizens' heart rate and daily activity

Although predictions based on the statistical data may not be 100% accurate, they can nevertheless enable the decision makers in smart cities to institute some preventive and proactive measures.

4.2. Prediction Models. Medical science (especially dermatology) was one of the real-world fields where AI and machine learning approaches were successfully implemented. Computer vision and machine learning prediction models can identify patients' most common dermatological diseases simply by learning from images. In the case of COVID-19, based on some set of crucial features (set of symptoms), machine learning approaches can help in identifying and predicting the following:

(i) A person infected with COVID-19

(ii) A positively diagnosed COVID-19 patient who needs to be hospitalized

(iii) According to the range of treatments available, the chances of a COVID-19 patient being successfully cured or dying

Pourhomayoun and Shakibi [57] used machine learning techniques to predict the mortality rate of patients affected by COVID-19. They used machine learning algorithms such as random forest, logistic regression, decision tree, support vector machines, and artificial neural networks to give up to 93% total accuracy in predicting the mortality rate. Moreover, the study also used machine learning models to extract the essential and unique symptoms and features to detect the virus.

4.2.1. Prediction of COVID-19 Pandemic. Different studies have been conducted to predict the likely occurrence of the COVID-19 pandemic [58]. For instance, Ndiaye, Tendeng, and Seck [59] conducted a global prediction study on the COVID-19 pandemic between January and April 2020. The study employed prophet [60], a tool for predicting time series data; it depends on the additive model that fits real nonlinear trends with daily, weekly, and annual seasonality and holiday effects. Four countries of Italy, China, Senegal, and Iran were selected as case studies for the research. However, the predictive performance of the study showed that the COVID-19 pandemic in countries like China could be optimistically estimated to end in a few weeks. In another perspective, Wang and Wong [26] proposed COVID-Net, which uses a convolutional neural network design to identify COVID-19 cases from chest X-ray (CXR) images. The study utilized the CXR dataset, which comprised 13,800 chest radiography images obtained from 13,725 patients from three public datasets. The experimental analysis shows that the proposed COVID-Net attained a predictive accuracy of 92.6% on the test data, indicating the importance of combining human and machine collaboratively in the design strategy for building modified deep neural network architectures faster fitted around the data, task and working requirements. In another study, Yang et al. [61] proposed a modified susceptible, exposed infections removed (SEIR) model and AI prediction of COVID-19 pandemics. The study employed the most up to date COVID-19 epidemiological data together with population migration data obtained prior to and after 23rd January 2020 into the SEIR model. In addition, a machine learning approach was employed to train on the 2003 SARS data for the pandemic prediction. The predictive result of the study shows that the pandemic of China was expected to be at peak by late February and then show a gradual decline by the end of April. However, the COVID-19 cases would have risen higher than expected in mainland China should the implementation of the proposed model have been delayed for as little as five days.

In their study on the outbreak of COVID-19, Gozes et al. [16] developed an artificial intelligence-based automated computer tomography (CT) image analysis tool using a deep learning approach for the detection, tracking, and quantification of COVID-19, which could distinguish patients infected with COVID-19 and those who were not. The study utilized various global datasets, which included those for Chinese disease-infected areas. Various retrospective deep learning experiments were performed to analyze the system performance in identifying speculated thoracic computer tomography features of the COVID-19 for the evaluation of disease evolution in each patient. One hundred and fifty-seven (157) patients were selected from the US and China for the testing sets. The model's classification performance attained 0.996% AUC, 92.2% specificity, and 98.2% sensitivity on Chinese control and infected patient datasets. This shows that the proposed model could attain high predictive performance in identifying, tracking, and quantifying the COVID-19 cases. Similarly, Narin, Kaya, and Pamuk [62], in their proposal for automatic prediction of COVID-19, employed a deep convolutional neural network built on chest X-ray image and a pretrained transfer model that includes the InceptionV3, Inception-ResNetV2, and ResNet50 models to attain higher predictive performance with a small size of X-ray dataset. The experimental results attained an optimum result of 98% accuracy on the ResNet50 pretrained model among the three selected models. The research results show that the model can assist doctors with decision-making in clinical practice as it uses transfer learning to detect the early stage of COVID-19 in infected patients.

4.2.2. Forecasting of Mortality. Since the outbreak of the COVID-19 pandemic in Wuhan city, China, in December 2019 to the time of the study, the number of confirmed deaths has risen to over 115,000, which indicates a weekly doubling of the number of deaths [63]. There appear to have been discrepancies between the mortality statistics and reported cases, which may have resulted from test policies. Thus, the daily report of deaths on COVID-19 has been at variance with the actual deaths over time. Accurate death rate estimation is important as it is a key factor in deciding whether a highly infectious disease should be a public

Table 1: Summary of the application of machine learning in drugs discovery.

Reference	Study approach	Advantage	Challenge
[67]	Deep learning-based computational drug discovery	The generated molecules could serve as a good drug potential for SARS CoV and COVID-19. The results also show that a deep learning model is important in testing the current compound and obtaining a new molecule for use in COVID-19 drugs	By utilizing limited data for model training, there could be a degree of error due to inaccuracies in parameter values
[68]	Target-specific and selective drug design for COVID-19 using a deep generative model	The framework can take care of candidate generations for multiple targets by employing trained mode. Besides, it can also permit explicit accounting for offtarget selectivity	Large scale data size is required for an effective study

concern. Consequently, there is a need for reliable estimates of numbers for mortality from COVID-19, the date for the peak of deaths, and the period of highest mortality, which all assist decision makers in responding to the present and future pandemics.

A statistical model, known as Global-19, was developed by Brown et al. [63], which gives an estimate of mortality trends between 12th April 2020 and 1st October 2020 for 12 countries (USA, China (Hubei), Italy, Spain, France, UK, Belgium, Iran, Netherlands, Germany, Canada, and Switzerland). The mortality data were collected from the WHO daily reports for each country and some other online data. Similarly, Wang et al. [26] employed the patient information-based algorithm (PIBA) to determine the mortality rate for the COVID-19 pandemic in real-time with forecasts of future deaths. The data was collected from three public sites for COVID-19 patients in Wuhan, China. The data consisted of daily numbers of patients newly infected with COVID-19, the patients that had died of the infection, the patients in a critical condition who had been admitted to an intensive care unit (ICU), and people who had been in close contact with the source of infection. The findings show that the average time between the beginning of the infection to the time of death was 13 days. This prediction is based on the data collected from Wuhan, which was the first city with confirmed records of deaths related to the COVID-19 pandemic.

4.3. Resource Allocation. Resources required to manage the COVID-19 pandemic become scarce due to the very high number of people needing them. These resources include ventilators, masks, testing kits, personal protection equipment, and sanitizers. The problem of resource allocation is an NP-Hard problem, and it is impossible to solve in a polynomial time. Considering the urgency of the emergency created by the COVID-19 pandemic, machine learning can be very beneficial in predicting the best allocation of resources using linear and logistic regression. The machine learning model may provide feasibility and close to optimal resource allocation even on a small training dataset (as is the case for COVID-19).

4.4. Vaccine Development. The process of discovering a new vaccine based on the available clinical data could take a long time. But with the help of machine learning approaches, the overall process can be reduced significantly without sacrificing the quality of the vaccine. For example, Ekins et al. [64] report on the use of the Bayesian machine learning model in a study to develop a vaccine for Ebola. Also, Zhang et al. [65] also used the random forest algorithm to improve the accuracy of the scores while working on the H7N9 virus. Currently, there is much effort in the scientific community applying machine learning to search for a design for the COVID-19 vaccine. Gonzalez-Dias et al. [66] reported on the stages of using machine learning to predict vaccine immunogenicity and reactogenicity signatures. The stages involve data preparation, vaccines, and relevant gene selection, selecting the suitable machine learning algorithm for modeling and performance evaluation of the predictive model.

4.5. Drug Discovery for COVID-19. Since the start of the COVID-19 pandemic, it has become necessary to identify drugs that can be employed to treat the disease. In this regard, machine learning can help identify existing drugs that may be effective in treating COVID-19. Machine learning can learn from drug and protein structures and predict their interaction to warrant clinical studies. Various approaches have been used to find the right drugs, either by repurposing existing ones (therapeutic) or discovering a new one. The application of machine learning and the development of the new models have made researchers focus on the application of machine and deep learning models to discover drugs that could bring a cure for COVID-19. A review of some of the studies that applied the machine or deep learning approach for drug discovery and vaccine for COVID-19 is summarized in Table 1.

5. Results and Discussion

In this section, we present the study's findings and reinforce the importance of the proposed framework by illustrating its use in cases where machine learning techniques are applied to help in combating COVID-19 in smart cities.

The key outcome of this study is the proposed framework and its wide-ranging applicability to the advancement of global efforts in curbing the devastating effects of COVID-19. The findings from this study have showed that handcrafted and manually driven mechanisms for managing COVID-19 remain ineffective as the outbreak has been overwhelming, defying such mechanisms. The pervasive nature of data-driven smart devices and applications continue to provide an intelligent and scalable solution for achieving a secure city in the event of local epidemics or global pandemics such as COVID-19. This finding is further supported by Shorfuzzaman et al. [69], who argued that mass video surveillance has great potential for managing social distancing as a panacea for propagating the disease. Their study echoes the aim of this study, which is to demonstrate that data-driven driven machine learning frameworks, dependent or sensory devices, will allow for their artificial deployment in smart cities for curbing COVID-19. This study has also shown that, in addition to the benefit of managing social distancing through surveillance, such video files could also provide contact tracing applications with input in chronicling events and persons needing tracing. Confirming the methods used in this study and its findings concerning the profitability of machine learning to cities overwhelmed by COVID-19, another related study by Allam et al. [70] noted that 6G technology, including digital twins and immersive realities (XR) would support the socioeconomic position of its population.

Countries across the globe have had their share of first, second, and even third waves of the pandemic and are beginning to look towards a postpandemic era with the digitalization of the city system to manage a future outbreak. Again, this concretizes the critique presented in this study, highlighting the need to “smart up” the systems that drive cities. The study of Graziano et al. [71] supports the potency of the framework proposed in this study. The authors noted that governments are now considering a more inclusive techled urban development, in other words, smart cities. We argue that in developing such a techurban settlement, the outcome of this study presents authorities with a machine learning-driven framework for curtailing and managing any subsequent waves of the disease. For instance, in leveraging the real-time data collection through sensory devices in smart cities, studies using machine and deep learning algorithms have built temporal learning algorithms. Sun et al. [72] have supported this claim by applying deep learning, a submodel of the machine learning model, in projecting the level of COVID-19 disease outbreak using temporal data that are richly generated in a smart city such as that driven by our proposed framework. On the imaging and preprocessing case proposed by the framework in this study, Lassau et al. [73] showed that by intelligently integrating the performance of deep learning models with related variables (e.g. both clinical and biological), the severity of the disease in patients can be predicted ahead of time. Again, the adoption of the framework proposed in this study presents city officials with a potent tool for managing future events. Furthermore, the works of Shorten et al. [74] and Pan et al. [75] demonstrate the importance of integrating machine learning algorithms in the city-wide management of COVID-19, as can be achieved by adopting the framework proposed in this study.

The user cases discussed below can help readers understand exactly how machine learning could assist in fighting the COVID-19 pandemic. As such, other nations can share their expertise in fighting the COVID-19 by applying the machine learning approaches.

5.1. Case Study: New York City. In New York City, there are heavy numbers of cases of COVID-19 patients and those exhibiting the symptoms. The medical staff in New York hospitals is overwhelmed by the extremely high numbers of COVID-19 patients. As a result, the medical staffs face difficulties in deciding which of the COVID-19 patients requires emergency treatment and which patients might be beyond medical intervention. To speed up such decision making for the medical staff in New York hospitals, a machine learning system has been developed through training of the system to provide clinical decisions that support the triage of patients. This system is now used in hospitals to assist with clinical decisions [20].

5.2. Case Study: China. China is well known for the massive amount of data generated from its citizens. China installed a network of over 200 million surveillance cameras across the country. In addition to these video surveillance cameras, biometric scanners were installed in the doorways of residential complexes. As a form of registration, any resident or person that is leaving the residential building must present his or her face to the biometric scanner. After that, the embedded intelligent systems process the data and track the person location through video surveillance. All the information is stored in a central database in which the machine learning algorithms run the data to determine the possible social interactions of the person when they leave the residential building [76].

5.3. Case Study: Canada. Human movement across the globe contributed significantly to the COVID-19 pandemic spreading throughout the world. BlueDot, based in Canada, applied machine learning and natural language processing to track, recognize, and report the spread of COVID-19, which they accomplished faster than the WHO or Centre for Disease Control and Prevention (CDCP) in the United States of America. It is projected that this technology, which is based on machine learning and natural language processing, can be leveraged in the future to predict zoonotic infection risk to humans using climate and human activities as variables. The prediction of individual risk profiles using the data extracted from social media such as family history and lifestyle as well as clinical, personal, and travel data can provide precise and accurate predictions. However, such technology can trigger privacy concerns [55]. Similarly, virtual healthcare assistant is a multilingual healthcare agent that has been developed based on natural language processing. It is a question–answering system that responds to questions related to COVID-19; it delivers trustworthy information on COVID-19 guidelines, protection measures, symptom monitoring and checking, and provides advice to individuals on their need for screening in the hospital or self-isolation. The virtual healthcare assistant was developed by a Canada-based organization [55].

5.4. Case Study: United States of America. In the United States, many medical centres are modifying their existing intelligent systems that were purposely meant to predict patients' illnesses. These intelligent systems are now being modified to predict specific types of COVID-19 outcomes, like the need for intubation. The intelligent systems are trained to learn the illness patterns by feeding the system with thousands of patient records as training data. However, there is insufficient data to build an entirely new intelligent system for predicting COVID-19. Therefore, researchers are assessing the existing tools with the aim of customizing them to help in the fight against the COVID-19 pandemic [20].

6. Conclusions

This paper proposes a solution framework based on machine learning for integration of the fight against COVID-19 in smart cities, from different viewpoints such as predicting COVID-19 vaccine immunogenicity, detecting COVID-19 severity, predicting COVID-19 mortality, COVID-19 resource allocation, COVID-19 drug discovery, COVID-19 contact tracing, detecting social distancing and wearing of masks, triage of COVID-19 patients, identifying COVID-19 patients requiring a ventilator, and predicting which COVID-19 patients might be beyond medical intervention. The paper presented a comprehensive guide for implementing the machine learning framework in smart cities. The solution framework can potentially automate the means of fighting the COVID-19 pandemic in smart cities from multiple dimensions. This would ease the fatigue of the healthcare workers due to the very high number of COVID-19 patients requiring medical attention simultaneously and provide widespread access to a quality healthcare system. In addition, this study foresees that the proposed smart city machine learning-based framework used in combatting COVID-19 will be an essential guide for the research community in developing more compartmentalized forecasting and analyzing tools, with the prospect of mitigating the spread of the COVID-19 pandemic and the occurrences of any similar future disease pandemics. The limitation of the study is inherent in the pieces of the framework being still in their design form and so not yet pieced together in implementation form. However, as seen in several studies, components of the framework have already been implemented, which confirms its applicability. In future work, it will be interesting to see the real-world application of the proposed framework and further investigate the practicality of the model and its efficiency in smart city environments. This deployment of the proposed framework should generate policies to allow for effective integration into the city's existing social and health systems.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] X. Duan, X. Guo, and J. Qiang, "A retrospective study of the initial 25 COVID-19 patients in Luoyang, China," *Japanese Journal of Radiology*, vol. 38, no. 7, pp. 683–690, 2020.

[2] Y. Dong, X. Mo, Y. Hu et al., "Epidemiological characteristics of 2143 pediatric patients with 2019 coronavirus disease in China," *Pediatrics*, vol. 145, no. 6, article e20200702, 2020.

[3] Q. X. Zhang, Q. Xu, Y. Y. Chen et al., "Clinical characteristics of 41 patients with pneumonia due to 2019 novel coronavirus disease (COVID-19) in Jilin, China," *BMC Infectious Diseases*, vol. 20, no. 1, article 5677, p. 961, 2020.

[4] P. Hao, W. Zhong, S. Song, S. Fan, and X. Li, "Is SARS-CoV-2 originated from laboratory? A rebuttal to the claim of formation via laboratory recombination," *Emerging microbes & infections*, vol. 9, no. 1, pp. 545–547, 2020.

[5] ECDC, "European Centre for Disease Prevention and Control," 2020, https://www.ecdc.europa.eu/en/covid-19/latest-evidence.

[6] H. A. Rothan and S. N. Byrareddy, "The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak," *Journal of Autoimmunity*, vol. 109, article 102433, 2020.

[7] L. Lin, L. Lu, W. Cao, and T. Li, "Hypothesis for potential pathogenesis of SARS-CoV-2 infection–a review of immune changes in patients with viral pneumonia," *Emerging microbes & infections*, vol. 9, no. 1, pp. 727–732, 2020.

[8] I. Chakraborty and P. Maity, "COVID-19 outbreak: migration, effects on society, global environment and prevention," *Science of the Total Environment*, vol. 728, article 138882, 2020.

[9] WHO, "MINIMUM REQUIREMENTSfor infection prevention and control programmes," 2019, https://www.who.int/infection-prevention/publications/MinReq-Manual_2019.pdf?ua=1.

[10] A. Alimadadi, S. Aryal, I. Manandhar, P. B. Munroe, B. Joe, and X. Cheng, *Artificial Intelligence and Machine Learning to Fight COVID-19*, American Physiological Society, Bethesda, MD, 2020.

[11] Y. Ge, T. Tian, S. Huang et al., "A data-driven drug repositioning framework discovered a potential therapeutic agent targeting COVID-19," 2020, https://www.biorxiv.org/content/10.1101/2020.03.11.986836v1.

[12] H. C. Metsky, C. A. Freije, T. S. Kosoko-Thoroddsen, P. C. Sabeti, and C. Myhrvold, "CRISPR-based COVID-19 surveillance using a genomically-comprehensive machine learning approach," 2020, https://www.biorxiv.org/content/10.1101/2020.02.26.967026v1.

[13] R. Pandey, V. Gautam, K. Bhagat, and T. Sethi, "A machine learning application for raising wash awareness in the times of covid-19 pandemic," 2020, https://arxiv.org/abs/2003.07074.

[14] L. Yan, H. T. Zhang, Y. Xiao et al., "Prediction of survival for severe Covid-19 patients with three clinical features: development of a machine learning-based prognostic model with clinical data in Wuhan," 2020, https://www.medrxiv.org/content/10.1101/2020.02.27.20028027v2.

[15] G. S. Randhawa, M. P. Soltysiak, H. el Roz, C. P. de Souza, K. A. Hill, and L. Kari, "Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study," *PLoS One*, vol. 15, no. 4, article e0232391, 2020.

[16] O. Gozes, M. Frid-Adar, H. Greenspan et al., "Rapid ai development cycle for the coronavirus (covid-19) pandemic: initial results for automated detection & patient monitoring using deep learning ct image analysis," 2020, https://arxiv.org/abs/2003.05037.

[17] O. N. Oyelade and A. E. Ezugwu, "COVID19: a natural language processing and ontology oriented temporal case-based

framework for early detection and diagnosis of novel coronavirus," 2020, https://www.preprints.org/manuscript/202005.0171.

[18] O. N. Oyelade, A. E. Ezugwu, and H. Chiroma, "CovFrameNet: An Enhanced Deep Learning Framework for COVID-19 Detection," *IEEE Access*, vol. 9, pp. 77905–77919, 2021.

[19] A. Zhavoronkov, V. Aladinskiy, A. Zhebrak et al., *Potential COVID-2019 3c-like protease inhibitors designed using generative deep learning approaches*, vol. 307, article E1, Insilico Medicine Hong Kong Ltd A, 2020.

[20] I. Spectrum, "AI can help hospitals triage COVID-19 patients," 2020, https://spectrum.ieee.org/the-human-os/artificial-intelligence/medical-ai.

[21] F. Sullivan, *Strategic opportunity analysis of the global smart city market smart city market is likely to be worth a cumulative $1.565 trillion by 2020*, vol. 19, The Growth Partnership Company, 2008.

[22] S. Yu, C. Wang, K. Liu, and A. Y. Zomaya, "Editorial for IEEE access special section on theoretical foundations for big data applications: challenges and opportunities," *IEEE Access*, vol. 4, pp. 5730–5732, 2016.

[23] L. van Dorp, M. Acman, D. Richard et al., "Emergence of genomic diversity and recurrent mutations in SARS-CoV-2," *Infection, Genetics and Evolution*, vol. 83, article 104351, 2020.

[24] M. Bassetti, A. Vena, and D. R. Giacobbe, "The novel Chinese coronavirus (2019-nCoV) infections: challenges for fighting the storm," *European Journal of Clinical Investigation*, vol. 50, no. 3, article e13209, 2020.

[25] R. Lu, X. Zhao, J. Li et al., "Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding," *The Lancet*, vol. 395, no. 10224, pp. 565–574, 2020.

[26] L. Wang and A. Wong, "COVID-net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest radiography images," 2020, https://arxiv.org/abs/2003.09871.

[27] M. Coccia, "An index to quantify environmental risk of exposure to future epidemics of the COVID-19 and similar viral agents: theory and practice," *Environmental Research*, vol. 191, article 110155, 2020.

[28] M. Coccia, "Effects of the spread of COVID-19 on public health of polluted cities: results of the first wave for explaining the dejà vu in the second wave of COVID-19 pandemic and epidemics of future vital agents," *Environmental Science and Pollution Research*, vol. 28, no. 15, article 11662, pp. 19147–19154, 2021.

[29] J. Cui, F. Li, and Z.-L. Shi, "Origin and evolution of pathogenic coronaviruses," *Nature Reviews Microbiology*, vol. 17, no. 3, pp. 181–192, 2019.

[30] WHO, "Coronavirus Disease (COVID-2019) Situation Reports. Geneva," 2020, https://www.who.int/emergencies/diseases/novel-coronavirus-2019?gclid=Cj0KCQjwnv71BRCOARIsAIkxW9HqYxLRYA3R2Sz_jVGDPecH0r-RV5Fe7dd4XAO5ObXn9kYYivrYKnIaApArEALw_wcB.

[31] S. Duffy, "Why are RNA virus mutation rates so damn high?," *PLoS Biology*, vol. 16, no. 8, article e3000003, 2018.

[32] Z. Wu and J. M. McGoogan, "Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China," *JAMA*, vol. 323, no. 13, pp. 1239–1242, 2020.

[33] S. Angeletti, D. Benvenuto, M. Bianchi, M. Giovanetti, S. Pascarella, and M. Ciccozzi, "COVID-2019: the role of the nsp2 and nsp3 in its pathogenesis," *Journal of medical virology*, vol. 92, no. 6, pp. 584–588, 2020.

[34] C. Liu, Q. Zhou, Y. Li et al., *Research and Development on Therapeutic Agents and Vaccines for COVID-19 and Related Human Coronavirus Diseases*, ACS Publications, 2020.

[35] E. Prompetchara, C. Ketloy, and T. Palaga, "Immune responses in COVID-19 and potential vaccines: lessons learned from SARS and MERS epidemic," *Asian Pacific Journal of Allergy and Immunology*, vol. 38, no. 1, pp. 1–9, 2020.

[36] G. Kampf, D. Todt, S. Pfaender, and E. Steinmann, "Persistence of coronaviruses on inanimate surfaces and their inactivation with biocidal agents," *Journal of Hospital Infection*, vol. 104, no. 3, pp. 246–251, 2020.

[37] G. Lippi and M. Plebani, "Laboratory abnormalities in patients with COVID-2019 infection," *Clinical chemistry and laboratory medicine (CCLM)*, vol. 58, no. 7, pp. 1131–1134, 2020.

[38] L. Gao, D. Jiang, X. S. Wen et al., "Prognostic value of NT-proBNP in patients with severe COVID-19," *Respiratory Research*, vol. 21, no. 1, pp. 83–87, 2020.

[39] S. Y. Xiao, Y. Wu, and H. Liu, "Evolving status of the 2019 novel coronavirus infection: proposal of conventional serologic assays for disease diagnosis and infection monitoring," *Journal of Medical Virology*, vol. 92, no. 5, pp. 464–467, 2020.

[40] X. Li, M. Geng, Y. Peng, L. Meng, and S. Lu, "Molecular Immune Pathogenesis and Diagnosis of COVID-19," *Journal of Pharmaceutical Analysis*, vol. 10, no. 2, pp. 102–108, 2020.

[41] E.-C. Smart and G. C. Cooperation, *Comparative Study of Smart Cities in Europe and China*, Current Chinese Economic Report Series, Springer, 2014.

[42] B. Cohen, *Copenhagen Is the 8th Smartest City in the World*, Copenhagen Capacity, 2012, https://www.copcap.com/news/copenhagen-is-the-8th-smartest-city-in-the-world/.

[43] C. Manville, G. Cochrane, J. Cave et al., "Mapping smart cities in the EU," 2014, https://smartcities.at/assets/Publikationen/Weitere-Publikationen-zum-Thema/mappingsmartcities.pdf.

[44] Y. S. Abu-Mostafa, M. Magdon-Ismail, and H. T. Lin, *Learning from data*, vol. 4, AMLBook, New York: NY, USA, 2012.

[45] I. Portugal, P. Alencar, and D. Cowan, "The use of machine learning algorithms in recommender systems: a systematic review," *Expert Systems with Applications*, vol. 97, pp. 205–227, 2018.

[46] R. C. Deo, "Machine learning in medicine," *Circulation*, vol. 132, no. 20, pp. 1920–1930, 2015.

[47] L. Fuyan, "An attribute selection approach and its application," in *2005 International Conference on Neural Networks and Brain*, pp. 636–640, Beijing, China, 2005.

[48] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Science & Business Media, 2009.

[49] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning. Adaptive computation and machine learning*, vol. 31, MIT Press, 2012.

[50] O. Tsur, D. Davidov, and A. Rappoport, "ICWSM—A Great Catchy Name: Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews," in *Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010*, Washington, DC, USA, 2010.

Hindawi
BioMed Research International
Volume 2024, Article ID 9895849, 1 page
https://doi.org/10.1155/2024/9895849

Retraction

Retracted: Dysregulated Circulating Apoptosis- and Autophagy-Related lncRNAs as Diagnostic Markers in Coronary Artery Disease

BioMed Research International

Received 12 March 2024; Accepted 12 March 2024; Published 20 March 2024

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope

(2) Discrepancies in the description of the research reported

(3) Discrepancies between the availability of data and the research described

(4) Inappropriate citations

(5) Incoherent, meaningless and/or irrelevant content included in the article

(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] L. Zhang, D. Lou, D. He et al., "Dysregulated Circulating Apoptosis- and Autophagy-Related lncRNAs as Diagnostic Markers in Coronary Artery Disease," *BioMed Research International*, vol. 2021, Article ID 5517786, 19 pages, 2021.

Hindawi
BioMed Research International
Volume 2021, Article ID 5517786, 19 pages
https://doi.org/10.1155/2021/5517786

Research Article

Dysregulated Circulating Apoptosis- and Autophagy-Related lncRNAs as Diagnostic Markers in Coronary Artery Disease

Lijiao Zhang,[1] Dayuan Lou,[1] Dan He,[2] Ying Wang,[3] Yuhang Wu,[3] Xinyang Cao,[3] and Peng Qu[1]

[1] *Department of Cardiology, The Second Affiliated Hospital of Dalian Medical University, Dalian, 116027 Liaoning, China*
[2] *Department of Physiology and Pathophysiology, School of Basic Medical Sciences, Peking University, Beijing 100191, China*
[3] *Institute of Heart & Vessel Diseases, The Second Affiliated Hospital of Dalian Medical University, Dalian, 116027 Liaoning, China*

Correspondence should be addressed to Peng Qu; qu13304261697@163.com

Received 2 February 2021; Revised 3 April 2021; Accepted 16 August 2021; Published 1 September 2021

Academic Editor: Parameshachari B. D.

Objective. Increasing evidence emphasizes the implications of dysregulated apoptosis and autophagy cellular processes in coronary artery disease (CAD). Herein, we aimed to explore apoptosis- and autophagy-related long noncoding RNAs (lncRNAs) in peripheral blood of CAD patients. *Methods*. The mRNA and lncRNA expression profiles were retrieved from the Gene Expression Omnibus (GEO) database. With $|\text{fold change}| > 1.5$ and adjusted p value < 0.05, differentially expressed apoptosis- and autophagy-related mRNAs were screened between CAD and healthy blood samples. Also, differentially expressed lncRNAs were identified for CAD. Using the psych package, apoptosis- and autophagy-related lncRNAs were defined with Spearson's correlation analysis. Receiver operating characteristic (ROC) curves were conducted for the assessment of the diagnosed efficacy of these apoptosis- and autophagy-related lncRNAs. *Results*. Our results showed that 24 apoptosis- and autophagy-related mRNAs were abnormally expressed in CAD than normal controls. 12 circulating upregulated and 1 downregulated apoptosis- and autophagy-related lncRNAs were identified for CAD. The ROCs confirmed that AC004485.3 (AUC = 0.899), AC004920.3 (AUC = 0.93), AJ006998.2 (AUC = 0.776), H19 (AUC = 0.943), RP5-902P8.10 (AUC = 0.956), RP5-1114G22.2 (AUC = 0.883), RP11-247A12.1 (AUC = 0.885), RP11-288L9.4 (AUC = 0.928), RP11-344B5.2 (AUC = 0.858), RP11-452C8.1 (AUC = 0.929), RP11-565A3.1 (AUC = 0.893), and XXbac-B33L19.4 (AUC = 0.932) exhibited good performance in differentiating CAD from healthy controls. *Conclusion*. Collectively, our findings proposed that circulating apoptosis- and autophagy-related lncRNAs could become underlying diagnostic markers for CAD in clinical practice.

1. Introduction

Coronary artery disease (CAD), as a commonly diagnosed heart disease, contributes to the dominant cause of cardiovascular-related deaths [1]. This disease mainly occurs when the myocardial blood supply decreases [2]. It is composed of myocardial infarction and stable and unstable angina, as well as sudden cardiac death [3]. The etiology of CAD remains little understood due to complex causes such as environmental or genetic risk factors [4]. Hence, it requires exploring in depth for the pathogenesis of CAD. Despite much progress in CAD management, the prevalence is still rising and clinical outcomes are unsatisfactory. Currently, the gold standard for diagnosing CAD is still coronary angiography, and a peripheral blood biochemical test is only used for evaluating the risk factors of CAD. Increasing evidence highlights that circulating biomarkers that can be detected in peripheral blood can be applied for early detection in patients with high-risk CAD [5]. The noninvasive early diagnosis may prevent the progression of CAD, thereby validly lowering its mortality [6]. Nevertheless, there is still lack of circulating markers with high diagnostic value for CAD in clinical practice [7].

Apoptosis and autophagy, as two types of programmed cellular deaths, are both involved in the development of CAD [8]. Undue apoptosis inevitably induces cell death under oxidative stress, ischemia conditions, and the like [9]. Meanwhile, autophagy is an evolutionarily conserved

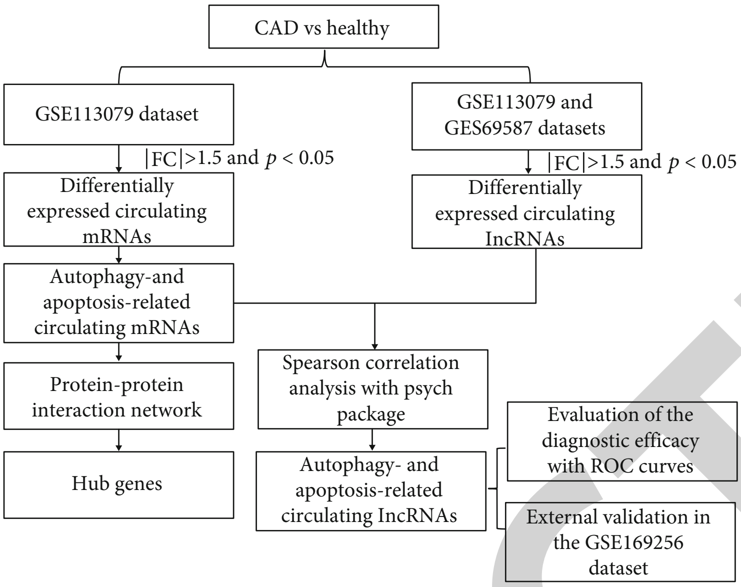

CAD vs healthy
GSE113079 dataset
GSE113079 and GES69587 datasets
|FC|>1.5 and p < 0.05
|FC|>1.5 and p < 0.05
Differentially expressed circulating mRNAs
Differentially expressed circulating lncRNAs
Autophagy-and apoptosis-related circulating mRNAs
Spearson correlation analysis with psych package
Protein-protein interaction network
Evaluation of the diagnostic efficacy with ROC curves
Hub genes
Autophagy- and apoptosis-related circulating lncRNAs
External validation in the GSE169256 dataset

Figure 1: The workflow of this study.

cellular process that participates in degrading and recycling the redundant or useless protein constituents, organelles, and the like [10]. This process is fundamental for maintaining intracellular homeostasis. Hence, its disorder in cardiac cells exerts destructive impacts on the cardiovascular system [11]. Currently, activation of autophagy has been a therapeutic approach for heart diseases [12]. Increasing evidence emphasizes the implications of the interplay between autophagy and apoptosis in CAD [13]. The balance between the two decides cell survival. Both serum levels in CAD patients are higher than healthy controls [14]. lncRNAs with >200 nucleotides may participate in the pathophysiological processes of CAD including autophagy and apoptosis [15]. On account of their tissue and cell specificity, circulating lncRNAs are promising diagnostic markers for various diseases [15]. A previous study has identified three lncRNAs including Chast, HULC, and DICER1-AS1 that are distinctly related to autophagy in blood circulation of CAD patients [16]. Among them, HULC and DICER1-AS1 may properly differentiate CAD individuals from healthy individuals. It has been demonstrated that apoptosis and autophagy may be mediated by several common lncRNAs in CAD. For example, lncRNA MALAT1 [17] or THRIL [18] inhibits autophagy and apoptosis of endothelial progenitor cells in CAD. However, it remains unclear what the clinical implications of autophagy- and apoptosis-related lncRNAs in CAD are. Herein, we firstly screened circulating dysregulated apoptosis- and autophagy-related mRNAs in CAD. Secondly, circulating abnormally expressed lncRNAs were identified in CAD compared to healthy subjects. Thirdly, Spearson's correlation analysis was employed for identifying circulating apoptosis- and autophagy-related lncRNAs, and ROC curves were conducted for evaluating their diagnostic efficacy for CAD. Finally, their expression was externally verified in blood specimens of CAD and healthy subjects. Figure 1 showed the workflow of this study. These lncRNAs proposed by our findings may reflect the pathologically relevant processes that occurred in CAD, which could provide a novel insight into the diagnosis and management of CAD.

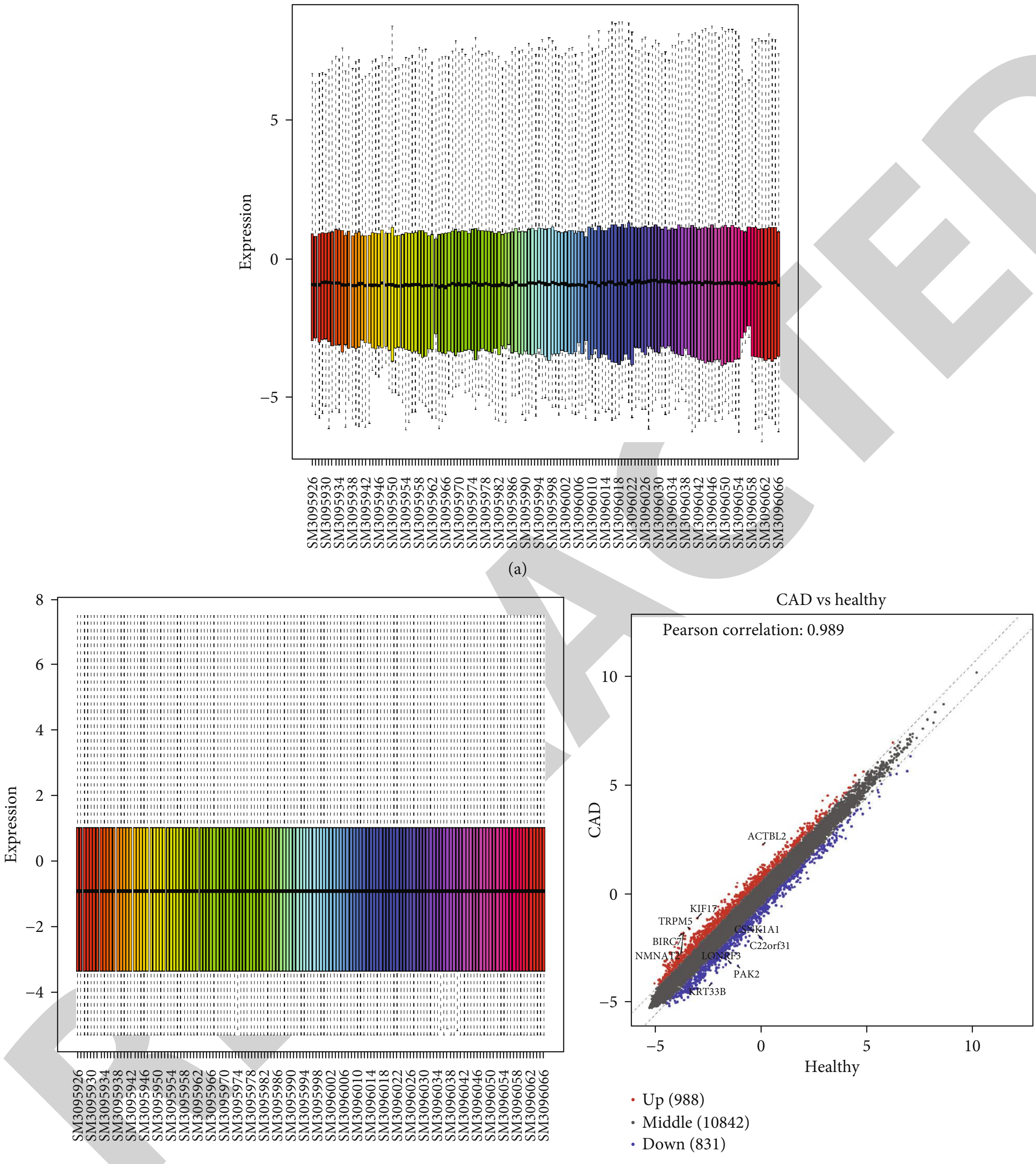
2. Materials and Methods

2.1. Datasets and Preprocessing. The mRNA and lncRNA expression profiles of CAD patients and healthy controls were searched from the Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi) according to the following criteria: organism—*Homo sapiens*; experiment type—noncoding RNA profiling by array; and disease—CAD. As a result, two datasets including GSE113079 and GSE69587 datasets were obtained for this study. The GSE113079 dataset included 93 CAD and 48 healthy blood samples based on the GPL20115 platform [19]. The GSE69587 dataset was composed of 3 CAD and 3 healthy blood specimens on the platform of GPL15314 [20]. The microarray data were normalized to quartile by the normalizeBetweenArrays in the limma package [21]. If the same gene corresponded to multiple IDs, the average value was calculated as the expression level of the gene.

2.2. Differential Expression Analysis. Differentially expressed mRNAs or lncRNAs were screened between CAD and healthy groups with the cutoff of $|\text{fold change (FC)}|>1.5$ and adjusted p value < 0.05 via the limma package, which were visualized into volcano and heatmaps [21].

2.3. Autophagy- and Apoptosis-Related mRNAs. Genes in autophagy (entry: map04140) and apoptosis (entry: map04210) were obtained from the Kyoto Encyclopedia of Genes and Genomes database (KEGG; https://www.kegg.jp/) [22]. They were overlapped by differentially expressed mRNAs called differentially expressed autophagy- and apoptosis-related mRNAs.

2.4. Protein-Protein Interaction (PPI). Physical or functional interactions between specified proteins were analyzed via the

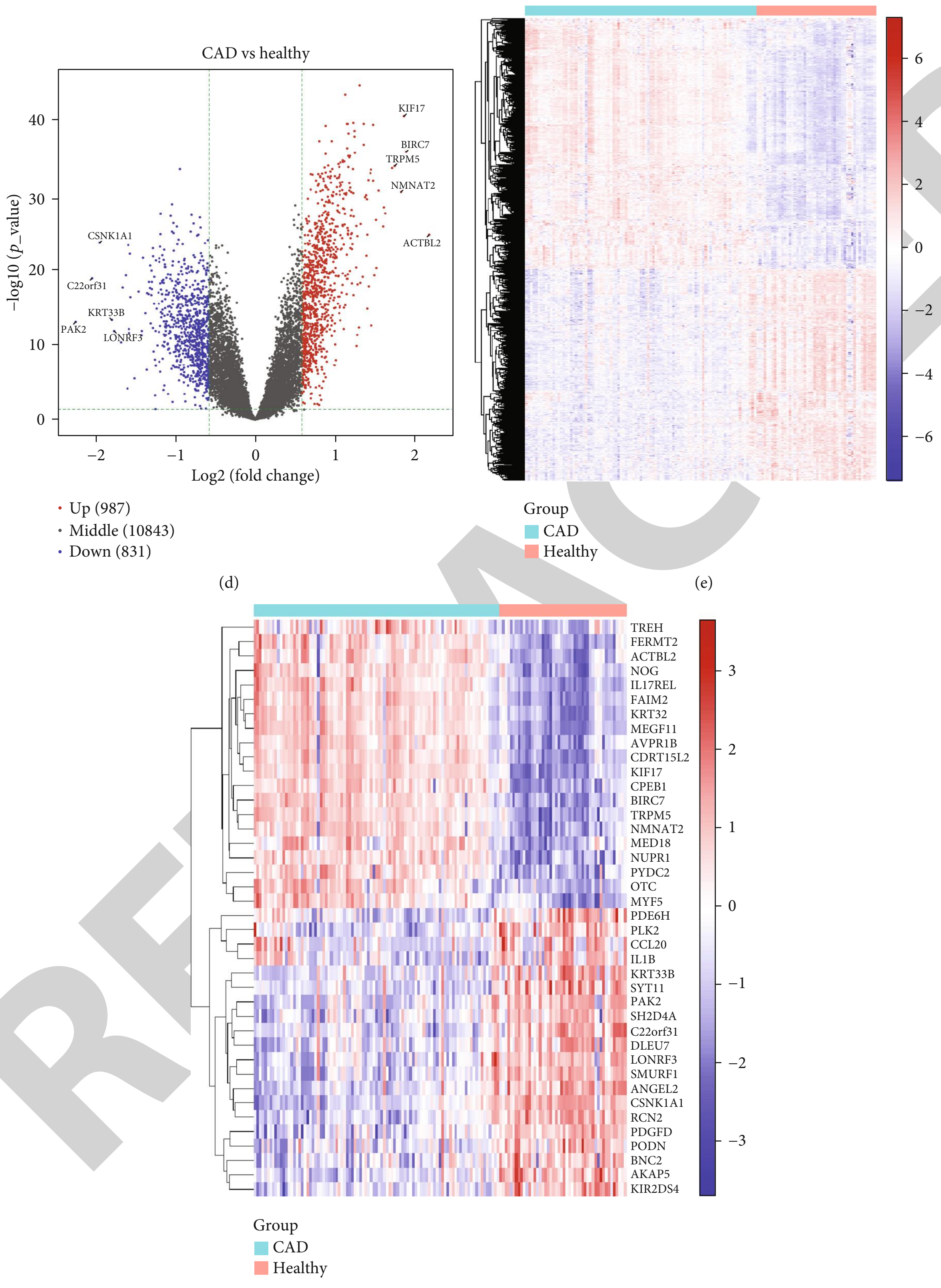


(a)

(b)

(c)

Figure 2: Continued.



(d)

(e)

(f)

Figure 2: Screening circulating abnormally expressed mRNAs for CAD. Box plots for the expression levels of mRNAs in CAD and healthy samples before (a) and after (b) normalization. (c) Scatter and (d) volcano plots for abnormally expressed mRNAs between CAD and healthy samples. (e) Heatmap for the expression patterns of these mRNAs in CAD and healthy samples. (f) Heatmap for the top 20 abnormally expressed mRNAs in CAD and healthy samples. Red: upregulation; blue: downregulation.

Table 1: The top ten circulating upregulated mRNAs in CAD than healthy controls.

Gene name	Log 2 FC	p value	Q value	CAD	Healthy
ACTBL2	2.175707163	2.35433E-25	1.3929E-23	2.274988093	0.09928093
BIRC7	1.89222112	2.12792E-36	1.5848E-33	-1.863763605	-3.755984724
KIF17	1.86560487	3.60156E-41	1.51998E-37	-1.141065467	-3.006670337
NMNAT2	1.826600047	4.98533E-31	7.5142E-29	-1.828796794	-3.655396841
TRPM5	1.745508109	1.50641E-34	6.57677E-32	-1.629455661	-3.374963771
NUPR1	1.713797044	2.91563E-34	1.11863E-31	-2.275289877	-3.98908692
AVPR1B	1.628310899	3.45489E-28	3.55629E-26	-1.949760523	-3.578071423
NOG	1.613371937	4.96571E-29	5.56379E-27	0.499062616	-1.114309321
PYDC2	1.604132397	1.92489E-26	1.40873E-24	-2.468823982	-4.072956379
CPEB1	1.548280137	3.14808E-31	5.10998E-29	-2.738597275	-4.286877412

STRING online tool (http://string-db.org/) [23]. Required confidence (combined score) > 0.7 was set as the cutoff of the interactions. Cytoscape software was utilized to visualize the PPI network [24]. Connectivity degree was calculated, and hub genes with degree ≥ 3 were obtained [25].

2.5. Correlation Analysis. Spearson's correlation analysis between differentially expressed lncRNAs and differentially expressed autophagy- and apoptosis-related mRNAs was presented via the psych package in R. lncRNAs with correlation p value < 0.05 with at least 50% of differentially expressed autophagy- and apoptosis-related mRNAs were considered as differentially expressed autophagy- and apoptosis-related lncRNAs.

2.6. External Validation. The expression of differentially expressed autophagy- and apoptosis-related lncRNAs was externally verified in blood samples from 5 CAD patients and 5 healthy controls in the GSE169256 dataset. Moreover, associations between their expression and clinical features (age) were analyzed by Spearson's correlation tests. Their expression was also compared between male and female patients.

2.7. Statistical Analysis. Based on the expression profiles of the differentially expressed autophagy- and apoptosis-related lncRNAs, relative operating characteristic curves (ROCs) were conducted via the pROC package in R in the GSE113079 dataset [26].

3. Results

3.1. Circulating Abnormally Expressed mRNAs for CAD. To explore CAD-related mRNAs, we screened abnormally expressed mRNAs between 93 CAD and 48 healthy blood samples in the GSE113079 dataset. Firstly, we normalized the microarray data via the limma package (Figures 2(a) and 2(b)). 988 up- and 831 downregulated mRNAs were obtained in CAD compared to normal samples (Figures 2(c) and 2(d)). The top five upregulated mRNAs included KIF17, BIRC7, TRPM5, NMNAT2, and ACTBL2. The top five downregulated mRNAs were as follows: CSNK1A1, C22orf31, KRT33B, PAK2, and LONRF3. Heatmaps demonstrated that these mRNAs clearly distinguished CAD samples into healthy samples (Figure 2(e)). Figure 2(f) visualized the top 20 abnormally expressed mRNAs in CAD and healthy blood samples. The details of the top ten up- and downregulated mRNAs were separately listed in Tables 1 and 2. There were distinct differences in their expressions between CAD and healthy samples indicating that they could participate in the progression of CAD.

3.2. Abnormally Expressed Autophagy- and Apoptosis-Related mRNAs in CAD. To find autophagy- and apoptosis-related mRNAs in CAD, we overlapped the abnormally expressed mRNAs and autophagy- and apoptosis-related mRNAs. As a result, 24 mRNAs were identified for CAD (Figure 3(a)), as follows: ATG2B, CAPN2, CASP8, CTSW, DFFB, FASLG, GABARAPL1, GZMB, HIF1A, ITPR3, JUN, LMNA, MAPK9, MTMR4, NGF, PIK3R2, PPP2CA, PRF1, PRKACA, RRAGB, RRAS2, TNFSF10, TP53AIP1, and TUBA8. We further analyzed whether the proteins encoded by them had physical or functional interactions. A PPI network was constructed based on them, which was made up of 15 nodes (Figure 3(b)). Among all nodes in the network, PRKACA (degree = 3), TNFSF10 (degree = 2), NGF (degree = 3), PIK3R2 (degree = 1), and TUBA8 (degree = 1) were highly expressed in CAD compared to healthy samples. MAPK9 (degree = 2), JUN (degree = 5), HIF1A (degree = 1), GABARAPL1 (degree = 1), ITPR3 (degree = 1), LMNA (degree = 1), PRF1 (degree = 2), GZMB (degree = 5), FASLG (degree = 5), and CASP8 (degree = 3) were poorly expressed in CAD compared to healthy samples.

3.3. Abnormally Expressed Circulating lncRNAs for CAD. Circulating lncRNAs have been considered as diagnosed biomarkers for CAD [27]. Herein, two datasets GSE113079 and GSE69587 were collected for screening abnormally expressed circulating lncRNAs for CAD. In the GSE113079 dataset, we normalized the microarray data of each sample (Figures 4(a) and 4(b)). Then, 1382 up- and 1356 downregulated lncRNAs were identified for CAD blood compared to healthy blood samples (Figure 4(c) and 4(d)). The top five upregulated lncRNAs included RP11-548O1.3, RP11-216N14.9, XLOC_I2_013427, RP11-370I10.2, and linc-

Table 2: The top ten circulating downregulated mRNAs in CAD than healthy controls.

Gene name	Log 2 FC	p value	Q value	CAD	Healthy
KPNA1	-0.585327427	3.78588E-05	9.95082E-05	-3.9803596	-3.395032172
GPRASP1	-0.585805079	4.35625E-10	2.28572E-09	-1.046965965	-0.461160886
RUFY2	-0.585946386	8.90116E-18	1.53958E-16	-0.659711497	-0.073765111
DONSON	-0.586018774	7.67172E-10	3.89149E-09	-1.745995947	-1.159977172
KLHDC1	-0.58626408	3.35774E-09	1.5431E-08	-1.889447587	-1.303183508
MED26	-0.586748565	1.78045E-10	9.89567E-10	-1.114735168	-0.527986603
KDM6A	-0.586918946	3.28334E-09	1.5122E-08	-1.981318284	-1.394399338
SUV39H1	-0.586939293	7.31103E-12	4.96061E-11	-1.621999657	-1.035060363
TAP2	-0.587219549	9.0103E-06	2.59879E-05	0.55477852	1.141998069
FGF7	-0.587788641	3.2713E-07	1.13785E-06	-3.743606566	-3.155817926

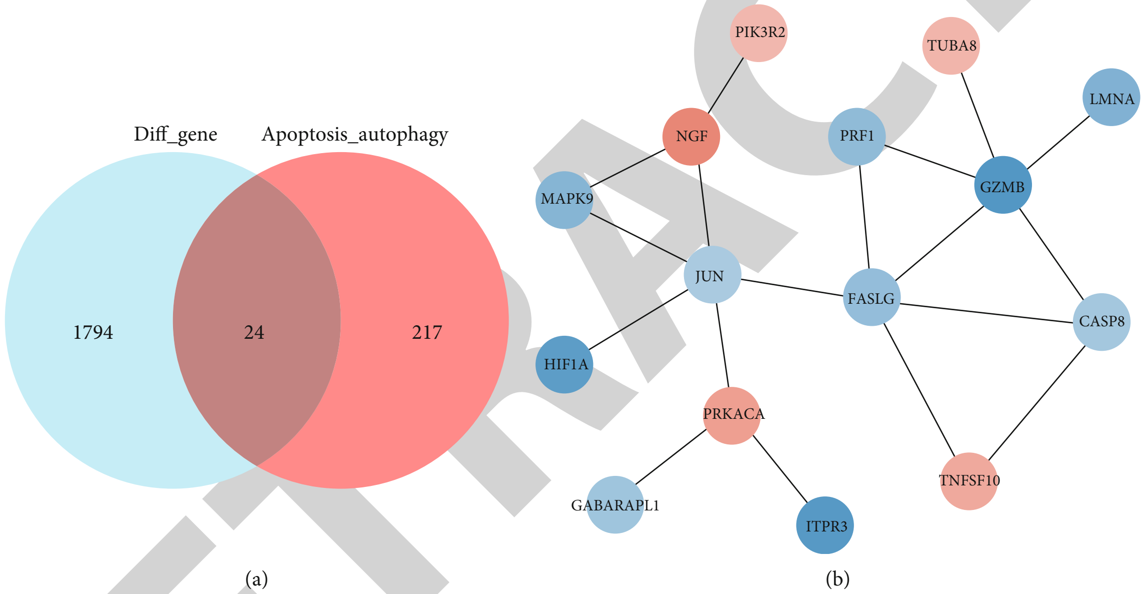
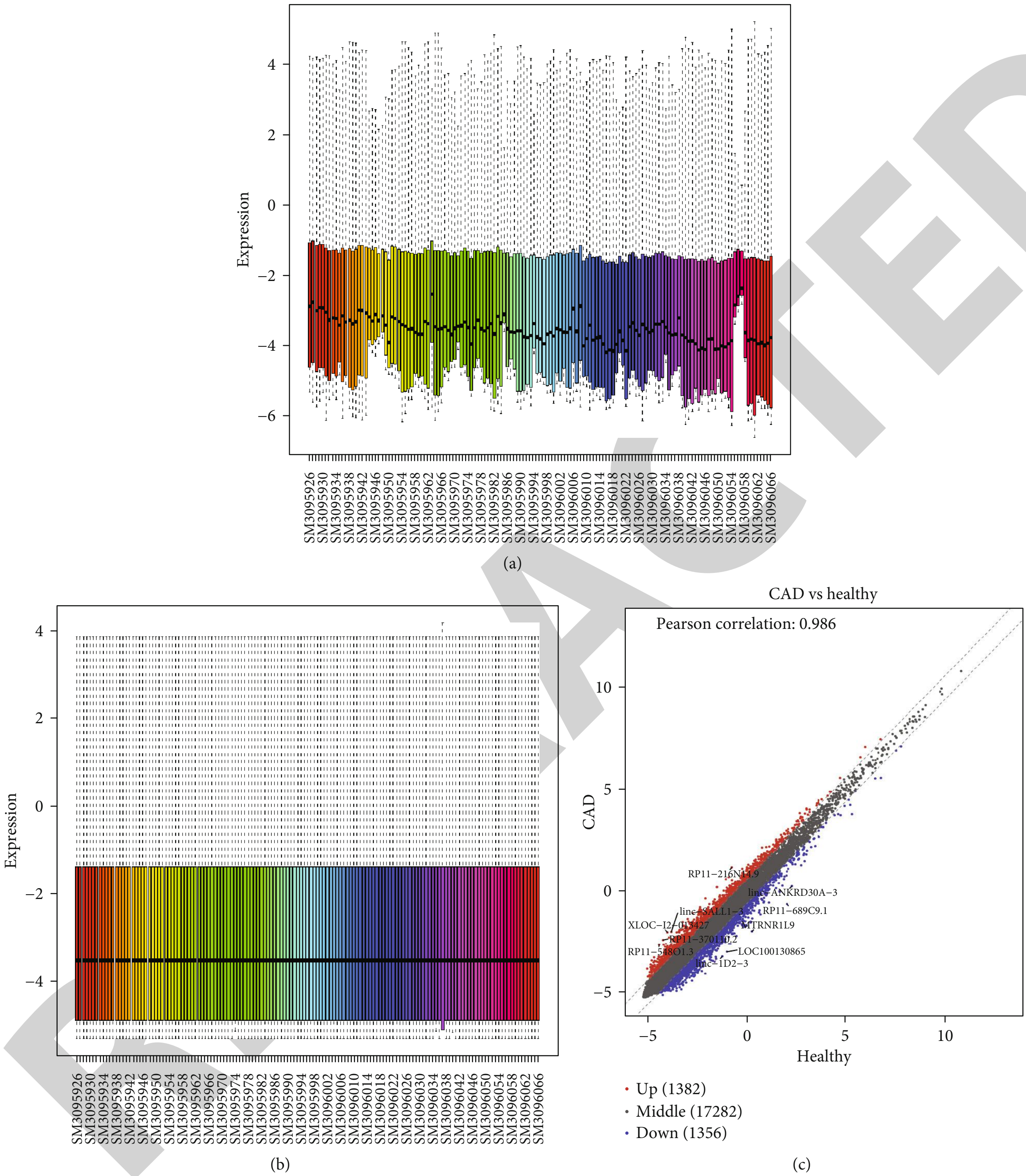


Figure 3: Identification of abnormally expressed autophagy- and apoptosis-related mRNAs in CAD. (a) Venn diagram for the 24 differentially expressed autophagy- and apoptosis-related mRNAs in CAD. (b) Construction of a PPI network based on them. Red: upregulation; blue: downregulation.

SALL1-3. Meanwhile, the top five downregulated lncRNAs covered linc-ID2-3, RP11-689C9.1, linc-ANKRD30A-3, LOC100130865, and MTRNR2L9. CAD samples were distinctly distinguished from healthy samples (Figure 4(e)). The top 20 abnormally expressed circulating lncRNAs were visualized in Figure 4(f). Table 3 listed the detailed information of the top ten circulating upregulated lncRNAs for CAD in the GSE113079 dataset. Meanwhile, the detailed information of the top ten circulating downregulated lncRNAs for CAD in the GSE113079 dataset is shown in Table 4.

Since the GSE69587 dataset has been standardized, this study no longer standardized the dataset. In total, 430 circulating lncRNAs were upregulated and 305 circulating lncRNAs were downregulated in CAD compared to healthy samples (Figures 5(a) and 5(b)). The top five up- (LOC284440, AK096649, AC118138.2, lincRNA-FYN-1, and RP11-372K14.2) and downregulated lncRNAs (AC005779.1, RP11-474J18.1, LOC400657, AK293020, and BC034788) were listed, respectively. Based on the expression levels of these lncRNAs, CAD samples were distinguished from healthy samples (Figure 5(c)). Figure 5(d) depicts the top 20 abnormally expressed circulating lncRNAs. To increase the reliability of the results, we overlapped the abnormally expressed lncRNAs in the GSE113079 and GSE69587 datasets. Consequently, 12 upregulated lncRNAs were obtained for CAD, including AC004485.3, AC004920.3, AJ006998.2, H19, RP11-247A12.1, RP11-288L9.4, RP11-344B5.2, RP11-452C8.1, RP11-565A3.1, RP5-1114G22.2, RP5-902P8.10, and XXbac-B33L19.4 (Figure 6(a)). Moreover, LOC338758 was downregulated in CAD blood samples (Figure 6(b)). These lncRNAs could be involved in CAD development.

3.4. Abnormally Expressed Autophagy- and Apoptosis-Related Circulating lncRNAs for CAD. We analyzed the

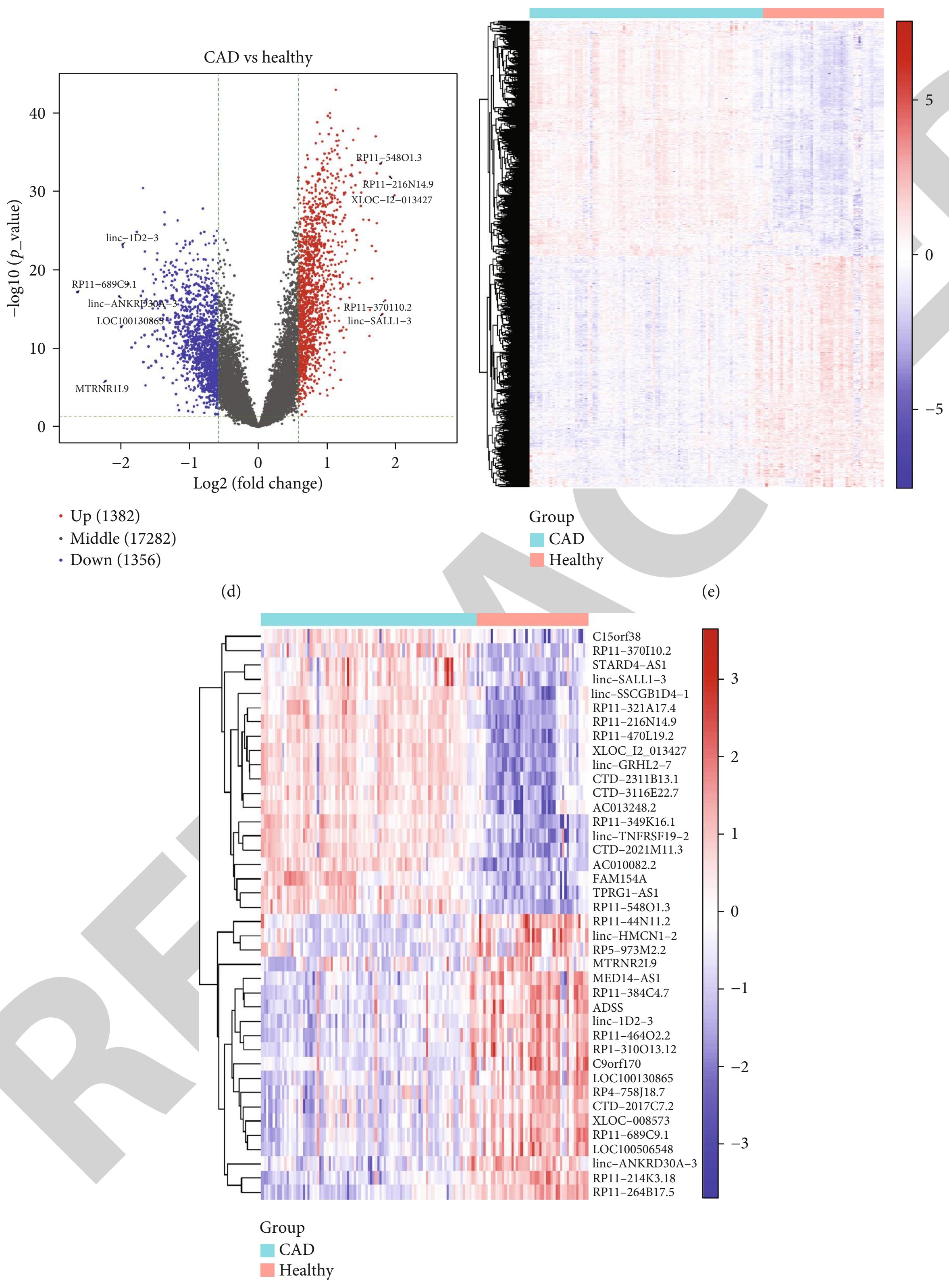


(a)

(b)

(c)

Figure 4: Continued.



(d)

(e)

(f)

Figure 4: Identification of abnormally expressed circulating lncRNAs for CAD in the GSE113079 dataset. Box plots depicting the expression levels of lncRNAs in CAD and healthy samples (a) before and (b) after normalization. (c) Scatter and (d) volcano plots showing all abnormally expressed lncRNAs between CAD and healthy blood samples. (e) Heatmap showing the expression patterns of these lncRNAs in CAD and healthy blood samples. (f) The top 20 abnormally expressed circulating lncRNAs between CAD and healthy groups. Red: upregulation; blue: downregulation.

Table 3: The top ten circulating upregulated lncRNAs for CAD in the GSE113079 dataset.

Gene name	Log 2 FC	*p* value	*Q* value	CAD	Healthy
XLOC_l2_013427	1.983060401	3.43721*E*-30	6.08964*E*-28	-2.043809987	-4.026870388
RP11-216N14.9	1.915753591	1.35955*E*-32	4.462*E*-30	1.134312816	-0.781440775
RP11-370I10.2	1.850499576	8.17738*E*-17	1.55324*E*-15	-2.42064409	-4.271143666
linc-SALL1-3	1.806903286	4.66588*E*-15	6.86341*E*-14	-2.027659452	-3.834562738
RP11-548O1.3	1.77753201	3.00599*E*-34	1.62648*E*-31	-2.686306726	-4.463838737
RP11-321A17.4	1.725530036	1.14947*E*-26	9.96205*E*-25	0.17188431	-1.553645726
CTD-2311B13.1	1.725127268	5.29805*E*-33	2.12134*E*-30	-0.390090969	-2.115218237
AC010082.2	1.710990299	9.98176*E*-38	1.66529*E*-34	0.180339682	-1.530650617
FAM154A	1.709708154	4.06108*E*-23	1.93118*E*-21	-1.554305626	-3.264013779
AC013248.2	1.656964824	1.22892*E*-19	3.52479*E*-18	-1.117728421	-2.774693245

Table 4: The top ten circulating downregulated lncRNAs for CAD in the GSE113079 dataset.

Gene name	Log 2 FC	*p* value	*Q* value	CAD	Healthy
RP11-689C9.1	-2.635616135	7.33574*E*-18	1.61386*E*-16	-0.646839565	1.98877657
MTRNR2L9	-2.223631802	1.54061*E*-06	6.75198*E*-06	-2.008594699	0.215037103
linc-ANKRD30A-3	-2.028353512	2.32287*E*-17	4.76444*E*-16	0.215910733	2.244264245
LOC100130865	-1.996896907	2.25419*E*-13	2.62836*E*-12	-3.03914794	-1.042251032
linc-ID2-3	-1.980200508	6.42151*E*-24	3.41911*E*-22	-3.262290981	-1.282090473
RP11-44N11.2	-1.973423235	1.24267*E*-23	6.26656*E*-22	-4.09545914	-2.122035905
RP11-464O2.2	-1.896597301	6.03265*E*-19	1.58288*E*-17	-3.53754729	-1.640949989
RP4-758J18.7	-1.845072205	8.60858*E*-11	6.927*E*-10	-1.180978882	0.664093323
C9orf170	-1.793813177	2.4523*E*-11	2.15897*E*-10	-3.381744032	-1.587930855
RP11-214K3.18	-1.770839873	1.37694*E*-25	1.02097*E*-23	-2.621810224	-0.850970351

correlation between 13 abnormally expressed circulating lncRNAs and autophagy- and apoptosis-related mRNAs. Herein, we found that PRKACA, PIK3R2, and NGF were positively related to the 12 upregulated lncRNAs (all $p < 0.05$; Figure 7 and Supplementary Table 1). TP53AIP1, RRAS2, PRF1, PPP2CA, MTMR4, MAPK9, LMNA, ITPR3, HIF1A, DFFB, CASP8, CAPN2, and ATG2B were negatively correlated to the 12 upregulated lncRNAs (all $p < 0.05$). Meanwhile, JUN and ITPR3 had positive correlations with downregulated LOC338758 (both $p < 0.05$). Thus, these lncRNAs could be distinctly related to autophagy and apoptosis in CAD.

3.5. Highly Expressed Autophagy- and Apoptosis-Related Circulating lncRNAs as Diagnostic Markers for CAD. In the GSE113079 dataset, we compared the differences in expression of the 12 upregulated autophagy- and apoptosis-related lncRNAs in CAD and healthy blood samples. Our results showed that 11 lncRNAs were distinctly highly expressed in CAD compared to controls, including AC004485.3 (log 2 FC = 1.048; $p = 1.32e$-25), AJ006998.2 (log 2 FC = 0.607; $p = 8.14e$-10), H19 (log 2 FC = 0.713; $p = 5.18e$-16), RP11-247A12.1 (log 2 FC = 0.622; $p = 3.15e$-17), RP11-288L9.4 (log 2 FC = 0.768; $p = 1.06e$-23), RP11-344B5.2 (log 2 FC = 0.968; $p = 2.52e$-11), RP11-452C8.1 (log 2 FC = 0.87; $p = 5.3e$-24), RP11-565A3.1 (log 2 FC = 0.618; $p = 1.29e$-15), RP5-1114G22.2 (log 2 FC = 0.717; p = 1.26e-11), RP5-902P8.10 (log 2 FC = 0.79; $p = 1.2e$-32), and XXbac-B33L19.4 (log 2 FC = 0.966; $p = 4.03e$-28; Figure 8). These lncRNAs could be related to CAD progression.

3.6. Validation of the Diagnostic Efficacy of Autophagy- and Apoptosis-Related Circulating lncRNAs for CAD. The diagnostic efficacy of the autophagy- and apoptosis-related circulating lncRNAs was assessed via ROCs. The areas under the curves (AUCs) are as follows: AC004485.3 (AUC = 0.899; 95%CI = 0.845-0.954; Figure 9(a)), AC004920.3 (AUC = 0.93; 95%CI = 0.885-0.974; Figure 9(b)), AJ006998.2 (AUC = 0.776; 95% CI = 0.691-0.861; Figure 9(c)), H19 (AUC = 0.943; 95%CI = 0.909-0.976; Figure 9(d)), RP5-902P8.10 (AUC = 0.956; 95% CI = 0.919-0.993; Figure 9(e)), RP5-1114G22.2 (AUC = 0.883; 95%CI = 0.827-0.939; Figure 9(f)), RP11-247A12.1 (AUC = 0.885; 95%CI = 0.828-0.942; Figure 9(g)), RP11-288L9.4 (AUC = 0.928; 95%CI = 0.881-0.975; Figure 9(h)), RP11-344B5.2 (AUC = 0.858; 95%CI = 0.789-0.926; Figure 9(i)), RP11-452C8.1 (AUC = 0.929; 95%CI = 0.885-0.972; Figure 9(j)), RP11-565A3.1 (AUC = 0.893; 95%CI = 0.824-0.962; Figure 9(k)), and XXbac-B33L19.4 (AUC =

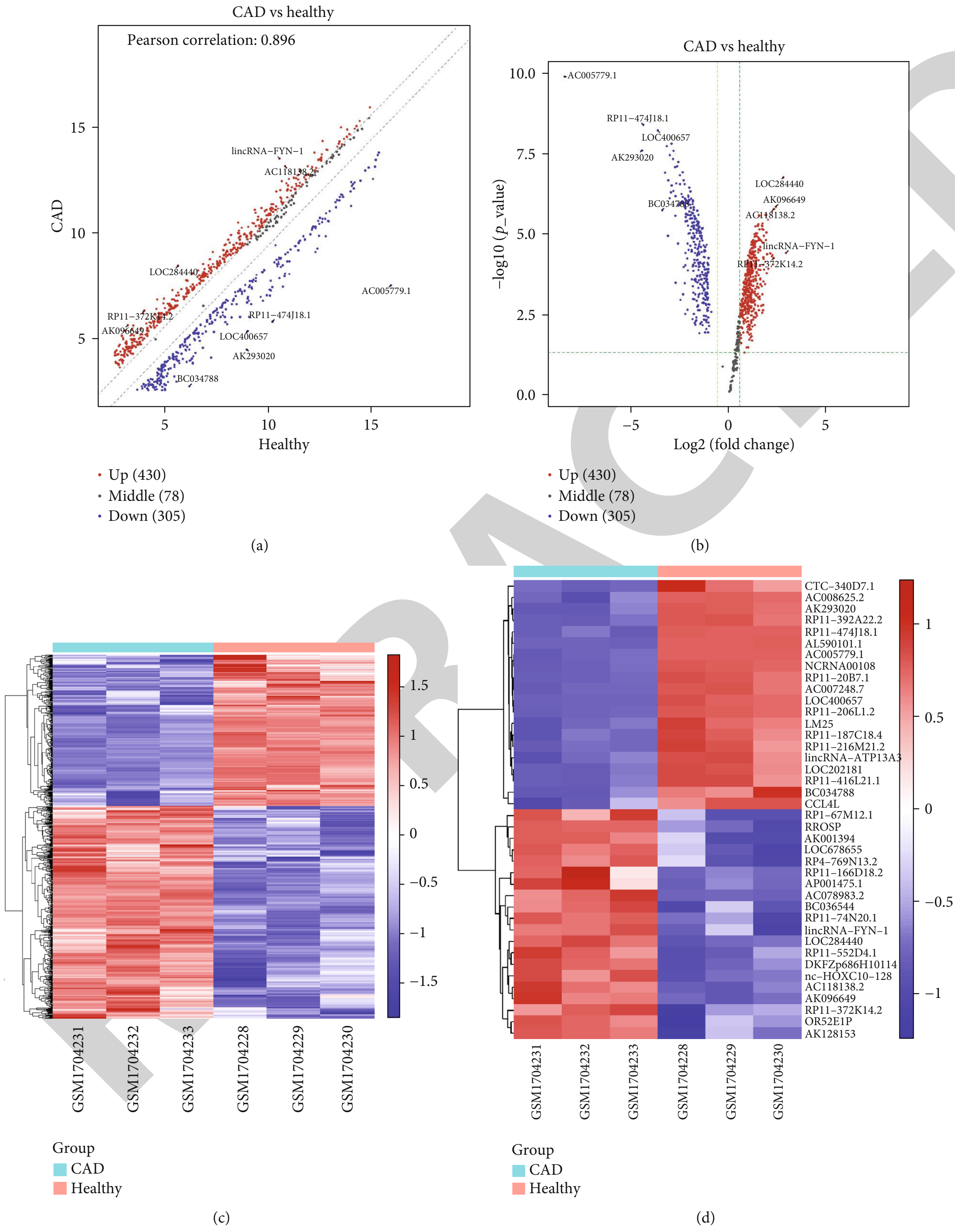


Figure 5: Identification of abnormally expressed circulating lncRNAs for CAD in the GSE69587 dataset. (a) Scatter and (b) volcano diagrams showing abnormally expressed circulating lncRNAs between CAD and healthy samples. (c) Heatmap depicting all abnormally expressed lncRNAs in CAD and healthy blood samples. (d) The top 20 circulating lncRNAs for CAD. Red: upregulation; blue: downregulation.

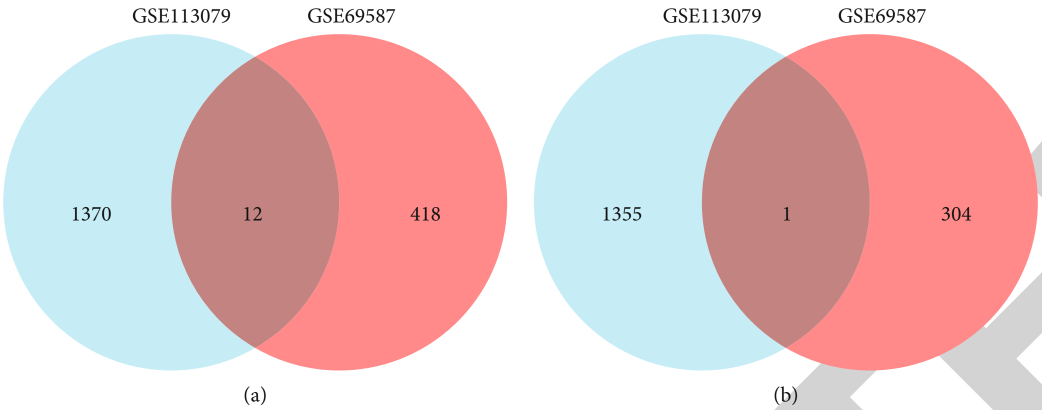


Figure 6: Common abnormally expressed circulating lncRNAs for CAD. (a) 12 upregulated lncRNAs both in the GSE113079 and GSE69587 dataset. (b) One downregulated lncRNA from both the GSE113079 and GSE69587 datasets.

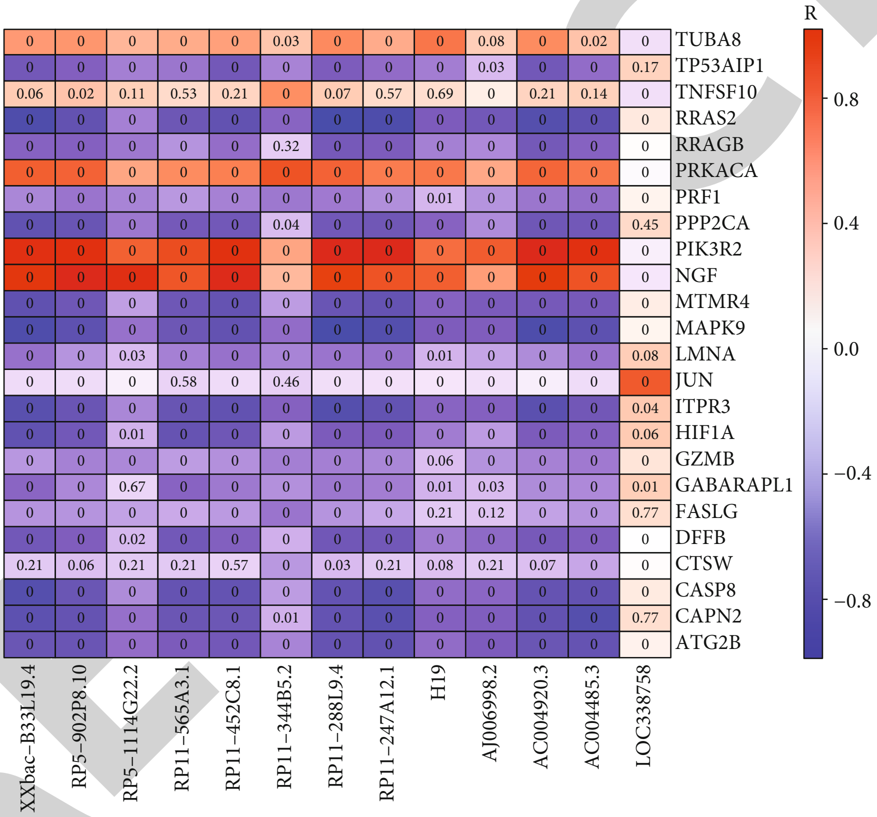


Figure 7: Heat map visualizing the correlation between 13 abnormally expressed circulating lncRNAs and autophagy- and apoptosis-related mRNAs in CAD. The color shade represents the absolute value of the correlation coefficient, and the value represents the p value. The columns represent lncRNAs, and the rows represent mRNAs.

0.932; 95%CI = 0.888-0.976; Figure 9(l)). The data above suggested that these lncRNAs accurately differentiated CAD from healthy controls. Thus, these lncRNAs could be underlying circulating diagnostic markers for CAD.

3.7. External Validation of Autophagy- and Apoptosis-Related Circulating lncRNAs in CAD. To further verify the expression of autophagy- and apoptosis-related circulating lncRNAs in CAD, we employed the GSE169256 dataset. Spearson's correlation analysis showed that AC004485.3 and AC004920.3 were both negatively correlated to age, while AJ006998.2, H19, LOC338758, RP11-247A12.1, RP11-288L9.4, RP11-452C8.1, RP11-565A3.1, RP5-1114G22.2, RP5-902P8.10, and XXbac-B33L19.4 were positively correlated to age (Figure 10(a)). Figure 10(b) shows the differences in expression of the above lncRNAs between male and female CAD patients. Furthermore, the abnormal expression of these lncRNAs was externally confirmed by comparing 5 CAD patients and 5 healthy controls in the GSE169256 dataset (Figure 10(c)).

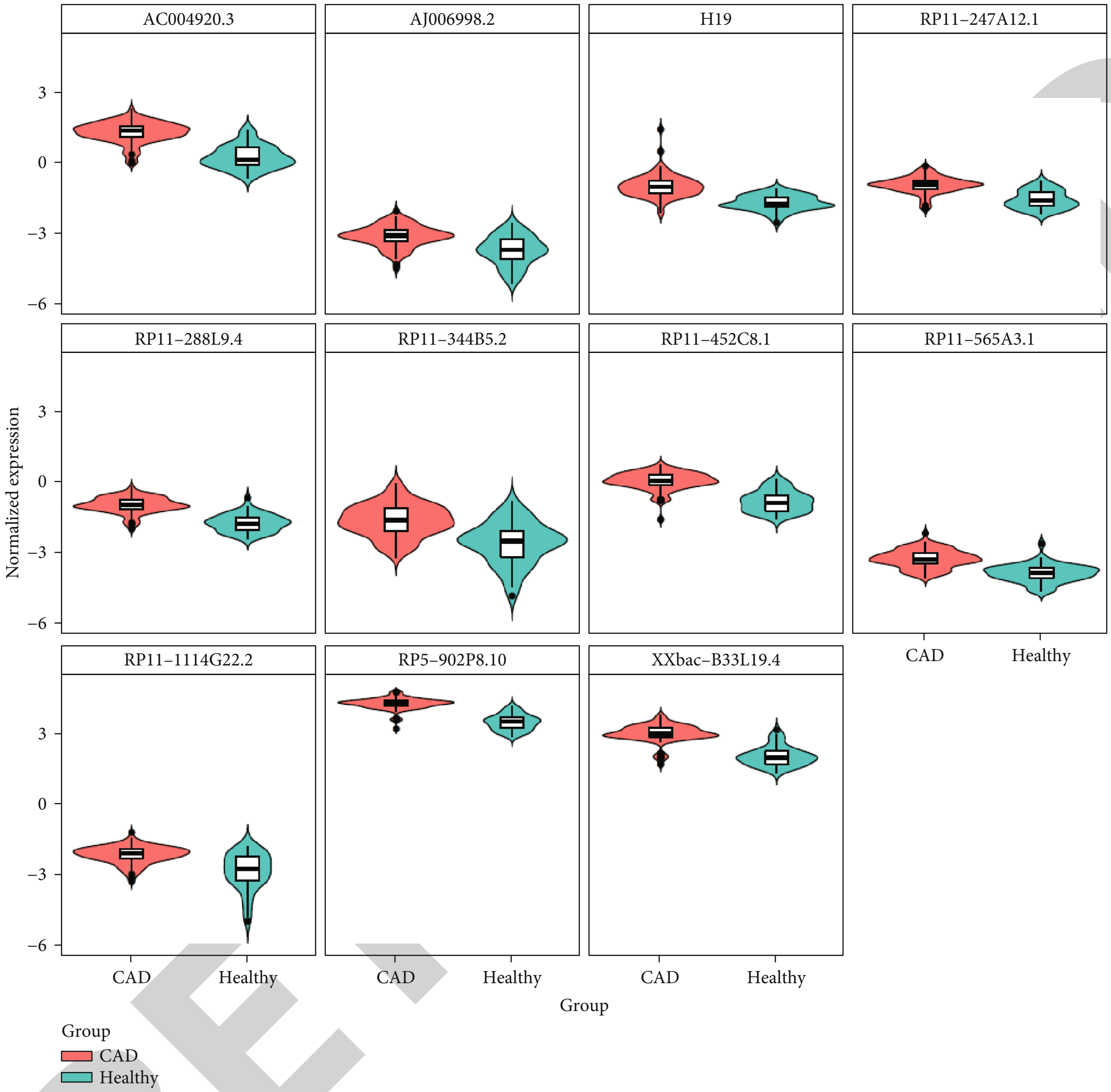
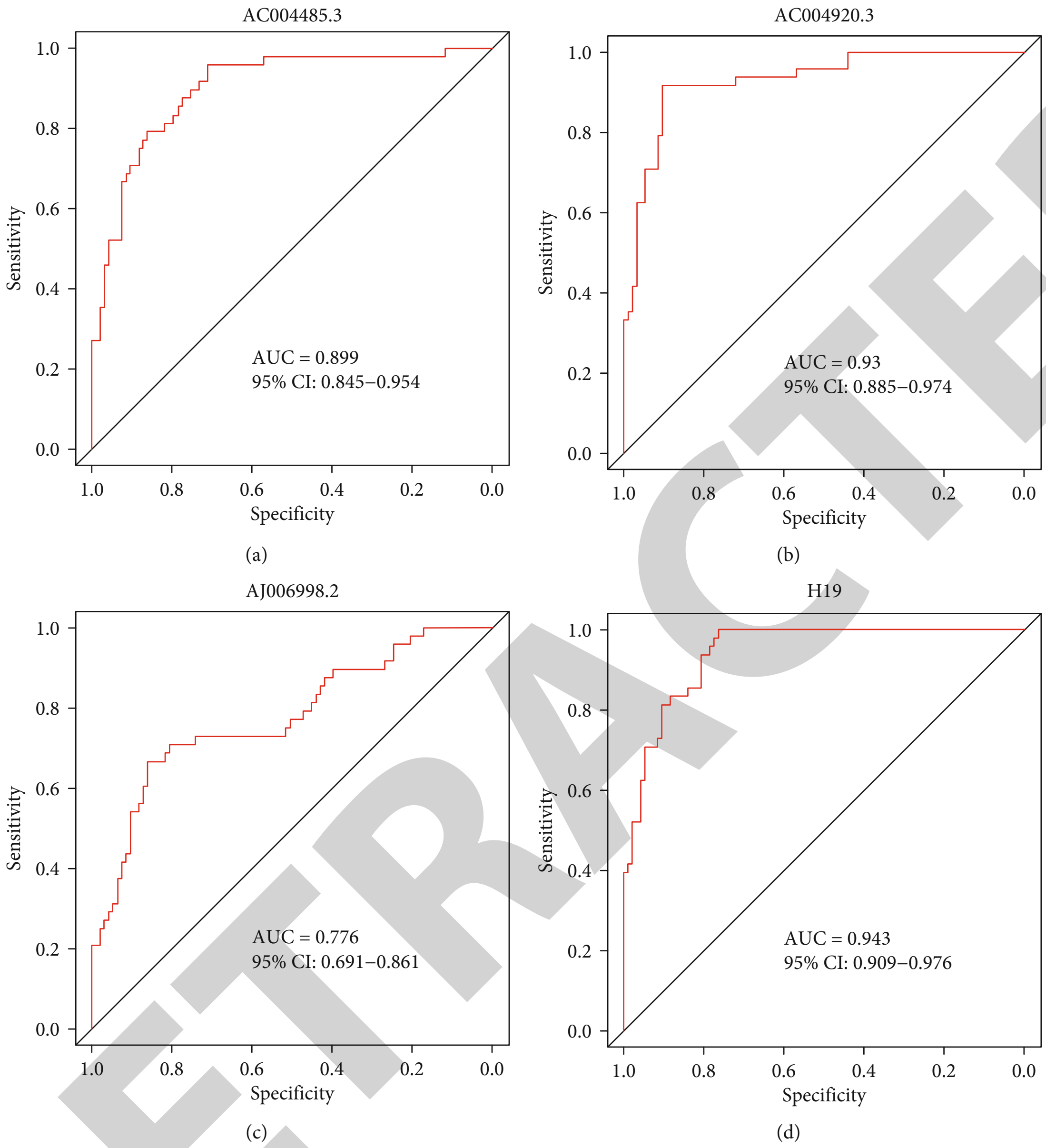


Figure 8: Violin diagram for the expression of 11 circulating lncRNAs in CAD and healthy blood samples, including AC004485.3 (log 2 FC = 1.048; *p* = 1.32*e*-25), AJ006998.2 (log 2 FC = 0.607; *p* = 8.14*e*-10), H19 (log 2 FC = 0.713; *p* = 5.18*e*-16), RP11-247A12.1 (log 2 FC = 0.622; *p* = 3.15*e*-17), RP11-288L9.4 (log 2 FC = 0.768; *p* = 1.06*e*-23), RP11-344B5.2 (log 2 FC = 0.968; *p* = 2.52*e*-11), RP11-452C8.1 (log 2 FC = 0.87; *p* = 5.3*e*-24), RP11-565A3.1 (log 2 FC = 0.618; *p* = 1.29*e*-15), RP5-1114G22.2 (log 2 FC = 0.717; *p* = 1.26*e*-11), RP5-902P8.10 (log 2 FC = 0.79; *p* = 1.2*e*-32), and XXbac-B33L19.4 (log 2 FC = 0.966; *p* = 4.03*e*-28).

4. Discussion

CAD is the most common cause of death globally, which usually kills approximately 17 million individuals each year [28]. Circulating lncRNAs, with tissue and cell specificity, may discern the risk of CAD and assist in formulating therapeutic therapy [29]. In comparison to the conventional diagnosed approach, circulating lncRNAs are noninvasive and innocuous, with highly sensitive and accurate advantages [30]. Furthermore, lncRNAs may participate in the progression of CAD via mediating apoptosis and autophagy, two forms of programmed cell deaths [15]. On account of these strengths, this study explored circulating lncRNAs related to apoptosis and autophagy for CAD diagnosis. However, so far, there is still a lack of circulating lncRNAs for the diagnosis of CAD. To fill the gap, our study identified 12 apoptosis- and autophagy-related circulating lncRNAs that had good performance in diagnosing CAD.

In this study, 988 up- and 831 downregulated mRNAs were screened for CAD compared to healthy controls in blood samples. Among them, KIF17, BIRC7, TRPM5, NMNAT2, ACTBL2, CSNK1A1, C22orf31, KRT33B,



(a) (b) (c) (d)

Figure 9: Continued.

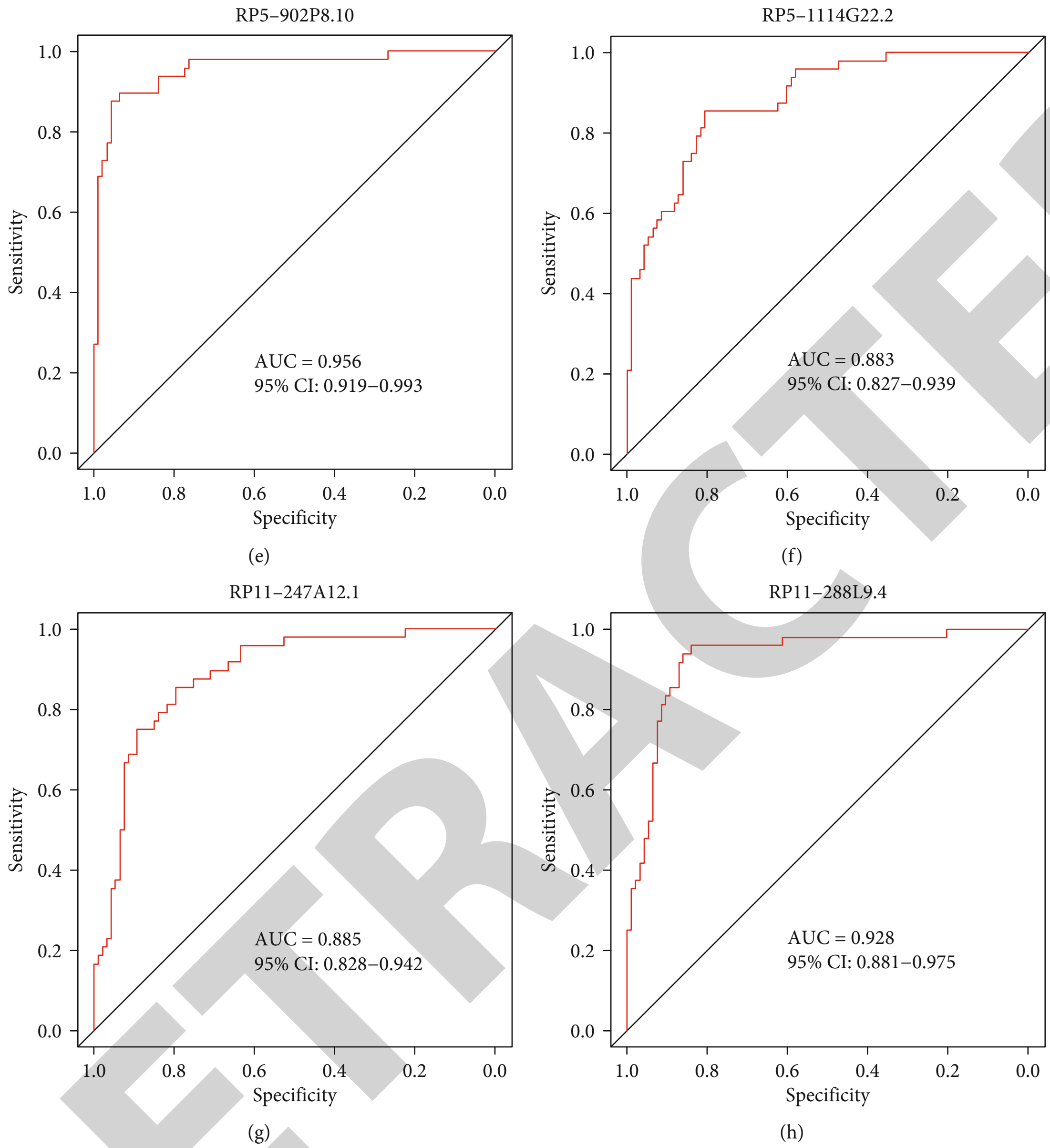
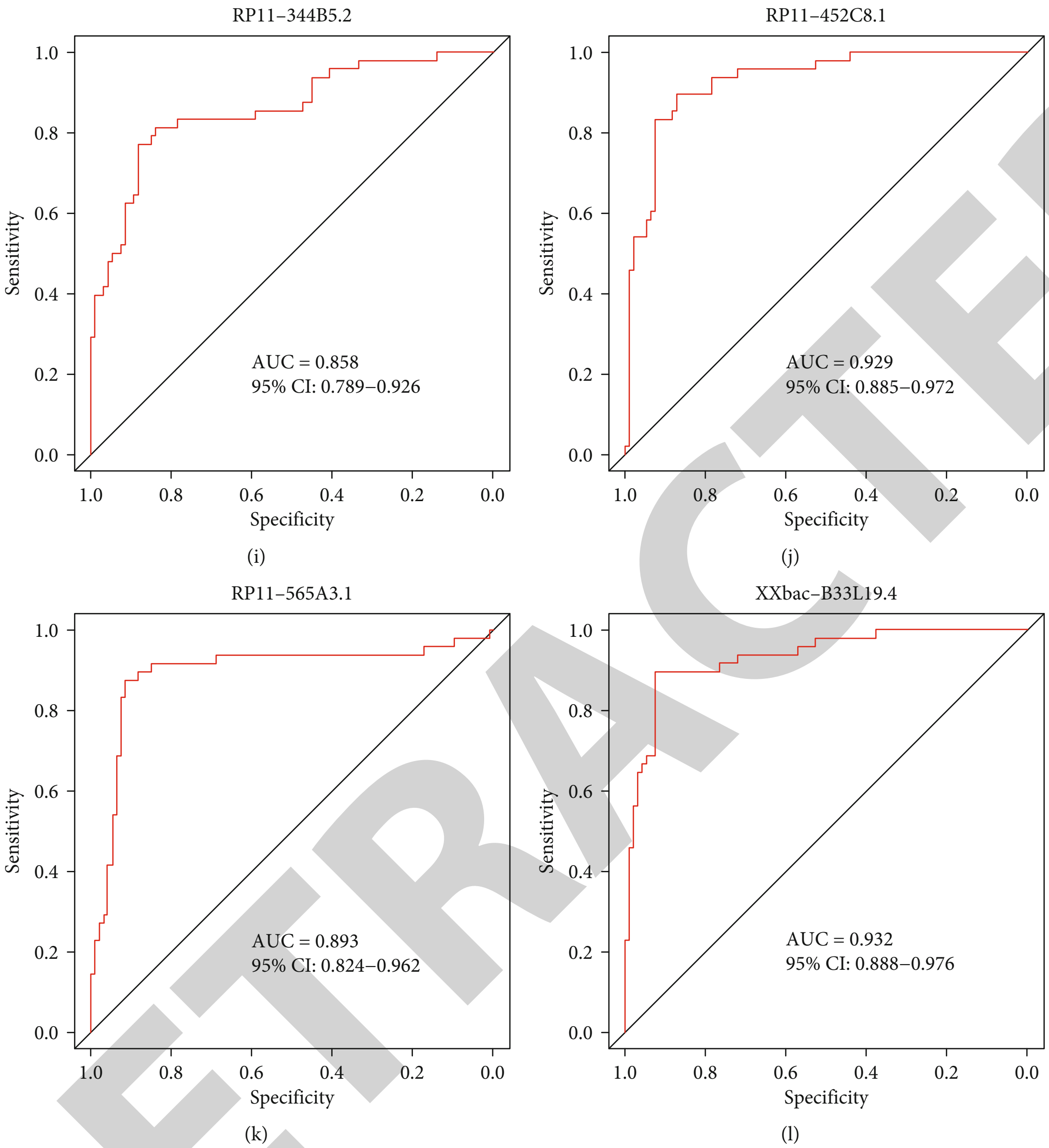


Figure 9: Continued.



(i)

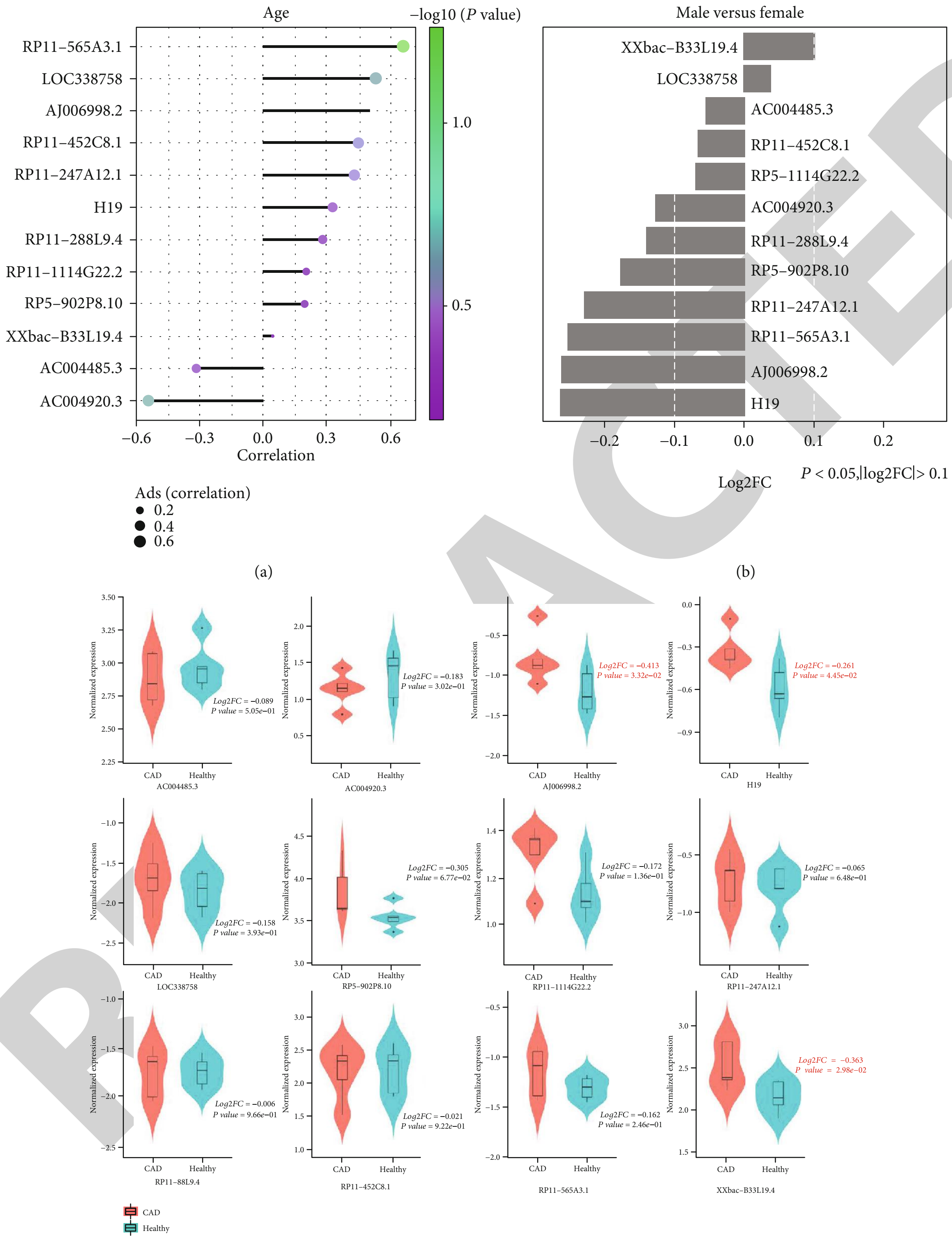
(j)

(k)

(l)

Figure 9: ROC validates 12 circulating lncRNAs as diagnostic markers for CAD: (a) AC004485.3, (b) AC004920.3, (c) AJ006998.2, (d) H19, (e) RP5-902P8.10, (f) RP5-1114G22.2, (g) RP11-247A12.1, (h) RP11-288L9.4, (i) RP11-344B5.2, (j) RP11-452C8.1, (k) RP11-565A3.1, and (l) XXbac-B33L19.4.

PAK2, and LONRF3 had the highest changes in expression between CAD and healthy controls. Among them, a previous study has found that PAK2 activated by METRNL may attenuate cardiomyocyte apoptosis induced by myocardial ischemia/reperfusion [31]. The balance between apoptosis and autophagy exerts a critical role on the pathological conditions of CAD [32]. Among all differentially expressed mRNAs, 24 mRNAs were on the apoptosis and autophagy pathways. Of them, silencing CAPN2 suppresses NF-κB activation as well as decreases myocardial infarction remodeling [33]. CASP8 polymorphic variants (-652 6N del/ins, IVS12-19G>A) could predict the risk of CAD [34]. Furthermore, high CASP8 levels have an association with elevated incidence of coronary diseases [35]. GzmB expression is increased in blood and tissues of CAD patients compared to healthy individuals [34]. H1F1A is significantly altered in CAD patients compared to controls [28]. ITPR3 single-nucleotide polymorphism rs2229634 could be indicative of an increased incidence in coronary artery aneurysm among youngsters [36]. Variants in LMNA are linked with lipodystrophy [37]. Combining previous research, these mRNAs identified by this study may possess tight links to CAD pathogenesis. We constructed a PPI network based on these apoptosis and autophagy mRNAs, which could help to study the pathogenesis of CAD from a systematic perspective. In the network, PRKACA, TNFSF10, NGF, PIK3R2, TUBA8, MAPK9, JUN, HIF1A, GABARAPL1, ITPR3, LMNA, PRF1, GZMB, FASLG, and CASP8 were considered hub genes for CAD. The protein products from these hub genes could have physical and functional associations, which



(a)

(b)

(c)

Figure 10: External validation of autophagy- and apoptosis-related circulating lncRNAs in CAD in the GSE169256 dataset. (a) Spearson's correlation analysis shows the associations between the circulating expression of AC004485.3, AC004920.3, AJ006998.2, H19, LOC338758, RP11-247A12.1, RP11-288L9.4, RP11-452C8.1, RP11-565A3.1, RP5-1114G22.2, RP5-902P8.10, and XXbac-B33L19.4 and age. (b) The differences in expression of the above lncRNAs between male and female CAD patients. (c) External validation of the above lncRNAs in 5 CAD patients and 5 healthy controls.

might play vital roles in the biological processes of CAD. In apoptosis and autophagy pathways, the regulation of other genes might be often affected by these hub genes.

Circulating lncRNAs have been proven as diagnosed biomarkers for CAD [19]. By comprehensive analysis of the two datasets, we identified 12 upregulated lncRNAs in CAD compared to controls, including AC004485.3, AC004920.3, AJ006998.2, H19, RP11-247A12.1, RP11-288L9.4, RP11-344B5.2, RP11-452C8.1, RP11-565A3.1, RP5-1114G22.2, RP5-902P8.10, and XXbac-B33L19.4. Moreover, one downregulated lncRNA, LOC338758, was identified in CAD blood samples. These lncRNAs could be involved in CAD progression. Among them, upregulated H19 has been detected in blood samples of CAD patients compared to heathy controls [38]. Other lncRNAs should be explored during CAD development in depth. Considerable research suggests that lncRNAs widely participate in biological processes in CAD, especially apoptosis and autophagy [17, 18, 39]. Here, we analyzed the associations between circulating abnormally expressed lncRNAs and apoptosis- and autophagy-related mRNAs in CAD blood samples. Our data suggested that PRKACA, PIK3R2, and NGF were positively linked to the 12 upregulated lncRNAs. TP53AIP1, RRAS2, PRF1, PPP2CA, MTMR4, MAPK9, LMNA, ITPR3, HIF1A, DFFB, CASP8, CAPN2, and ATG2B had negative correlation to the 12 upregulated lncRNAs. Meanwhile, JUN and ITPR3 exhibited positive relationships with downregulated LOC338758. These data indicated that these lncRNAs could be closely associated with the autophagy and apoptosis processes in CAD.

On account of the shortcomings of current diagnostic markers on CAD, circulating lncRNAs appear to have attracted close attention. After verification, our data demonstrated that AC004485.3 (AUC = 0.899), AC004920.3 (AUC = 0.93), AJ006998.2 (AUC = 0.776), H19 (AUC = 0.943), RP5-902P8.10 (AUC = 0.956), RP5-1114G22.2 (AUC = 0.883), RP11-247A12.1 (AUC = 0.885), RP11-288L9.4 (AUC = 0.928), RP11-344B5.2 (AUC = 0.858), RP11-452C8.1 (AUC = 0.929), RP11-565A3.1 (AUC = 0.893), and XXbac-B33L19.4 (AUC = 0.932) exhibited good performance to differentiate CAD from healthy controls. The above findings concerning circulating lncRNAs might possess effective diagnostic value on CAD, thereby reducing mortality. Among them, circulating H19 is correlated to risk of CAD among a Chinese cohort [40]. Additionally, H19 polymorphisms show a tight link to CAD occurrence [41, 42].

Circulating lncRNAs have received much attention in the past years due to their effectiveness and noninvasiveness. This study declared several apoptosis- and autophagy-related circulating lncRNAs with high sensitivity and accuracy. Hence, these lncRNAs might possess the clinical application value as diagnostic markers for CAD, thereby improving the diagnostic accuracy and prolonging patients' survival duration. Several limitations should be considered in this study. First, the conclusion of this study was based on retrospective studies. The diagnostic efficacy of these circulating lncRNAs will be validated in a large-scale, multicenter, and prospective cohort in our future research. Second, the functions of these lncRNAs in apoptosis and autophagy processes are not completely clear in CAD. Their specific mechanisms will be explored in our further experimental studies.

5. Conclusion

Collectively, this study identified and externally confirmed that 12 apoptosis- and autophagy-related circulating lncRNAs (AC004485.3, AC004920.3, AJ006998.2, H19, LOC338758, RP11-247A12.1, RP11-288L9.4, RP11-452C8.1, RP11-565A3.1, RP5-1114G22.2, RP5-902P8.10, and XXbac-B33L19.4) were distinctly upregulated in CAD compared to healthy controls. More importantly, they had good performance in distinguishing CAD from healthy individuals. Thus, these circulating lncRNAs could be promising diagnostic markers for CAD.

Abbreviations

CAD:	Coronary artery disease
lncRNAs:	Long noncoding RNAs
ROC:	Receiver operating characteristic
GEO:	Gene Expression Omnibus
FC:	Fold change
KEGG:	Kyoto Encyclopedia of Genes and Genomes database.
PPI:	Protein-protein interaction
AUCs:	Areas under the curves.

Data Availability

The data used to support the findings of this study are included within the supplementary information files.

Conflicts of Interest

The authors declare no conflicts of interest.

Authors' Contributions

Lijiao Zhang and Dayuan Lou contributed equally to this work.

Acknowledgments

This study was supported by National Natural Science Foundation of China (81670406 and 91739119).

Supplementary Materials

Supplementary Table 1: correlation analysis between 13 abnormally expressed circulating lncRNAs and autophagy- and apoptosis-related mRNAs. *(Supplementary Materials)*

References

[1] G. Agha, M. M. Mendelson, C. K. Ward-Caviness et al., "Blood leukocyte DNA methylation predicts risk of future myocardial infarction and coronary heart disease," *Circulation*, vol. 140, no. 8, pp. 645–657, 2019.

[2] C. Wang, Q. Li, H. Yang et al., “MMP9, CXCR1, TLR6, andMPOparticipant in the progression of coronary artery disease,” *Journal of Cellular Physiology*, vol. 235, no. 11, pp. 8283–8292, 2020.

[3] Z. Jia, Y. Zhang, Q. Li et al., “A coronary artery disease-associated tRNAThr mutation altered mitochondrial function, apoptosis and angiogenesis,” *Nucleic Acids Research*, vol. 47, no. 4, pp. 2056–2074, 2019.

[4] A. V. Khera, C. A. Emdin, I. Drake et al., “Genetic risk, adherence to a healthy lifestyle, and coronary disease,” *The New England Journal of Medicine*, vol. 375, no. 24, pp. 2349–2358, 2016.

[5] L. Miao, R.-X. Yin, Q.-H. Zhang et al., “A novel circRNA-miRNA-mRNA network identifies circ-YOD1 as a biomarker for coronary artery disease,” *Scientific Reports*, vol. 9, no. 1, p. 18314, 2019.

[6] Y. Hu and J. Hu, “Diagnostic value of circulating lncRNA ANRIL and its correlation with coronary artery disease parameters,” *Brazilian Journal of Medical and Biological Research*, vol. 52, no. 8, p. e8309, 2019.

[7] L. Zhang, Y. Zhang, Y. Zhao et al., “Circulating miRNAs as biomarkers for early diagnosis of coronary artery disease,” *Expert Opinion on Therapeutic Patents*, vol. 28, no. 8, pp. 591–601, 2018.

[8] Y. Dong, H. Chen, J. Gao, Y. Liu, J. Li, and J. Wang, “Molecular machinery and interplay of apoptosis and autophagy in coronary heart disease,” *Journal of Molecular and Cellular Cardiology*, vol. 136, pp. 27–41, 2019.

[9] X. Yang, T. He, S. Han et al., “The role of traditional Chinese medicine in the regulation of oxidative stress in treating coronary heart disease,” *Oxidative Medicine and Cellular Longevity*, vol. 2019, Article ID 3231424, 13 pages, 2019.

[10] B. Levine and G. Kroemer, “Biological functions of autophagy genes: a disease perspective,” *Cell*, vol. 176, no. 1-2, pp. 11–42, 2019.

[11] M. Abdellatif, S. Ljubojevic-Holzer, F. Madeo, and S. Sedej, “Autophagy in cardiovascular health and disease,” *Progress in Molecular Biology and Translational Science*, vol. 172, pp. 87–106, 2020.

[12] Q. Lu, Y. Yao, Z. Hu et al., “Angiogenic factor AGGF1 activates autophagy with an essential role in therapeutic angiogenesis for heart disease,” *PLoS Biology*, vol. 14, no. 8, p. e1002529, 2016.

[13] X. Wang, Z. Guo, Z. Ding, and J. L. Mehta, “Inflammation, autophagy, and apoptosis after myocardial infarction,” *Journal of the American Heart Association*, vol. 7, no. 9, 2018.

[14] O. Kaplan and G. Demircan, “Relationship of autophagy and apoptosis with total occlusion of coronary arteries,” *Medical Science Monitor*, vol. 24, pp. 6984–6988, 2018.

[15] Y.-. M. Ding, E. C. Chan, L.-. C. Liu et al., “Long noncoding RNAs: important participants and potential therapeutic targets for myocardial ischaemia reperfusion injury,” *Clinical and Experimental Pharmacology & Physiology*, vol. 47, no. 11, pp. 1783–1790, 2020.

[16] N. Ebadi, S. Ghafouri-Fard, M. Taheri, S. Arsang-Jang, S. A. Parsa, and M. D. Omrani, “Dysregulation of autophagy-related lncRNAs in peripheral blood of coronary artery disease patients,” *European Journal of Pharmacology*, vol. 867, p. 172852, 2020.

[17] Y. Zhu, T. Yang, J. Duan, N. Mu, and T. Zhang, “MALAT1/-miR-15b-5p/MAPK1 mediates endothelial progenitor cells autophagy and affects coronary atherosclerotic heart disease via mTOR signaling pathway,” *Aging (Albany NY)*, vol. 11, no. 4, pp. 1089–1109, 2019.

[18] J. Xiao, Y. Lu, and X. Yang, “THRIL mediates endothelial progenitor cells autophagy via AKT pathway and FUS,” *Molecular Medicine*, vol. 26, no. 1, p. 86, 2020.

[19] L. Li, L. Wang, H. Li et al., “Characterization of lncRNA expression profile and identification of novel lncRNA biomarkers to diagnose coronary artery disease,” *Atherosclerosis*, vol. 275, pp. 359–367, 2018.

[20] Y. Cai, Y. Yang, X. Chen et al., “Circulating “lncRNA OTTHUMT00000387022” from monocytes as a novel biomarker for coronary artery disease,” *Cardiovascular Research*, vol. 112, no. 3, pp. 714–724, 2016.

[21] M. E. Ritchie, B. Phipson, Y. H. Di Wu, C. W. Law, W. Shi, and G. K. Smyth, “limma powers differential expression analyses for RNA-sequencing and microarray studies,” *Nucleic Acids Research*, vol. 43, no. 7, p. e47, 2015.

[22] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima, “KEGG: new perspectives on genomes, pathways, diseases and drugs,” *Nucleic Acids Research*, vol. 45, no. D1, pp. D353–d361, 2017.

[23] D. Szklarczyk, J. H. Morris, H. Cook et al., “The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible,” *Nucleic Acids Research*, vol. 45, no. D1, pp. D362–d368, 2017.

[24] N. T. Doncheva, J. H. Morris, J. Gorodkin, and L. J. Jensen, “Cytoscape StringApp: network analysis and visualization of proteomics data,” *Journal of Proteome Research*, vol. 18, no. 2, pp. 623–632, 2019.

[25] X. He and J. Zhang, “Why do hubs tend to be essential in protein networks?,” *PLoS Genetics*, vol. 2, no. 6, p. e88, 2006.

[26] X. Robin, N. Turck, A. Hainard et al., “pROC: an open-source package for R and S+ to analyze and compare ROC curves,” *BMC Bioinformatics*, vol. 12, no. 1, 2011.

[27] L. Wang and Y. Jin, “Noncoding RNAs as biomarkers for acute coronary syndrome,” *BioMed Research International*, vol. 2020, Article ID 3298696, 2020.

[28] V. Alexandar, P. G. Nayar, R. Murugesan, S. Shajahan, J. Krishnan, and S. S. S. J. Ahmed, “A systems biology and proteomics-based approach identifies SRC and VEGFA as biomarkers in risk factor mediated coronary heart disease,” *Molecular BioSystems*, vol. 12, pp. 2594–2604, 2016.

[29] L. Miao, R.-X. Yin, Q.-H. Zhang et al., “A novel lncRNA-miRNA-mRNA triple network identifies lncRNA TWF1 as an important regulator of miRNA and gene expression in coronary artery disease,” *Nutrition & Metabolism (London)*, vol. 16, no. 1, 2019.

[30] J. Viereck and T. Thum, “Circulating noncoding RNAs as biomarkers of cardiovascular disease and injury,” *Circulation Research*, vol. 120, no. 2, pp. 381–399, 2017.

[31] L. Xu, Y. Cai, Y. Wang, and C. Xu, “Meteorin-like (METRNL) attenuates myocardial ischemia/reperfusion injury-induced cardiomyocytes apoptosis by alleviating endoplasmic reticulum stress via activation of AMPK-PAK2 signaling in H9C2 cells,” *Medical Science Monitor*, vol. 26, 2020.

[32] G. Gao, W. Chen, M. Yan et al., “Rapamycin regulates the balance between cardiomyocyte apoptosis and autophagy in chronic heart failure by inhibiting mTOR signaling,” *International Journal of Molecular Medicine*, vol. 45, no. 1, pp. 195–209, 2020.

Hindawi
BioMed Research International
Volume 2024, Article ID 9753878, 1 page
https://doi.org/10.1155/2024/9753878

Retraction

Retracted: A Permissioned Blockchain-Based Clinical Trial Service Platform to Improve Trial Data Transparency

BioMed Research International

Received 12 March 2024; Accepted 12 March 2024; Published 20 March 2024

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope

(2) Discrepancies in the description of the research reported

(3) Discrepancies between the availability of data and the research described

(4) Inappropriate citations

(5) Incoherent, meaningless and/or irrelevant content included in the article

(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] L. Hang, B. Kim, K. Kim, and D. Kim, "A Permissioned Blockchain-Based Clinical Trial Service Platform to Improve Trial Data Transparency," *BioMed Research International*, vol. 2021, Article ID 5554487, 22 pages, 2021.

Hindawi
BioMed Research International
Volume 2021, Article ID 5554487, 22 pages
https://doi.org/10.1155/2021/5554487

Research Article

A Permissioned Blockchain-Based Clinical Trial Service Platform to Improve Trial Data Transparency

Lei Hang,[1] **BumHwi Kim**,[2] **KyuHyung Kim**,[2] **and DoHyeun Kim**[3]

[1] *Shanghai Normal University Tianhua College, Shanghai 201815, China*
[2] *Daegu-Gyeongbuk Research Center, Electronics and Telecommunications Research Institute, Daegu 42994, Republic of Korea*
[3] *Department of Computer Engineering, Jeju National University, Jeju 63243, Republic of Korea*

Correspondence should be addressed to DoHyeun Kim; kimdh@jejunu.ac.kr

Received 26 February 2021; Accepted 5 July 2021; Published 30 July 2021

Academic Editor: B. D. Parameshachari

The clinical research faces numerous challenges, from patient enrollment to data privacy concerns and regulatory requirements to spiraling costs. Blockchain technology has the potential to overcome these challenges, thus making clinical trials transparent and enhancing public trust in a fair and open process with all stakeholders because of its distinct features such as data immutability and transparency. This paper proposes a permissioned blockchain platform to ensure clinical data transparency and provides secure clinical trial-related solutions. We explore the core functionalities of blockchain applied to clinical trials and illustrate its general principle concretely. These clinical trial operations are automated using the smart contract, which ensures traceability, prevents a posteriori reconstruction, and securely automates the clinical trial. A web-based user interface is also implemented to visualize the data from the blockchain and ease the interaction with the blockchain network. A proof of concept is implemented on Hyperledger Fabric in the case study of clinical management for multiple clinical trials to demonstrate the designed approach's feasibility. Lastly, the experiment results demonstrate the efficiency and usability of the proposed platform.

1. Introduction

Clinical trials generate a significant amount of clinical research data to approve new drugs, instruments, and medical or surgical treatments on human participants [1]. During the trial, investigators collect data from the subject at a fixed period, including vital signs, changes to symptoms, side effects, or complications caused by the study drug. Generally, the clinical trial requires collaboration among diverse stakeholders, including regulatory bodies, pharmaceutical companies, clinical sites, and most importantly subjects who participate in the clinical trial [2–4]. Besides, each stakeholder of a clinical trial offers a different set of tools to support all the bases that form the clinical trial infrastructure. The current workflow of clinical trials consists of different steps performed independently of each other [5], as shown in Figure 1. In this model, data is created from disparate sources (smart devices, clinical trial sites), processed, and analyzed by different organizations in their preferred way and format.

It is estimated that near 80% of clinical studies are nonreproducible [6]. The high error rate probably results from human faults, fraud, or misconduct. Clinical trials' data quality can be directly affected by the usual mistakes in data acquisition and transcription [7]. More strict supervision requirements are necessary due to these issues, increasing the further burden of document inspection for clinical research regulators. In general, checking the data quality is not a trivial job. It needs knowledge of both the business domain and data modeling to maintain the data quality at a reasonable level [8].

Blockchain technology can improve clinical trials' quality with better reproducibility and grant both researcher communities with secure data sharing and patient groups with tools guaranteeing their privacy [9]. Many experts and institutes have certificated blockchain technology as an emerging technology to ensure secure data transformation among different stakeholders via a distributed ledger in recent years. A consensus-driven approach is enforced to agree with multiple trustless stakeholders [10, 11]. Blockchain is secure,

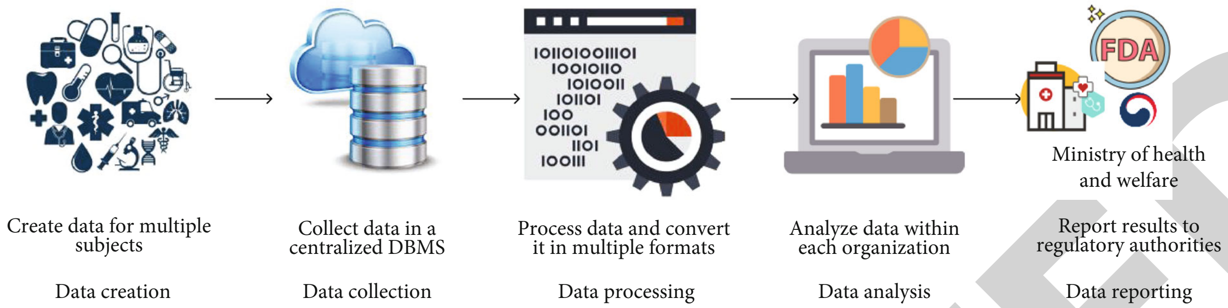

FDA
Ministry of health and welfare
Create data for multiple subjects
Data creation
Collect data in a centralized DBMS
Data collection
Process data and convert it in multiple formats
Data processing
Analyze data within each organization
Data analysis
Report results to regulatory authorities
Data reporting

Figure 1: Current clinical trial workflow.

decentralized info comprised of various peers, which provides storage to record data from an outsized form of entities [12]. It is a series of linked blocks of transferred data between different connected nodes that form the blockchain network. Among various blockchain characteristics, transparency is regarded as the critical characteristic since it enables instant access to data, replicated on all nodes without third parties' interventions [13, 14]. The smart contract [15] is a decentralized application compared with other software types as it resides on the blockchain. The smart contract plays multiple roles, including automating an application's business logic without the involvement of third parties, verifying the predetermined rules of operation, and enforcing obligations on an action if these rules are met [16].

Satoshi Nakamoto invented blockchain in 2008 to serve as the public ledger of the cryptocurrency Bitcoin [17]. As the technology evolves, various applications have been explored other than finance and banking services, for example, in the healthcare sector [18–21]. For example, blockchain-based models for electronic medical records have been proposed that would empower patients to exercise greater ownership of their medical data and enhance data sharing between platforms [22–25]. The characteristics of blockchain technologies such as data transparency and traceability have sparked a boom of the reformation in conventional healthcare organizations, including health data sharing and patient follow-up through transaction trace [26]. It is also indicated that the blockchain technology might also prove useful in supporting or even supplanting the traditional data infrastructure used in clinical trials [27–29]. Immutable clinical trial data recorded using a blockchain may inspire greater confidence in its integrity, resulting in better science, safer medicines, and enhanced public trust in biomedical research. However, most of the existing works either validate blockchain's feasibility in clinical trials or present the conceptual application of blockchain falling short of the specific implementation framework or approach. This paper is aimed at giving a practical way of building blockchain-based applications to offer an efficient, secure, and decentralized trace and track solution for clinical trials and further revolutionize the healthcare industry's developments.

This paper provides the following contributions to make clinical trials transparent and build trust among different stakeholders by using blockchain technologies:

(i) The structure and functionality of a blockchain-based architecture for managing the clinical trials among various stakeholders are presented. Besides, it demonstrates how pharmaceutical companies and other clinical trial investigators can collect and access clinical data in a secure, distributed manner

(ii) A smart contract is implemented to automate clinical trial-related operations without third party involvement. Besides, various service scenarios related to clinical trials, including subject enrollment, medical data collection, and audit query, are presented to demonstrate how the smart contract can handle these operations

(iii) A proof of concept is implemented using Hyperledger Fabric [30] to demonstrate the proposed solution's usability and effectiveness. This implementation is specified to perform clinical trial-related processes among multiple organizations for multiple clinical trials. Besides, a web-based application is implemented to enable end users to interact with the blockchain network, thereby increasing the operability in practice

The rest of this paper is structured as follows: Section 2 analyzes current issues of clinical trials. Section 3 overviews some existing studies related to this topic. Section 4 presents the proposed platform's design, while Section 5 describes the prototype application implemented on top of the designed platform with various snapshots of experiment results. Section 6 evaluates system performance using multiple performance metrics. Section 7 evaluates the security of the proposed platform. Lastly, Section 8 concludes the paper and points out research directions in future work.

2. Current Issues in Clinical Trials

A steady and efficient influx of medical innovations is critical to ensure that society's emerging medical needs are adequately met. Ongoing research has shown that drug development costs continue to rise with an average of 9% each year [31]. Unfortunately, drug development speed, costs, and success rates have not improved over the years, and in some instances, operating conditions have gotten worse. This has resulted in a situation where success rates for new candidate medicinal products entering clinical development are at an all-time low, with only 11% ultimately making it to the market.

Clinical research generates enormous amounts of trial data every day, which increases the pressures of regulatory agencies to overcome such significant data barriers. Besides, it is becoming evident that legacy data management systems are not strong enough to process and preserve the data extracted from current research studies. These systems lack relevant measures to build trust in the clinical trial industry among consumers and regulators. As a result, the three most giant pharmaceutical conglomerates, Pfizer, Amgen, and Sanofi, all carry out the plan to find an effective way of utilizing blockchain technology in clinical trials, from storing secure data to ultimately reducing research costs [32].

In the past decade, numerous eClinical solutions have emerged, aimed at smoothening trial operations and data management. However, most of these function-specific solutions are unable to communicate with each other. Indeed, most clinical researchers experience issues with keeping track of the status of documents and processes in their clinical operations [33], according to a recent industry-wide survey. Though fully integrated and all-encompassing eClinical platforms exist, in practice, these platforms are only accessible to large pharmaceutical developers for being prohibitively expensive as roughly 65% of trials are performed by academia, hospitals, and smaller industry stakeholders [34]. For regulators who experience difficulty auditing research data, there are no easy and secure way of viewing the complex network of data exchange, no real-time access to results once generated, and no easy way to track data historically [35]. Therefore, the FDA has listed a lack of traceability as one of the top data issues in clinical trials [36].

Patient data accessibility is another major problem in the current clinical trial industry as it involves many different specialty stakeholders and even crosses organizational and national boundaries. Clinical trial stakeholders operate in relative isolation, each applying their software systems, data formats, workflows, and organizational structures. In general, patient data is dispersed across multiple proprietary systems that have no relationship and cooperation with each other, making it exceedingly tough to recruit individuals for trials [37]. However, even worse, when investigators recruit enough patients to initiate a clinical trial, the medical condition of patients for that specific treatment could remain an issue and result in an incorrect study with false positive and false harmful errors.

3. Related Works

The clinical trial industry is held in high regard for its pivotal role in advancing human health. The trustworthiness of trial data is primary to modern medical theory and methods but cannot be assured due to various reasons, as described in the previous section. Some clinical trial management systems (CTMS) are used by biotechnology and pharmaceutical industries to manage clinical trials in clinical research. EasyTrial [38] is an online clinical trial management system for administering all clinical trial tasks (operational and logistical), so healthcare professionals can use their time efficiently. All trial data is easily accessible and presented simply and comprehensively from the database. The key features of EasyTrial include complete study overview, security monitoring, study data backup, design of eCRF and questionnaires, management of multiple sites, automatic subject invitation and notification by SMS/email, etc. EDETEK [39] is an innovative clinical solutions company that provides high-quality technology platforms and related clinical services to pharmaceutical, biotechnology, and medical device companies. The clinical data management system enables transforming the study protocol to CRF design and EDC setup, including medical coding and extensive metric dashboards and structure based on CDISC or the sponsor's standards. It also provides various functions such as patient management, site management, clinical data warehousing, and trial supply management. VOXCE [40] is an open-source clinical trial management system that offers complete control of data collection, operation, and analysis. It is built using a subscription model that allows multiple users to be part of that subscription, allowing users to utilize the same libraries and templates, which utilizes questions, sections, and forms. Phoenix CTMS [41] is a modern web application combining database software capabilities in clinical research in one modular system, including patient recruitment, clinical trial management, clinical data management, and electronic data capture. This unmatched feature set is geared to support all operational and regulatory requirements of the clinical front end in academic research, at CROs (Contract Research Organizations) and hospitals conducting clinical studies of any phase. However, these legacy systems are not built for today's complex trials. Even with the best IT support, clinical operation teams still get bogged down in ways that are easily overcome with new technology. For example, duplicative efforts and siloed processes cause inefficiencies and trial delays. The system's maintenance and upgrade raise costs and burdens since these systems are designed in a central architecture. Integration is another challenge that results in low data quality. Moreover, lack of insight can lead to uninformed decisions, and lack of oversight may result in noncompliance.

Blockchain is considered a foundation for improving the clinical research methodology and a step toward better transparency to enhance trust among research institutions, clinical sites, and patient populations. Adopting blockchain technology into the clinical industry brings obvious benefits from secure data tracking and sharing to users' data availability and privacy. A more in-depth exploration of blockchain values in clinical trials remains to be made by exchanging views, ideas, practical problems, and unmet needs between IT and clinical trial specialties. The authors in [42] describe the scope, requirements, system design, and challenges of blockchain technology specific to clinical trials and precision medicine. It is reported that several companies and institutions such as IBM Watson Health and the FDA are developing an initiative to define how blockchain can be used to work across the healthcare data from a variety of sources, including clinical trials, wearables, and electronic health records [43].

Table 1: Comparison with existing studies based on blockchain.

Name	Framework	Use of smart contract	Network type	Evaluation of security	Proof of concept
[44]	Bitcoin	N/A	Permissionless	No	No
[45]	Ethereum	Yes	Permissioned	No	No
[46]	Ethereum	Yes	Both	No	No
[47]	Ethereum	Yes	Permissioned	No	No
[48]	Bitcoin	Yes	Permissionless	No	No
[49]	N/A	No	Permissioned	Yes	No
Proposed approach	Hyperledger Fabric	Yes	Permissioned	Yes	Yes

The method for using blockchain to provide proof of prespecified endpoints in clinical trial protocols is first reported by Carlisle [44]. The author confirms the use of blockchain as a low-cost, independently verifiable method that could be widely and readily used to audit and verify scientific studies' reliability. The authors in [45] are the first to introduce smart contracts on the Ethereum network to address the data manipulation issues common to clinical trials. The results show that blockchain smart contracts can act as a trusted administrator and provide an immutable record of trial history, including trial registration, protocol, subject registration, and clinical measurements. A hybrid blockchain model is presented to tackle known issues in clinical trials [46]. A public blockchain approach is used for clinical trial recruitment, while a private blockchain approach is used for persistent monitoring. The smart contract feature of Ethereum is also utilized to automate the workflow of clinical trial operations. BlockTrial [47] is another similar system that runs trial-related smart contracts on the Ethereum network. This system allows patients to grant researchers permission to access their data and submit queries for data stored in the blockchain. Scrybe [48] is a blockchain ledger designed for clinical trials, paired with a novel Lightweight Mining (LWM) algorithm, to provide provenance on data with minimal system resource requirements. To demonstrate the superiority of the proposed LWM algorithm, the authors conduct a comprehensive security analysis. The verification results indicate that the algorithm can provide the legal and ethical framework for auditors to validate clinical trials, expedite the research process, and save costs in the process. The authors in [49] present a blockchain system for collecting consent from patients. Each step of the consent collection process appended to the blockchain is attached with a timestamp. In this way, the traceability of the patient's consent is established and maintained unobtrusively.

According to the literature overviewed above, numerous studies have shown that adopting blockchain technology in the clinical trial industry is growing up to a new embranchment and subject field to enhance data privacy, security, and transparency. Table 1 represents the comparison of the proposed approach with these existing blockchain-based studies. Most of the existing studies either remain just in the design phase without implementation or present a simple implementation. This paper proposes a proof of concept using blockchain to manage trial data and perform trial operations.

4. Clinical Trial Service Platform Based on Blockchain

4.1. Overview of the Clinical Trial Service Platform. The proposed blockchain-based clinical trial service platform consists of three layers: the physical layer, the service layer, and the application layer. As shown in Figure 2, the bottom layer is the physical layer, comprised of various smart devices for collecting the vital signals from subjects. These devices enable the collection of objective measures of intervention effects in both clinical and remote settings. The service layer adopts the modular design that makes the blockchain network more natural to maintain and extend. This layer encapsulates blockchain technologies' various characteristics into individual modules, including peer-to-peer (P2P) protocol, certificate authority, and consensus. The blockchain network consists of various entities, including distributed ledger, certificate authority, P2P protocol, consensus, and smart contract. The ledger is decentralized storage to maintain the replicated and shared data distributed across the entire network. The smart contract defines the business logic concerning all clinical trial-related operations, such as creating a patient record. The orderer is a particular node performing a consensus algorithm to guarantee the stable operation of the blockchain network and ensure that all peers maintain the data consistency. The event hub is responsible for emitting events whenever a new block is generated or the condition defined in the smart contract is triggered. The functions specified in the smart contract are encapsulated into REST APIs. The external smart devices and applications can integrate with the network by calling these APIs. The application layer describes the way that services provided by the blockchain are visualized to the end user. The blockchain network can be accessed either using responsive web-based applications or native applications on smartphones and tablets.

4.2. Roles in the Clinical Trial Service Platform. There are admins, clinical research associates (CRAs), clinical research coordinators (CRCs), principal investigators (PIs), subjects, and smart devices, all of which form the stakeholders of the proposed system, as shown in Figure 3. The pharmaceutical company is the sponsor responsible for creating experiment plans, developing clinical protocols, and preparing instruments and medicines for a clinical trial. The pharmaceutical company is not considered the stakeholder since the CRO provides clinical trial services for the pharmaceutical

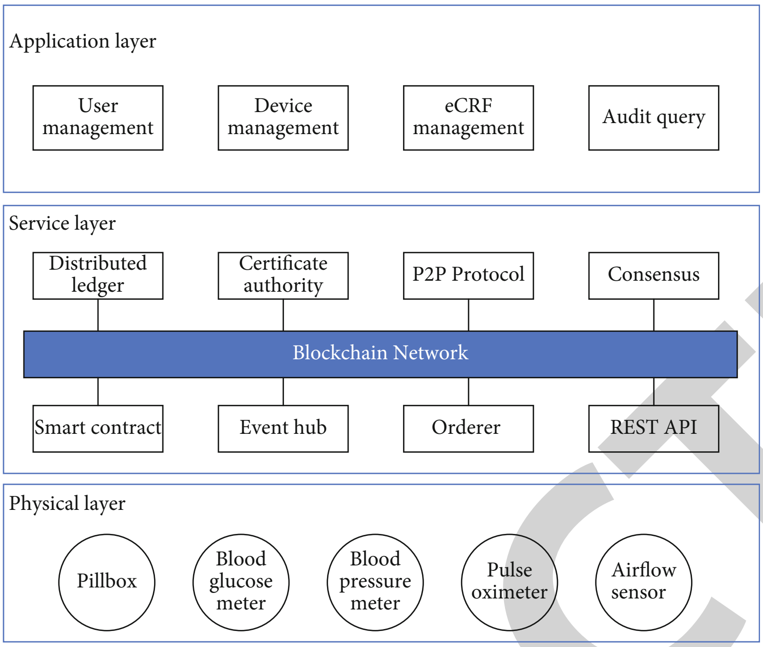


Figure 2: The layer-based architecture of the proposed clinical trial service platform.

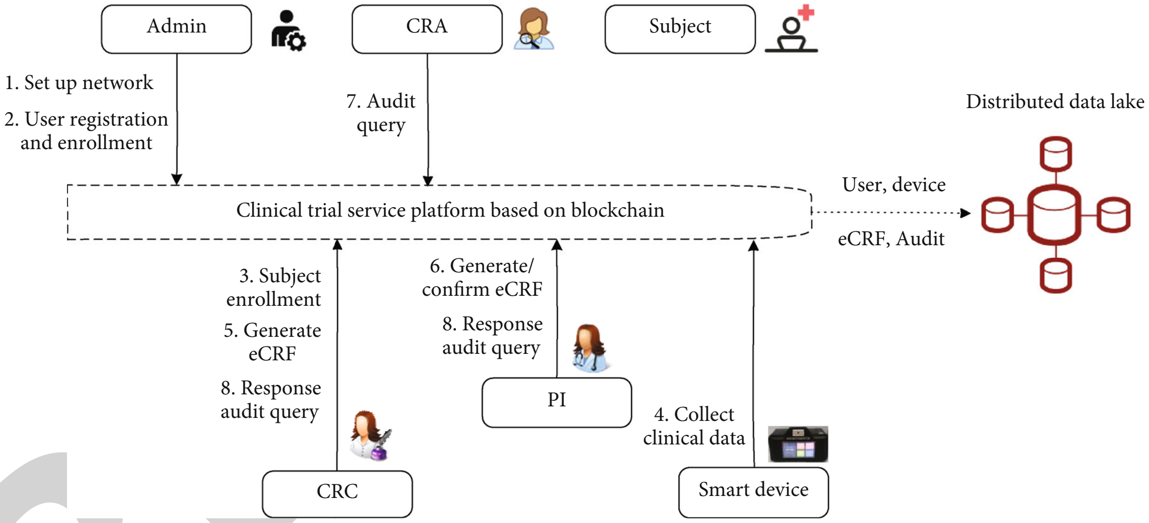


Figure 3: Roles in the proposed clinical trial service platform.

company on an outsource basis. In this paper, the blockchain management company's admin can set up the network to initiate a clinical trial but cannot perform transactions on the blockchain. Besides, the admin is the network manager of blockchain responsible for user registration and enrollment. CRC and PI are investigators responsible for the conduct of the clinical trial at a trial site. Generally, a clinical trial is conducted by a team of individuals at a clinical site. CRC interacts heavily with subjects, doing things like collecting and entering data. PI is the lead individual of the team responsible for all trial-related activities at the site. Their job is to ensure the protocol is executed precisely as written and may delegate trial-related activities to the clinical team members. CRA is the regulator who works in CRO to review submitted clinical data and those that conduct inspections. The subject is a direct participant of the clinical trial, either as a recipient of the investigational product or as a control. The process of subject enrollment is performed by CRC who has to screen the recruited subjects to see whether they meet the inclusion and exclusion criteria. As the most fundamental part of the clinical trial system, subjects need to transmit the biomedical data collected from smart devices throughout clinical trials. The distributed data lake serves as isolated storage that resides on the blockchain, also known as off-chain data storage. It is used to preserve all clinical data, covering user, device, eCRF, and audit data.

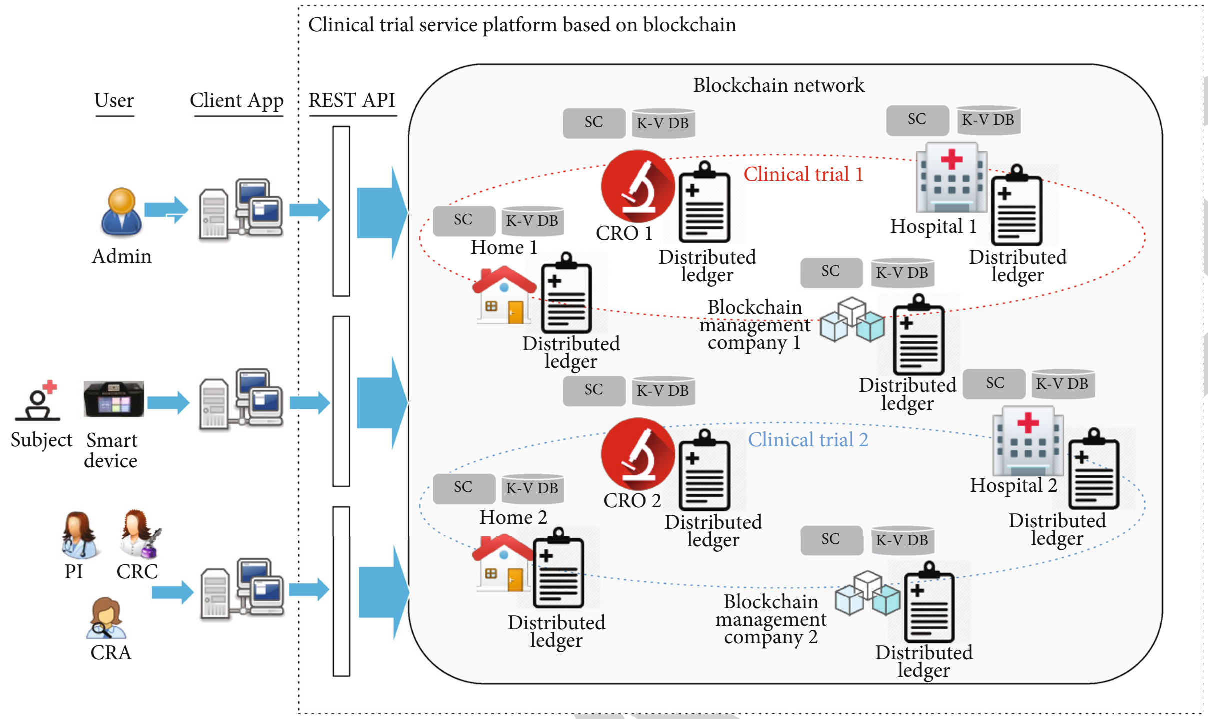


FIGURE 4: The system architecture of the clinical trial service platform.

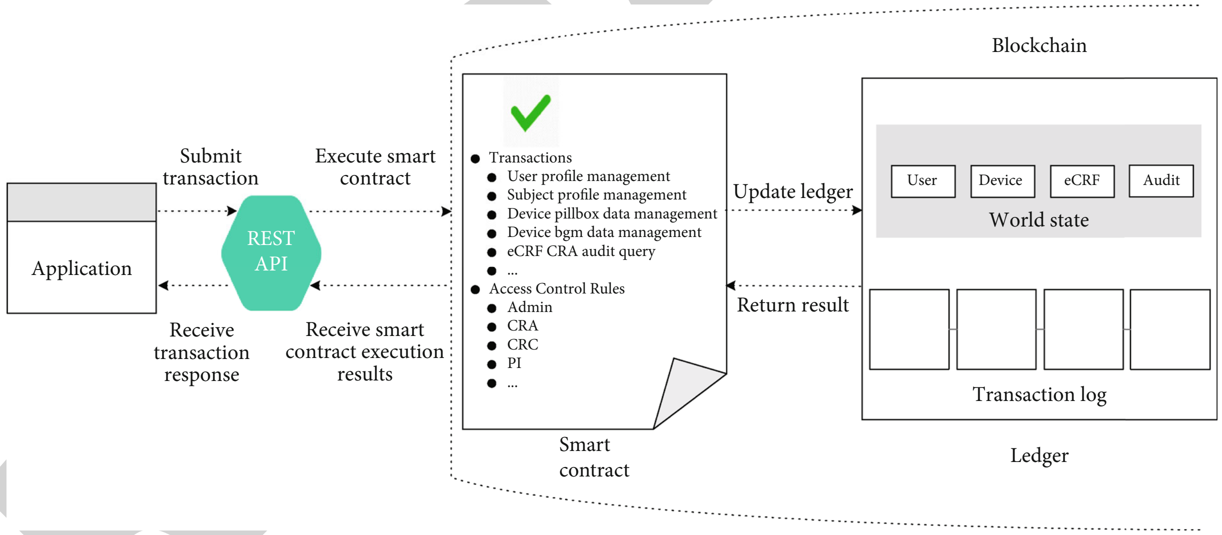


FIGURE 5: Interaction between application and blockchain via the smart contract.

4.3. System Architecture of the Clinical Trial Service Platform. As shown in Figure 4, these participants can access the blockchain network through the client applications that can communicate with REST APIs. REST APIs serve as an intermediate between external applications and the blockchain network. The smart contract (SC) is a decentralized application that defines the blockchain network's business logic according to the clinical protocol and automates the clinical trial process. The blockchain network appends an immutable record in the ledger to reflect changes resulting from transactions proposed by external applications, and a transaction response is returned as the response. The key value database (K-V DB) holds the current state of the ledger. Each time a new transaction is agreed upon and added, the K-V DB will update to reflect the latest transaction. The blockchain network comprises multiple channels with various organizations, identities, and data visibility rules. The proposed platform comprises multiple private networks, and each private network is specified for an individual clinical trial. In this way, this platform is appropriate for managing multiple clinical trials, and trial data is shared only between participants within the same network. This paper defines four organizations (blockchain management company, hospital, home, contract research organization), which are business entities

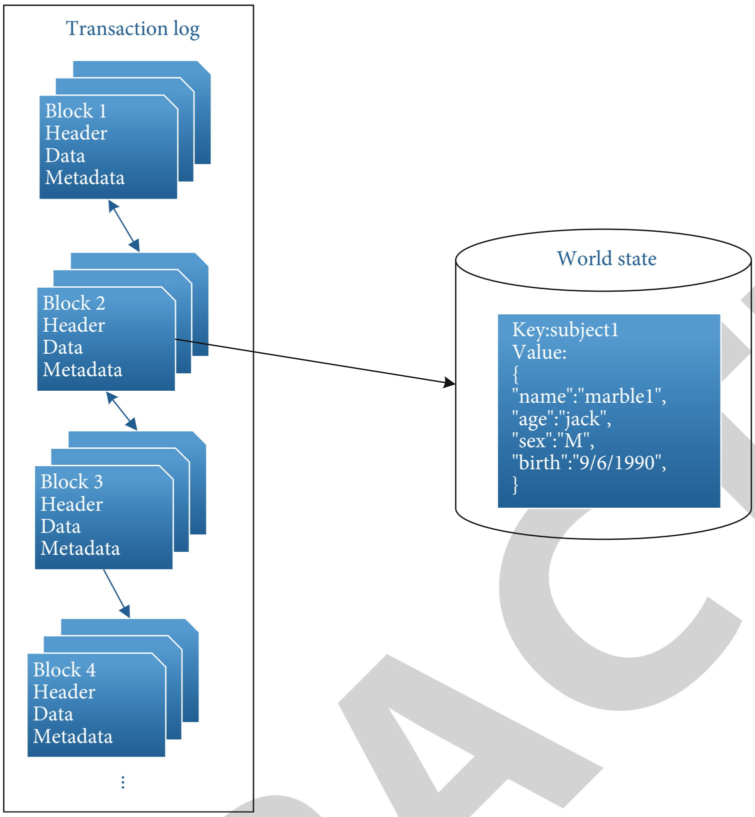


Figure 6: Sample ledger structure.

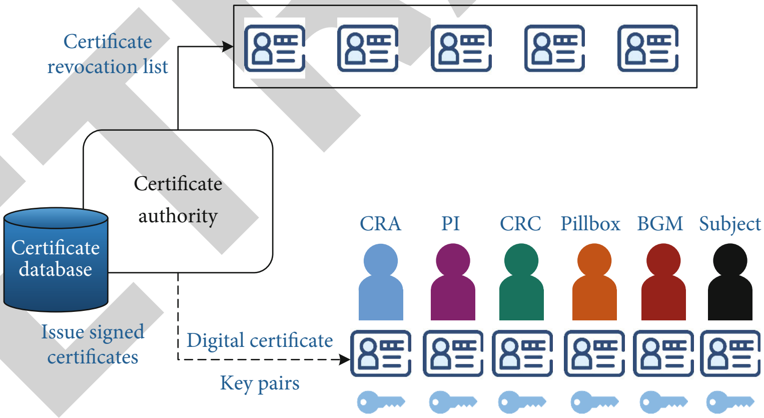


Figure 7: Identity management using certificate authorities.

that participate in the clinical trial. Each private network consists of four organizations, and each organization holds a copy of the ledger to maintain data consistency.

4.4. Smart Contract of the Clinical Trial Service Platform. Figure 5 illustrates the interaction between the application and the blockchain via the smart contract. The smart contract, together with the ledger, forms the blockchain network's heart from a high-level view. External applications invoke the smart contract by requesting the REST API to perform operations on the blockchain network. The smart contract programmatically accesses two distinct parts of the ledger: a transaction log that immutably holds the history of all transactions that ever happened in the blockchain and a world state that records these states' latest value. It can perform actions on the states stored in the world state or query transaction records in the transaction log.

The smart contract is defined in the package, and once deployed to the business network, all transactions are made available to applications. For example, CRC can manage the profile of a subject by invoking the corresponding transaction. Besides, the smart contract provides a specific rule list to evaluate whether or not the user can access or manipulate network resources. For instance, the network administrator

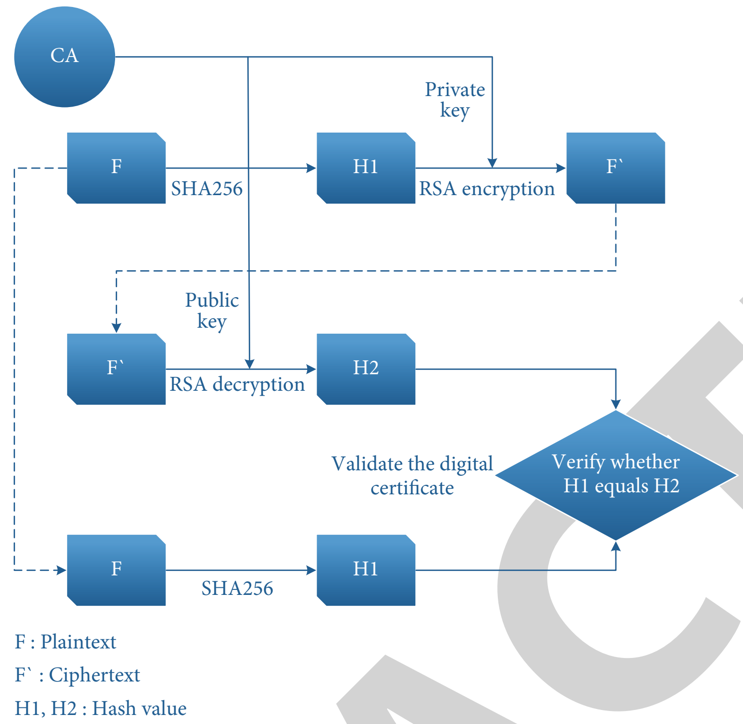


Figure 8: Process of certificate issuance and validation.

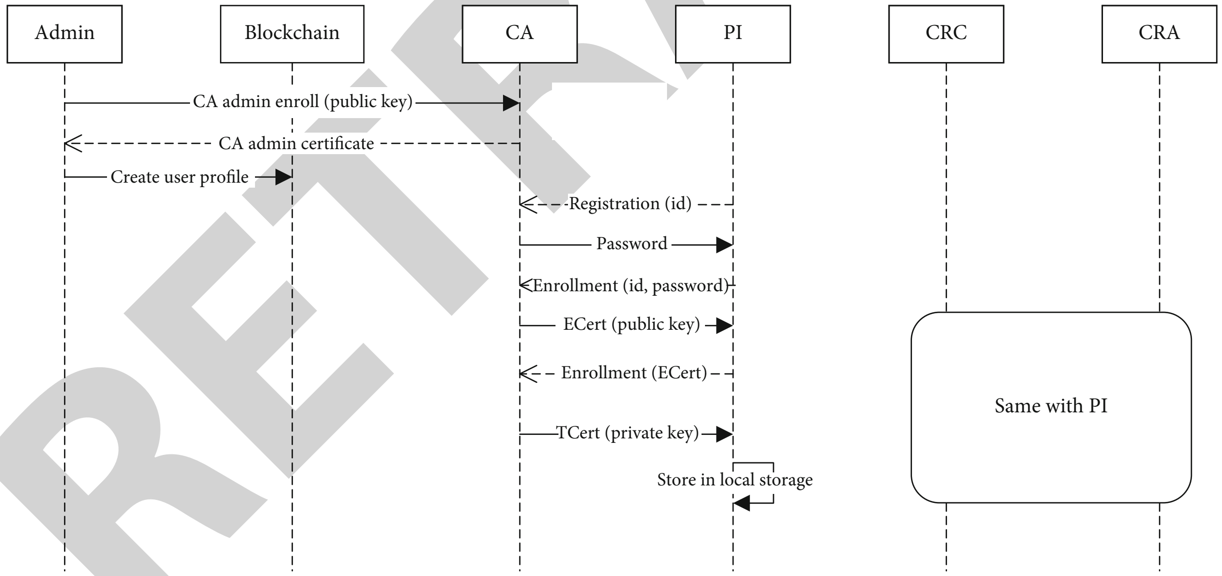


Figure 9: Execution process of user registration and enrollment.

is granted permission to perform operations on a network level like starting or stopping the network. Simultaneously, general users are only allowed to access specific resources or perform transactions on them.

The structure of the ledger comprises the transaction log and world state, as shown in Figure 6. This diagram indicates the contents inside the block and the world state associated with the block. A block consists of a header, a data section that contains multiple transactions, and the metadata. The world state stores the current state value of the ledger and changes incessantly against the updated state value, such as when a new subject is created or the subject information is modified. The world state provides rich query support that is flexible and efficient against large index data stores when users want to query the actual data value instead of the keys. The world state can significantly improve the transaction processing throughput since it is unnecessary to traverse the overall transaction log.

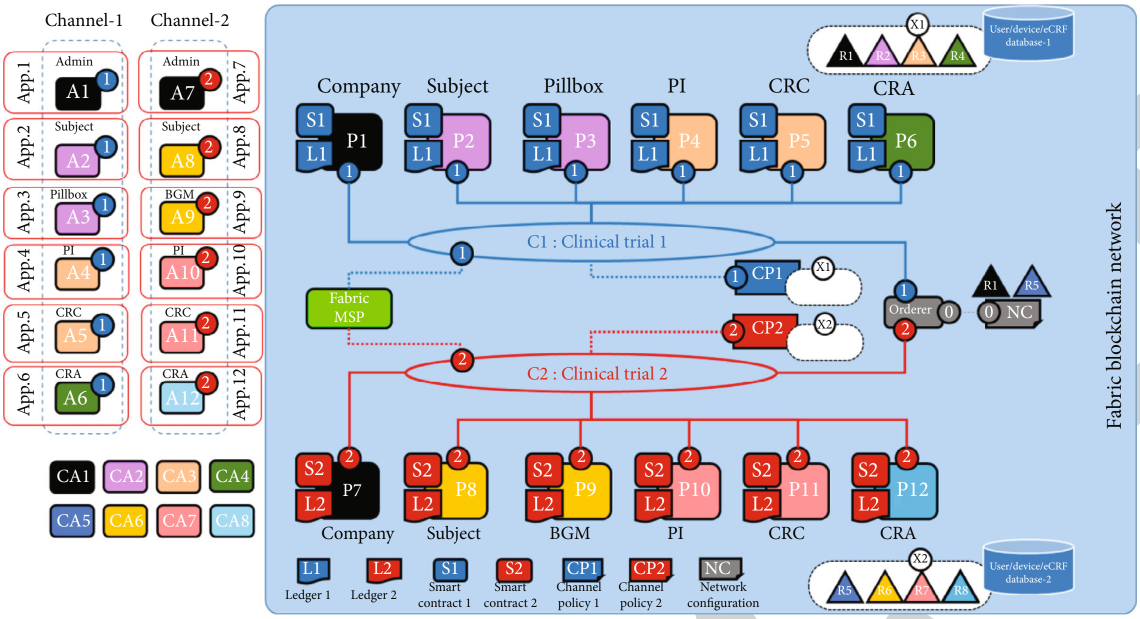


Figure 10: Prototype blockchain network topology.

4.5. Identity Management of the Clinical Trial Service Platform. Public Key Infrastructure (PKI) is aimed at facilitating the secure transfer of information for the clinical trial's network activities. In this paper, the PKI architecture utilizes certificate authorities (CA) responsible for the registration and issuance of digital certificates, as shown in Figure 7. The digital certificate certifies the ownership of a public key by the named subject of the certificate. All participants in the network (e.g., CRC, PI, and CRA) use the issued certificate to certify themselves in the messages they exchange in the network. This approach allows all participants to rely upon the digital signature made about the private key that corresponds to the issued certificate. Recipients of digitally signed messages can verify the received message's origin and integrity by checking whether the signature is valid with the sender's public key. The certificate database provides storage to preserve information about issued certificates, such as the status and validity period. The Certificate Revocation List (CRL) is a secure location where invalid certificates are stored and referred. This paper utilizes the X.509 standard [50], which defines the most commonly used public key certificate format.

Figure 8 illustrates the entire process from certificate issuance to validation in CA. The digital certificate is comprised of three elements: plaintext, ciphertext, and encryption method. The plaintext is sent to CA, which signs the certificate with its private key. Specifically, CA performs a SHA256 function on the plaintext to calculate the hash value H1. Then, CA uses its private key to encrypt the hash value H1 using the RSA algorithm to get the ciphertext F'. When validating the digital certificate, the ciphertext F' is decrypted with CA's public key to get a hash value H2. Then, CA performs a SHA256 function again on the plaintext F to get the hash value H1. If the value of H1 equals H2, it means the certificate is verified, indicating that the client holds a certificate issued by CA.

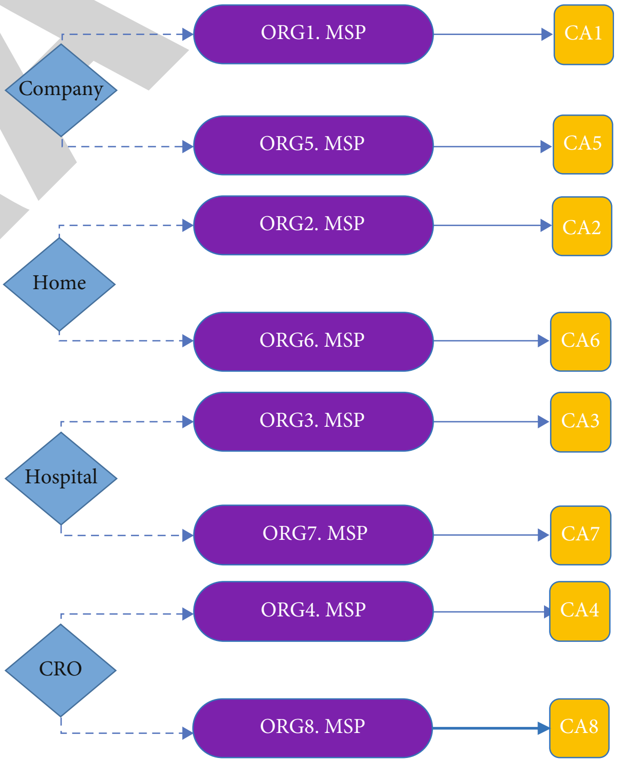


Figure 11: Mapping MSPs to organizations in the prototype.

Figure 9 represents the process of user registration and enrollment in the clinical trial service platform. When the network is set up, an admin is created as the registrar for the CA. The first step is to enroll the admin using the private and public key generated locally when the network is initialized. The public key is then sent to the CA, which returns an encoded certificate for the admin. The admin sends the

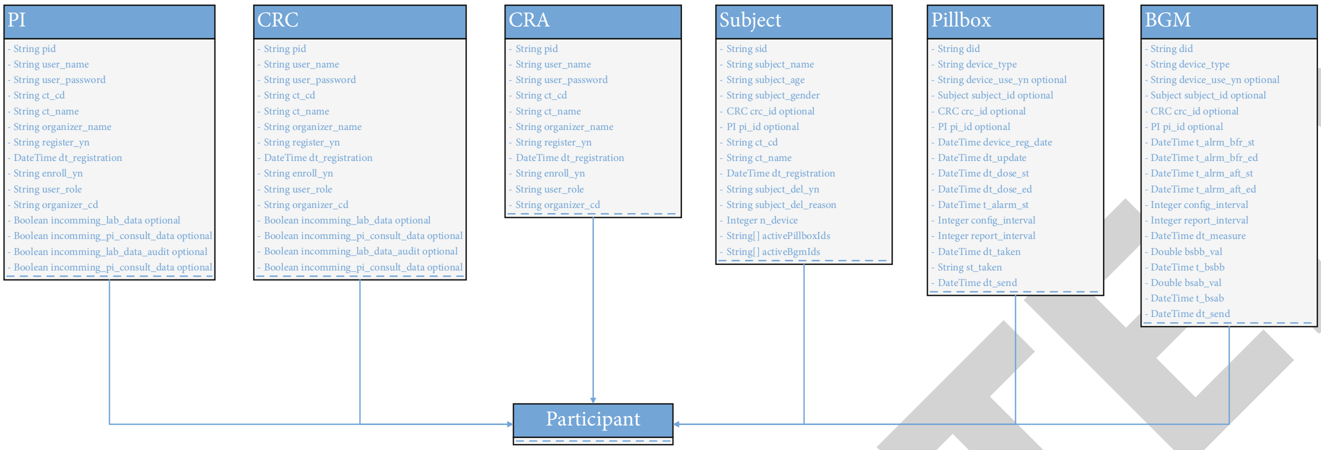


Figure 12: Class diagram showing the participant definition.

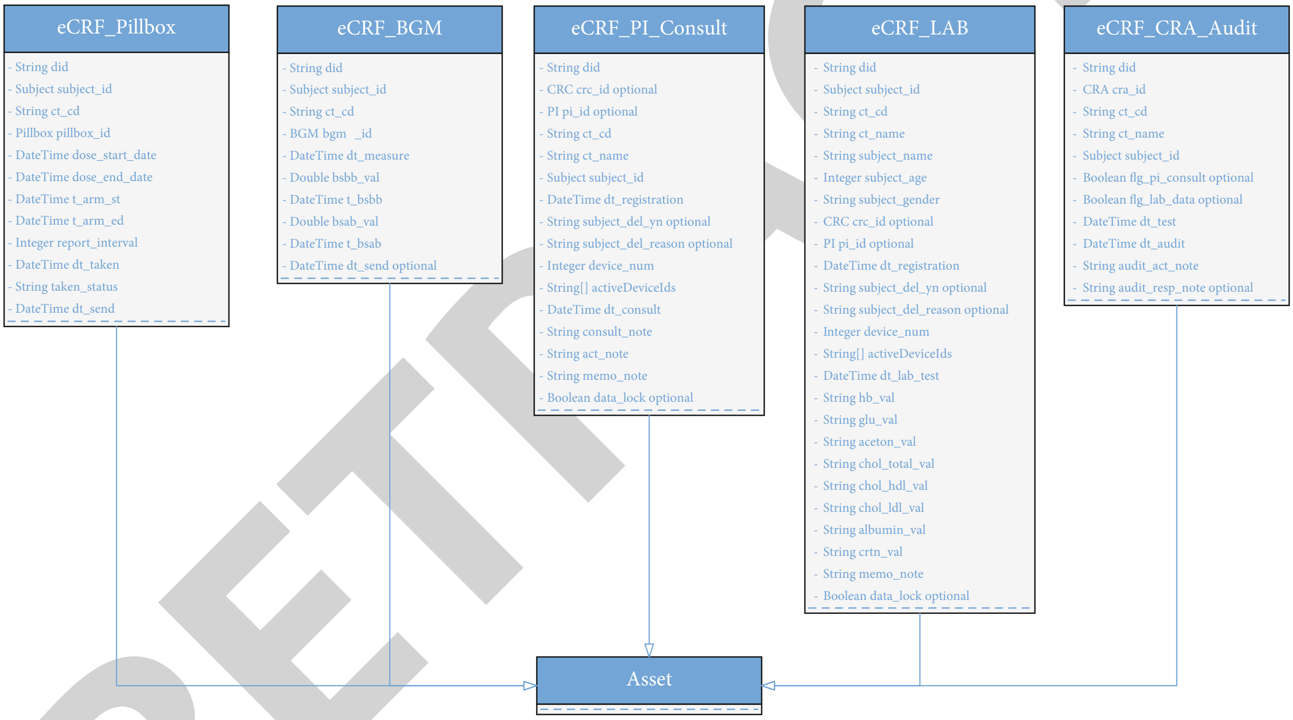


Figure 13: Class diagram showing the asset definition.

Table 2: Transaction definitions in the smart contract.

Transaction	Participant	Operation	Resource (participant, asset)
User profile management	Admin	ALL	CRC, CRA, PI
Subject management	CRC, PI	ALL	Subject
Device pillbox profile management	CRC, PI	ALL	Pillbox
Device BGM profile management	CRC, PI	ALL	BGM
eCRF pillbox data management	CRC, PI, pillbox	READ, CREATE	eCRF pillbox data
eCRF BGM data management	CRC, PI, BGM	READ, CREATE	eCRF BGM data
eCRF PI consult data management	CRC, PI, CRA	ALL	eCRF PI consult data
eCRF LAB data management	CRC, PI, CRA	ALL	eCRF lab data
eCRF CRA audit	CRC, PI, CRA	ALL	eCRF audit

```
rule NetworkAdminSystem {
            description: “Grant business network administrators full access”
            participant: “org.hyperledger.composer.system.NetworkAdmin”
            operation: ALL
            resource: “org.hyperledger.composer.system.**”
            action: ALLOW
}
rule CRC_to_Subject {
            description: “Grant CRC access to the subjects created”
            participant: “org.clinical.trial.CRC”
            operation: ALL
            resource: “org.clinical.trial.Subject”
            action: ALLOW
}
rule CRA_to_Subject {
            description: “Grant CRA access to the subjects created”
            participant: “org.clinical.trial.CRA”
            operation: READ
            resource: “org.clinical.trial.Subject”
            action: ALLOW
}
…
```

Algorithm 1: Access control rules in the smart contract.

Table 3: Sample REST API endpoints.

URI	Verb	Description
/api/CreateBgmTransaction	POST	Create device BGM profile
/api/UpdateBgmTransaction	POST	Update device BGM profile
/api/DeleteBgmTransaction	POST	Delete device BGM profile
/api/CreatePillboxTransaction	POST	Create device pillbox profile
/api/UpdatePillboxTransaction	POST	Update device pillbox profile
/api/DeletePillboxTransaction	POST	Delete device pillbox profile
/api/CreateSubjectTransaction	POST	Create subject profile
/api/UpdateSubjectTransaction	POST	Update subject profile
/api/DeleteSubjectTransaction	POST	Delete subject profile
/api/CreateECRFpillboxTransaction	POST	Create eCRF pillbox data
/api/CreateECRFbgmTransaction	POST	Create eCRF BGM data
/api/CreateECRFpiConsultTransaction	POST	Create eCRF PI consult data
/api/CreateECRFlabTransaction	POST	Create eCRF lab data
/api/CreateCRAauditTransaction	POST	Create eCRF audit data
/api/ConfirmECRFpiConsultTransaction	POST	Confirm eCRF PI consult data
/api/ConfirmECRFlabTransaction	POST	Confirm eCRF lab data
/api/system/historian	GET	Retrieve all historian transactions
/api/system/identities/{id}/revoke	POST	Revoke the specified identity
/api/systemidentities/issue	POST	Issue an identity to the specified participant

blockchain’s request to create a new user profile, which contains an id. For example, the participant can send a registration request to the CA with this id, and the CA returns a password accordingly. An enrollment request is then sent to the CA, and the id and password are obtained in the registration process. In response, the CA returns the enrollment certificate (ECert) along with the public key. The ECert is used to request the Transaction Certificate (TCert), and the CA returns the TCert along with the private key for signing the transactions. These credentials are then stored in the wallet, and the user can use them to interact with the blockchain.

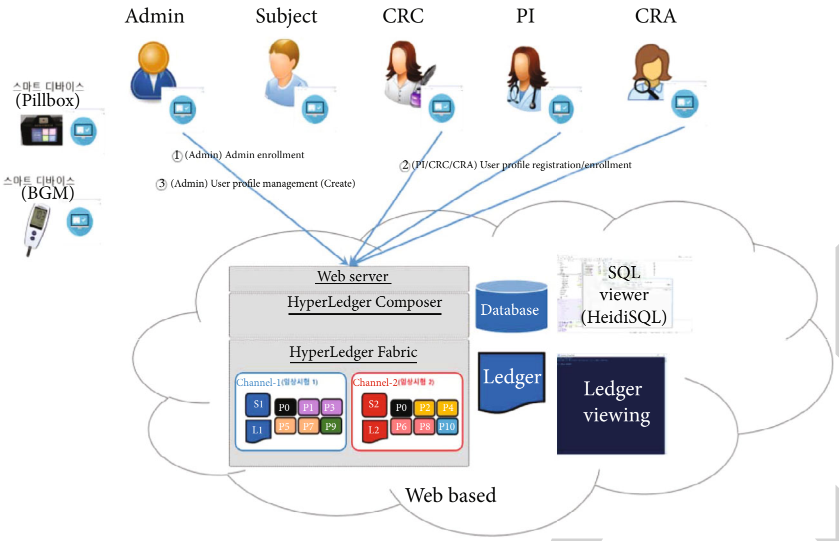


Figure 14: User registration and enrollment.

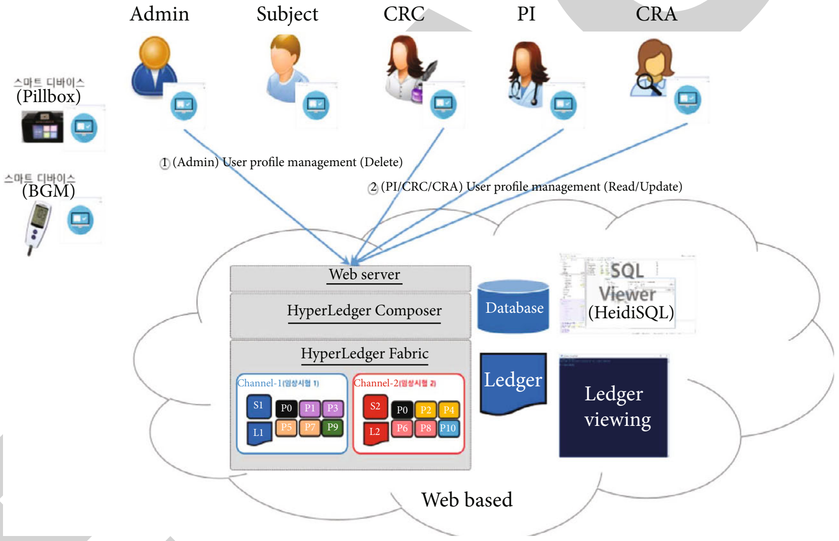


Figure 15: User profile management.

5. Experiment: Applying the Proposed Platform to Create Prototype Application

5.1. Overview of the Prototype Application. We utilize the Hyperledger Composer [51] to implement the blockchain application based on Hyperledger Fabric. All of the fabric network entities ran in the Docker environment hosted in a single Linux virtual machine. The smart contract is written in JavaScript by using the Visual Studio Code and further deployed to all the network peers via composer-cli. A REST server is generated by the composer-rest-server from the deployed business network to be consumed by HTTP or REST clients. The REST server provides create, read, update, and delete (CRUD) operations to manipulate the states of assets or participants and allows transactions to be submitted or retrieved with queries. The data transmission between the external client and the REST is secured using the GitHub Passport middleware; thus, every client must be authenticated before they are permitted to call the REST server's APIs. Hyperledger Explorer [32] is used as a visualization tool to observe blocks, transactions, and other relevant information stored in the ledger. The blockchain web application is implemented with HTML, CSS, and JavaScript. The Apache Tomcat web server ships as a host environment capable of serving the web application. This web application can interact with the fabric network to operate on the resources and perform functions via the REST server's endpoint APIs.

As shown in Figure 10, the fabric network is set up and exploited by eight organizations with corporately decided and signed agreements. An organization refers to a managed group of members, such as the blockchain management company, home, hospital, and CRO. A single membership service provider (MSP) defines the list of members of an organization,

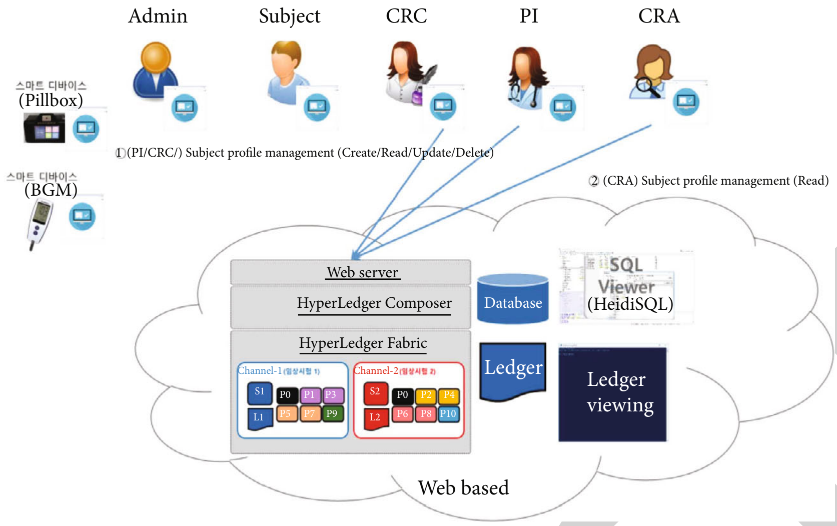


Figure 16: Subject profile management.

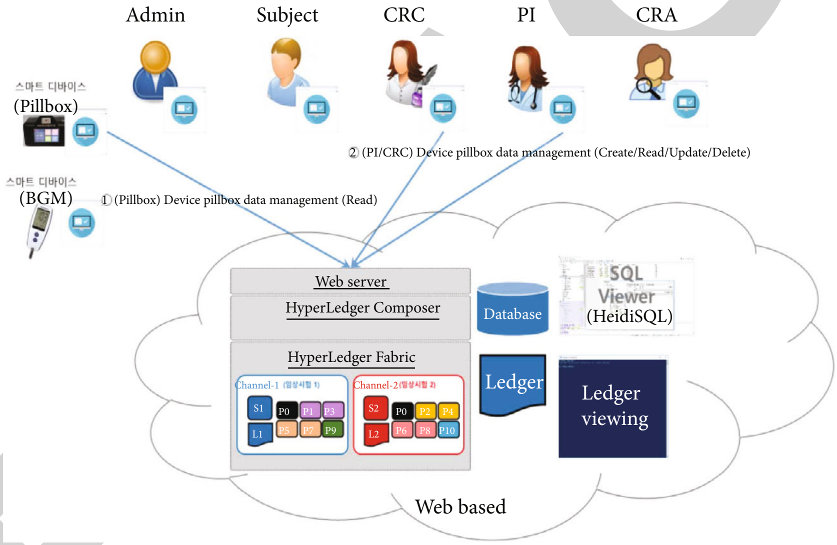


Figure 17: Device pillbox profile management.

as described in Figure 11. In this paper, the organizations manage their members under MSPs, representing different organizational groups in independent clinical trials. It is worth noting that the different MSPs can be used to present the same organization group. For example, the company organization consists of two MSPs, ORG1.MSP and ORG5.MSP, which represent the same blockchain management company that performs different clinical trials. Organizations R1 and R5, which refer to the blockchain management companies, have been empowered to initialize the network.

Each organization (e.g., organization R1) in channel C1 is connected with a client application that can submit transactions and other organizations within the same channel. Similarly, client applications connected with organizations in channel C2 can perform transactions within channel C2. It is worth mentioning that one organization can also have multiple client applications such as R2 and R6. Each peer in channel C1 keeps the same copy of the ledger L1 while peer nodes in channel C2 keep the same copy of the ledger L2. The network is under the control of policy rules specified in network configuration (NC), governed by R1 and R5. Channel C1 is governed in terms of the rules defined by channel policy CP1.

Similarly, channel C2 is governed in terms of the rules defined by channel policy CP2. The ordering service supports applications in both channels and orders transactions into blocks. Besides, each of the eight organizations has a preferred CA.

5.2. Smart Contract Implementation. The smart contract is modeled and packaged as a compressed business network definition, consisting of participants, assets, and transactions. Figures 12 and 13 describe the unified class diagram of participant and asset definition, respectively. In this business

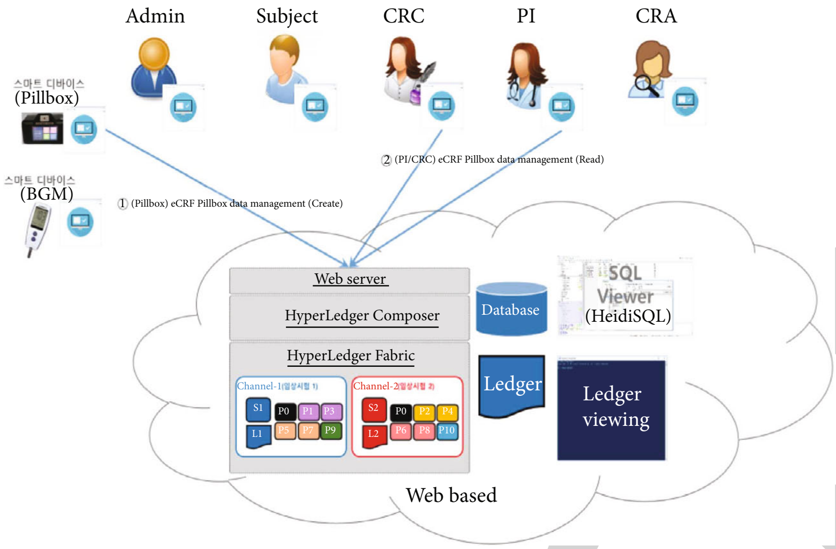


Figure 18: eCRF pillbox data management.

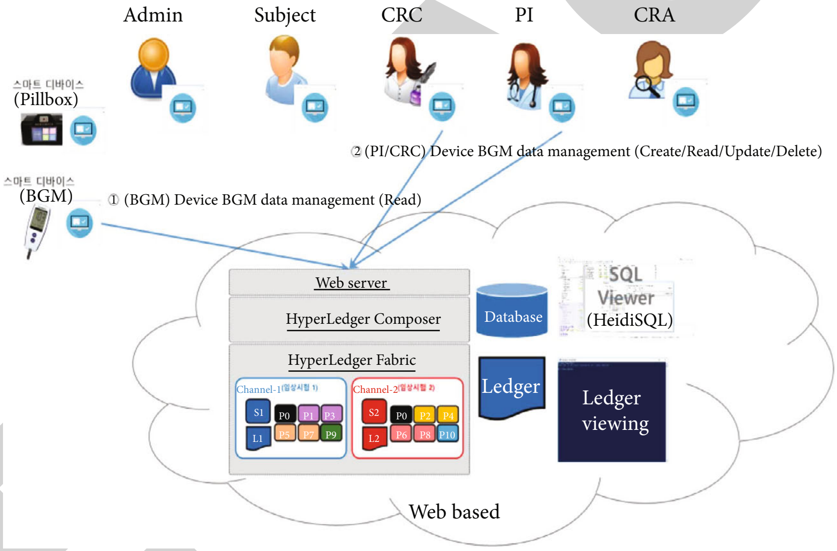


Figure 19: Device BGM profile management.

network, a participant or an asset defined in the business network will generate a corresponding instance founded in the model.

Table 2 describes the defined transactions in the smart contract. The participant is the individual (e.g., CRC and CRA) who enrolls in the business network. The asset represents either a tangible property (e.g., pillbox) or intangible data (e.g., eCRF pillbox data and eCRF BGM data). A participant proposes transactions to operate against a specified resource, such as modifying the participant's info. The prototype application supports four types of operations (read, create, update, delete) that the transaction can perform on the particular network resource. ALL is a particular term to determine that the transaction supports all operations. For example, the user can have full and unhindered access to its profile. Once the smart contract is deployed to the blockchain network, all transactions described in the smart contract are made available to applications.

We specify various access control rules to allow or deny access to resources depending on the user's identity, bound to a specific participant. As shown in Algorithm 1, the description states the rule in plain language; participant defines the type of participant affected by the rule, resource defines the type of resource affected by the rule, condition defines the condition that triggers the rule, action describes the permission type (ALLOW or DENY), and the operation is the action allowed or denied by the rule. For example, the CRC has full operation permissions on the subject while the CRA can only read its profile.

Table 3 presents the sample list of some REST API endpoints used to call transactions provided by the smart contract. Each API contains a URI and verbs such as GET,

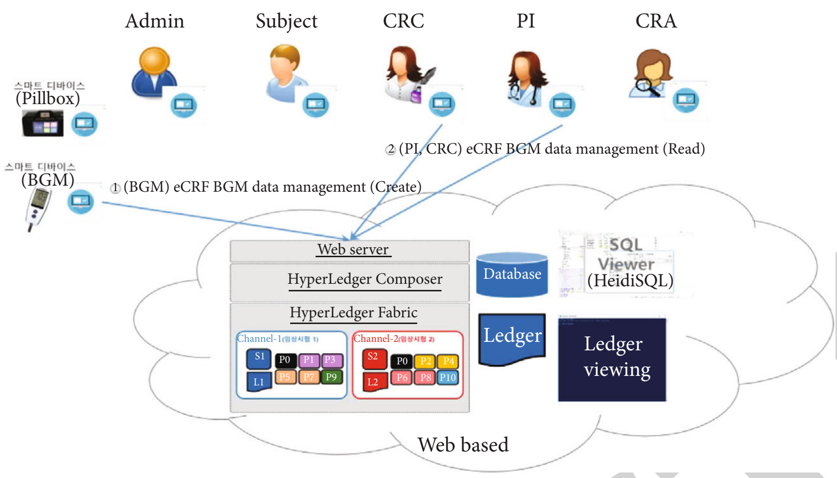


Figure 20: eCRF BGM data management.

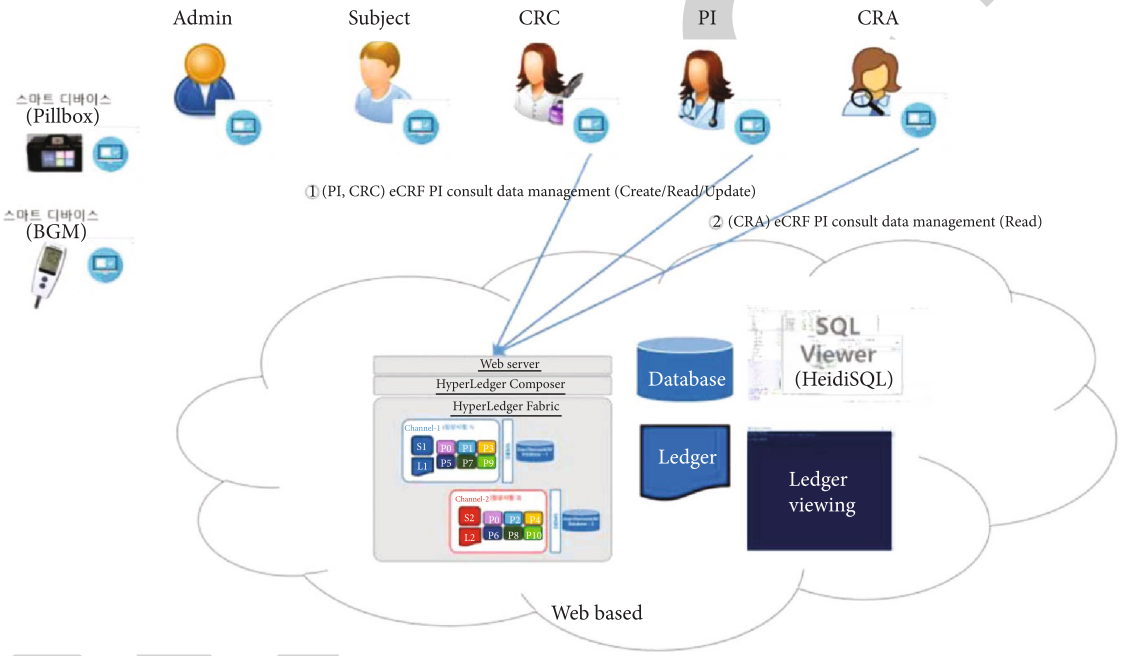


Figure 21: eCRF PI consult data management.

POST, PUT, and DELETE. The URI specifies the endpoint path, and the verb presents the specific operation to be performed on the resource.

5.3. Service Scenario of the Clinical Trial Service Platform. The proposed platform contains ten different service scenarios in the clinical trial. The first scenario, referred to as *user registration and enrollment*, shows that the admin can create the user profile for each participant, as shown in Figure 14. This operation requires the admin to log onto the network before PI, CRC, and CRA can register their identity and enroll in the CA certificate system. Figure 15 illustrates the second scenario, *user profile management*. The admin can delete the user profile while the PI, CRC, and CRA can only read and update their profiles. Figure 16 shows the third scenario, *subject profile management*, in which the CRC and PI can create, read, update, and delete a subject while the CRA can only read the subject's profile. Figure 17 illustrates the fourth scenario, named *device pillbox profile management*, in which the CRC and PI can create, read, update, and delete the device pillbox profile while the pillbox can read its profile. Figure 18 describes the fifth scenario, *eCRF pillbox data management*, where the pillbox generates the eCRF pillbox data, and the CRC and PI can read the eCRF pillbox data. Figures 19 and 20 show the sixth and seventh scenarios, namely, device BGM profile management and eCRF BGM data management, similar to the fourth and fifth scenarios but with the BGM (blood glucose meter). Figure 21 describes the eighth scenario, *eCRF PI consult data management*, in which the PI and CRC can create, read, and update the eCRF PI consult data while the CRA can read the eCRF PI consult data. Figure 22 shows the ninth scenario, *eCRF LAB data*

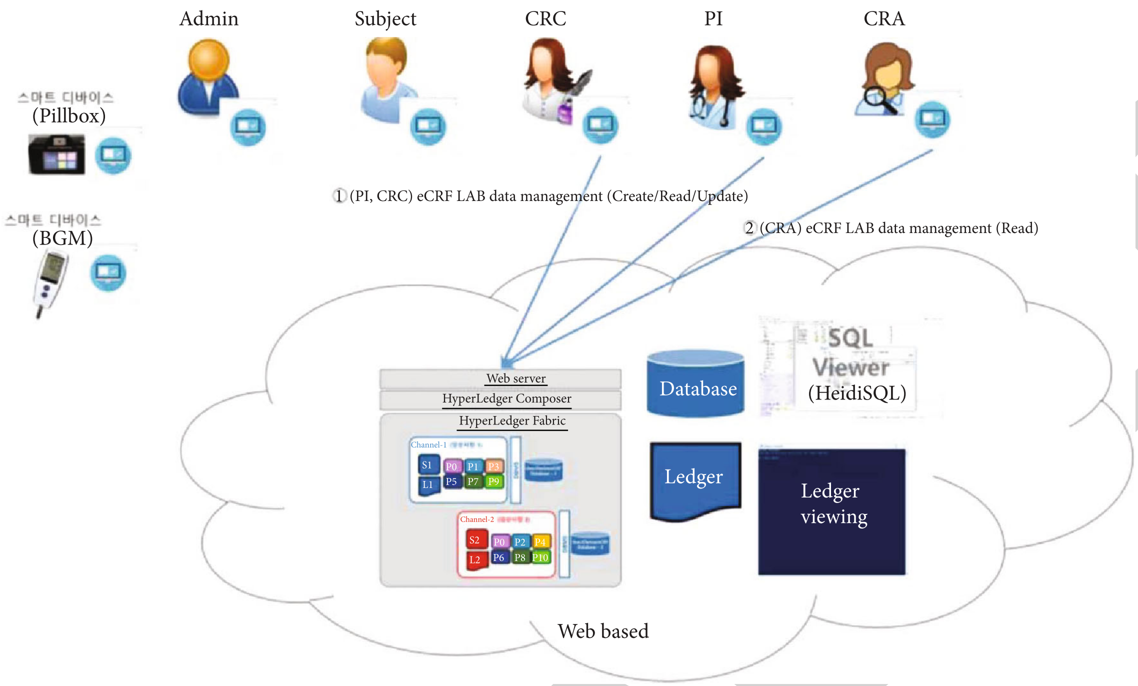


Figure 22: eCRF lab data management.

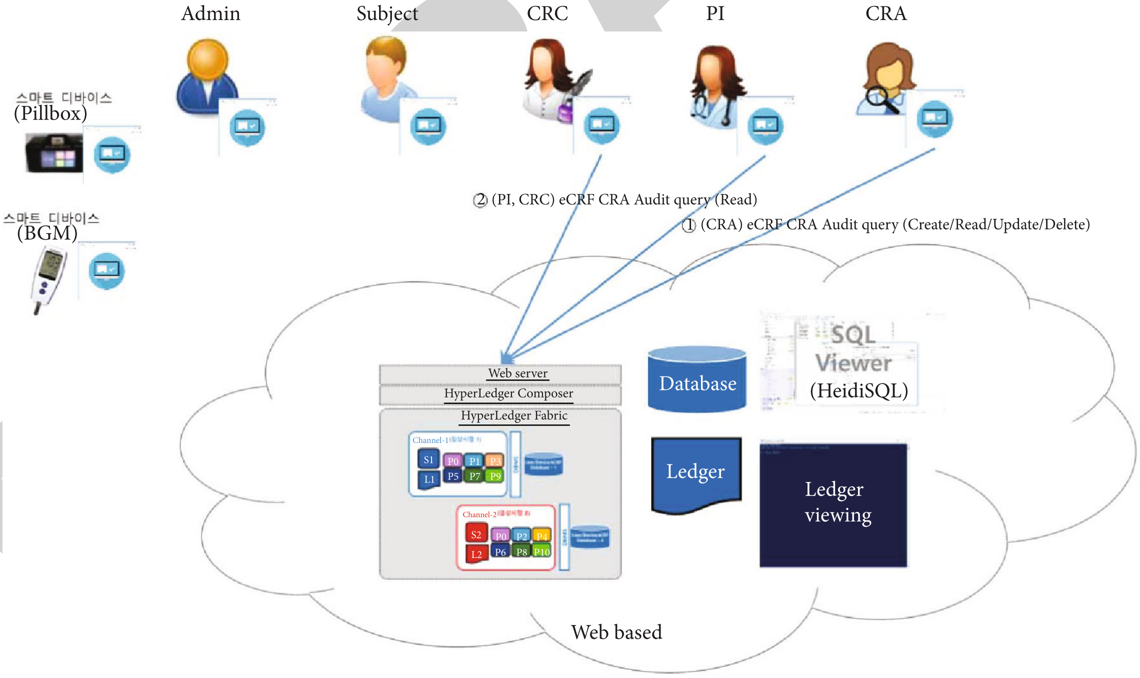
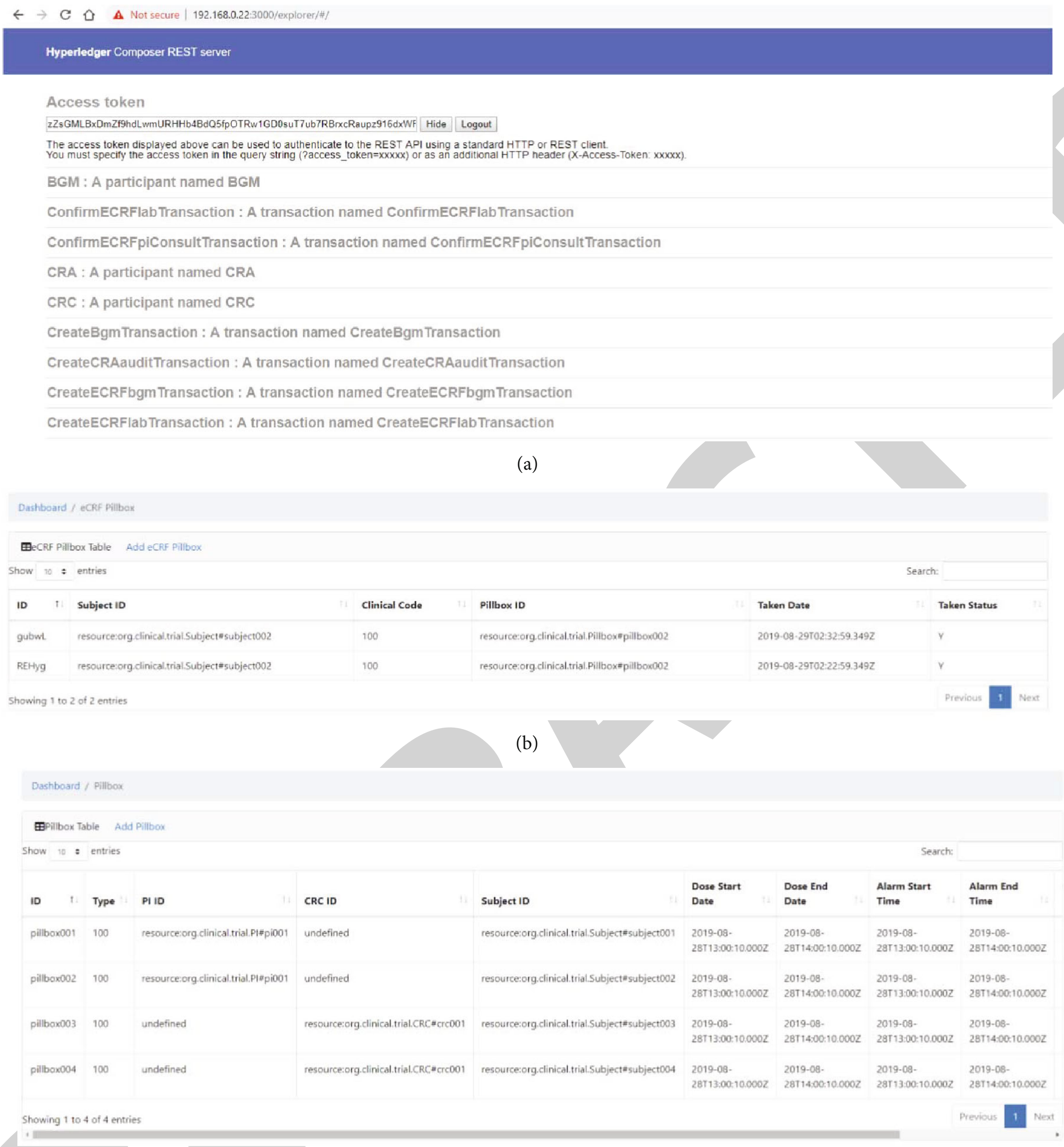


Figure 23: eCRF CRA audit query.

management, in which the PI and CRC can create, read, and update the eCRF lab data while the CRA can read the eCRF lab data. Figure 23 describes the last scenario, *eCRF CRA audit query*, where the CRA can create, read, and update the eCRF audit query while the CRC and PI can read and update the audit data.

5.4. Experiment Results. Figure 24 is a screenshot of the experiment results, presenting the following features (a) REST server interface, (b) eCRF pillbox data, and (c) device pillbox data. The web application authenticates the REST server by navigating to the Github OAuth provider's path. The REST server will redirect the request to GitHub to



Hyperledger Composer REST server

Access token

zZsGMLBxDmZf9hdLwmURHHb4BdQ5fpOTRw1GD0suT7ub7RBrxcRaupz916dxWF

The access token displayed above can be used to authenticate to the REST API using a standard HTTP or REST client.
You must specify the access token in the query string (?access_token=xxxxx) or as an additional HTTP header (X-Access-Token: xxxxx).

BGM : A participant named BGM

ConfirmECRFlabTransaction : A transaction named ConfirmECRFlabTransaction

ConfirmECRFpiConsultTransaction : A transaction named ConfirmECRFpiConsultTransaction

CRA : A participant named CRA

CRC : A participant named CRC

CreateBgmTransaction : A transaction named CreateBgmTransaction

CreateCRAauditTransaction : A transaction named CreateCRAauditTransaction

CreateECRFbgmTransaction : A transaction named CreateECRFbgmTransaction

CreateECRFlabTransaction : A transaction named CreateECRFlabTransaction

(a)

ID	Subject ID	Clinical Code	Pillbox ID	Taken Date	Taken Status
gubwL	resource:org.clinical.trial.Subject#subject002	100	resource:org.clinical.trial.Pillbox#pillbox002	2019-08-29T02:32:59.349Z	Y
REHyg	resource:org.clinical.trial.Subject#subject002	100	resource:org.clinical.trial.Pillbox#pillbox002	2019-08-29T02:22:59.349Z	Y

Showing 1 to 2 of 2 entries

(b)

ID	Type	PI ID	CRC ID	Subject ID	Dose Start Date	Dose End Date	Alarm Start Time	Alarm End Time
pillbox001	100	resource:org.clinical.trial.PI#pi001	undefined	resource:org.clinical.trial.Subject#subject001	2019-08-28T13:00:10.000Z	2019-08-28T14:00:10.000Z	2019-08-28T13:00:10.000Z	2019-08-28T14:00:10.000Z
pillbox002	100	resource:org.clinical.trial.PI#pi001	undefined	resource:org.clinical.trial.Subject#subject002	2019-08-28T13:00:10.000Z	2019-08-28T14:00:10.000Z	2019-08-28T13:00:10.000Z	2019-08-28T14:00:10.000Z
pillbox003	100	undefined	resource:org.clinical.trial.CRC#crc001	resource:org.clinical.trial.Subject#subject003	2019-08-28T13:00:10.000Z	2019-08-28T14:00:10.000Z	2019-08-28T13:00:10.000Z	2019-08-28T14:00:10.000Z
pillbox004	100	undefined	resource:org.clinical.trial.CRC#crc001	resource:org.clinical.trial.Subject#subject004	2019-08-28T13:00:10.000Z	2019-08-28T14:00:10.000Z	2019-08-28T13:00:10.000Z	2019-08-28T14:00:10.000Z

Showing 1 to 4 of 4 entries

(c)

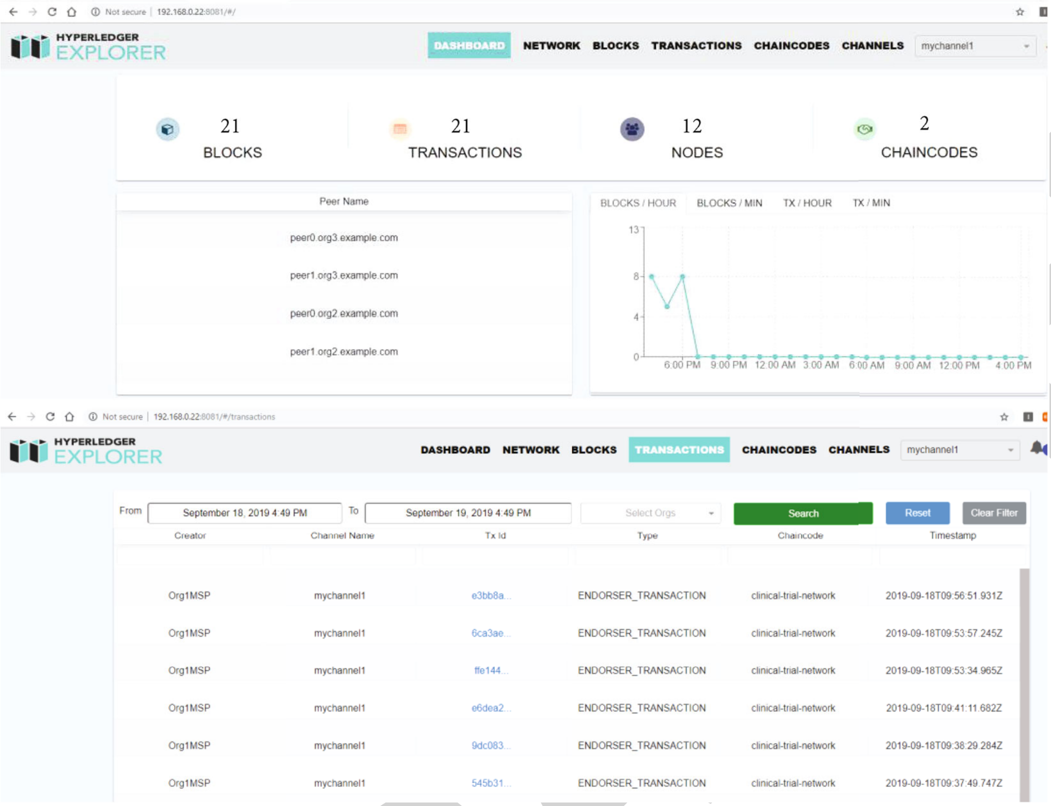
Figure 24: Screenshot of experiment results.

perform the OAuth authentication flow. After generating the token, Github will redirect back to the REST server and display the access token. The web application then forwards the user requests along with the access token to the REST server. The REST server invokes the smart contract's relevant functions to perform business transactions and returns responses routed back to the web application. For the sake of simplicity, the source code of the prototype application was uploaded to GitHub [52]. The implementation results of our prototype application were videotaped and put on the Internet (https://www.youtube.com/watch?v=Nk3P6PGSksY).

As shown in Figure 25, we can monitor various blockchain data on the browser as Hyperledger Explorer is integrated with the network. Hyperledger Explorer provides a dashboard that gives an overview of the network, such as the number of blocks, transactions, and peer nodes. It also provides an entry to access the detailed information; for example, we can get the details of a transaction in terms of transaction ID, type, creator, channel, and timestamp.

6. Performance Evaluation

6.1. Evaluation Setup. The prototype application's performance was evaluated using an open-source benchmark simulation tool called Hyperledger Caliper [53]. The experiment was performed using 10 clients in a two-channel network,



21 BLOCKS

21 TRANSACTIONS

12 NODES

2 CHAINCODES

Peer Name
peer0.org3.example.com
peer1.org3.example.com
peer0.org2.example.com
peer1.org2.example.com

From: September 18, 2019 4:49 PM To: September 19, 2019 4:49 PM

Creator	Channel Name	Tx Id	Type	Chaincode	Timestamp
Org1MSP	mychannel1	e3bb8a...	ENDORSER_TRANSACTION	clinical-trial-network	2019-09-18T09:56:51.931Z
Org1MSP	mychannel1	6ca3ae...	ENDORSER_TRANSACTION	clinical-trial-network	2019-09-18T09:53:57.245Z
Org1MSP	mychannel1	ffe144...	ENDORSER_TRANSACTION	clinical-trial-network	2019-09-18T09:53:34.965Z
Org1MSP	mychannel1	e6dea2...	ENDORSER_TRANSACTION	clinical-trial-network	2019-09-18T09:41:11.682Z
Org1MSP	mychannel1	9dc083...	ENDORSER_TRANSACTION	clinical-trial-network	2019-09-18T09:38:29.284Z
Org1MSP	mychannel1	545b31...	ENDORSER_TRANSACTION	clinical-trial-network	2019-09-18T09:37:49.747Z

Figure 25: Screenshot of Hyperledger Explorer.

consisting of 8 organizations with 12 endorser peer nodes in total, as depicted in Table 4. The block size is set to 10 transactions per block, and a new block is formed every 250 ms. The ordering service is in solo mode, which consists only of a single ordering node. The experiment's scripts were specified to target two functions of our prototype: eCRF lab data generation and eCRF lab data query transaction since the user most frequently invokes these two transactions. Ten rounds of tests with a fixed number of transactions were performed by varying the send rate from 100 tps to 1000 tps using different transaction data sizes. The experiment results were averaged to reduce the probability of errors resulting from system overload and network congestion.

6.2. Throughput and Network Latency Evaluation. The throughput and latency are two standard performance metrics to evaluate the performance of the blockchain network. The throughput can be further divided into two subcategories concerning the operations to deal with. Read throughput is a specific measure to count the number of read operations completed in a defined period, expressed as read per second (rps). Read throughput is not used as a central performance parameter to measure the blockchain. Most of the systems are typically deployed adjacent to the blockchain to achieve significant reading and query efficiency. Transaction throughput is the rate at which valid transactions are committed by the blockchain in a defined period, expressed as transaction per second (tps). Transaction throughput is not the measure at a single node but across all nodes of the whole network.

Table 4: Experiment configuration parameters for performance evaluation.

Configurable parameters	Values
Number of clients	10
Number of channels	2
Number of organizations	8
Number of peers	12
Number of transactions	5000
Send rate	100–1000 tps
Transaction data size	50 kb, 250 kb, 500 kb
Block timeout	250 ms
Block size	10 transactions per block
Orderer type	Solo

$$\text{Read Throughput} = \text{Total readoperations/timein seconds}, \tag{1}$$

$$\begin{aligned}&\text{Transaction Throughput}\\&\quad = \text{Total validtransactions/totaltime in seconds}.\end{aligned} \tag{2}$$

Latency can also be separated into two subcategories in terms of the type of operations. Read latency measures the total time to submit a read request and receive the reply.

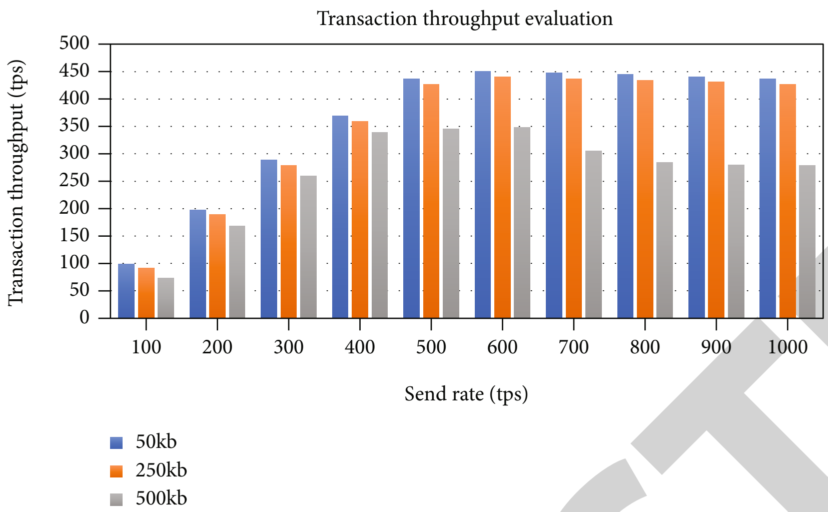


Figure 26: Impact of the send rate and transaction data size on blockchain transaction throughput.

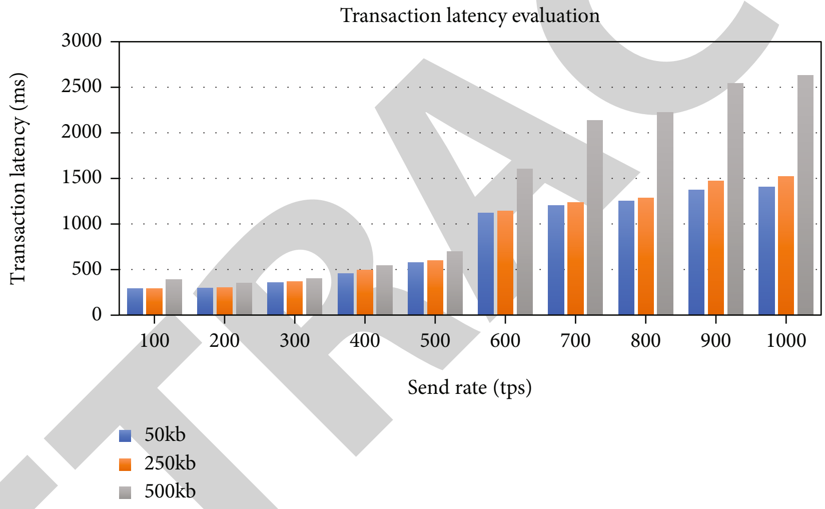


Figure 27: Impact of the send rate and transaction data size on blockchain transaction latency.

Transaction latency measures the time the entire network takes to validate a transaction, covering the broadcasting time and the allocation time spent by the consensus algorithm.

$$\text{Read Latency} = \text{Response received time} - \text{submission time}, \tag{3}$$

$$\begin{aligned}&\text{Transaction Latency}\\&\quad = \text{Confirmation time} - \text{submission time}.\end{aligned} \tag{4}$$

The transaction throughput increases linearly with the increase in send rate until it flattens out at around 600 tps, the saturation point, as plotted in Figure 26. Figure 27 plots the evaluation results of the observed transaction latency. The latency rises slowly as the send rate increases; however, it increases significantly when the send rate overs the saturation point.

The evaluation of read throughput and read latency is performed using the same experimental method. As shown in Figure 28, the average read throughput increases linearly as the send rate increases. The graph in Figure 29 shows that average read latency has a relatively small increase with the send rate increase. Unlike the eCRF lab data generation transaction, the throughput of eCRF lab data query transaction increases linearly with the increase in send rate. The reason is that eCRF lab data generation transaction requires much more computing power as this function directly modifies the ledger state. In contrast, eCRF lab data query transaction only performs read operations on the ledger. Moreover, it is evident from the results that the transaction data size has a strong impact on the network performance. The transaction throughput decreases and the transaction latency increases with the growth of the transaction data size.

7. Security Verification

7.1. Data Privacy. In the prototype application, data privacy is achieved using either channel at the network level or the

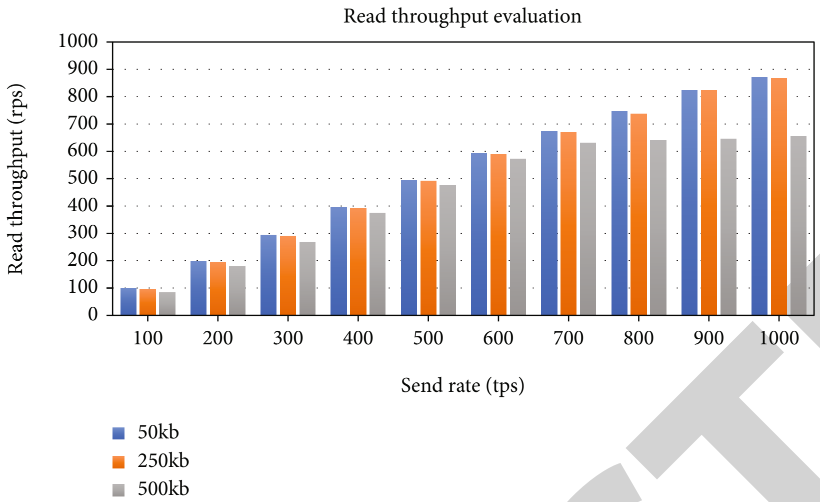


FIGURE 28: Impact of the send rate and transaction data size on blockchain read throughput.

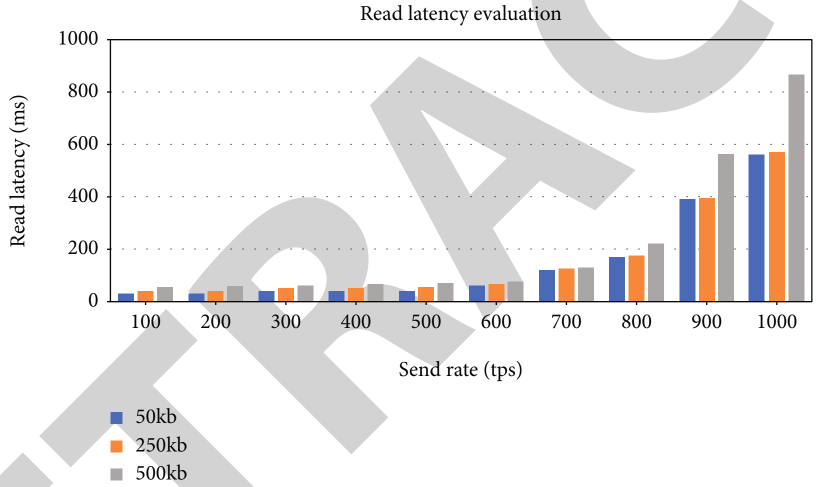


FIGURE 29: Impact of the send rate and transaction data size on blockchain read latency.

smart contract's access control rules. Each of these channels represents an isolated clinical trial test in which only authorized participants are allowed to access the data from the ledger; hence, data visibility is limited. The peer on the same channel shares a ledger, and the transaction peer needs to obtain the channel's recognition before it can join the channel and transact with others. The access control rules provide declarative access control over the elements of the business network. By defining access control rules, we can determine which users can perform the specific operation on elements over the network. The prototype application differentiates between access control for elements within a business network and access control for network administrative changes. These rules are evaluated in order, and the first rule whose condition matches determines whether access is granted or denied.

7.2. Data Transmission. In the prototype application, the client and the REST server's data transmission is secured using the proper authentication strategy. There are many authentication strategies one can choose from, including a mix of social media such as Facebook, Google, GitHub, and enterprise strategies such as SAML, JSON Web Tokens (JWT), or LDAP. To simplify the implementation, the Github authentication provider can authenticate the client to the REST server before it is permitted to access the business network elements. The service owner (Github account) can grant consent to the client application. The Github authorization server requests consent of the service owner and issues access tokens to clients. The issued token allows the client to access the APIs protected by OAuth2.0. These tokens are stored in a cookie in the local storage of the web browser. The access token is retrieved from the web browser's local storage instead of reauthenticating the user whenever they make a subsequent request. Besides, the business network cards used to connect to the business network are preserved by the REST server itself using the local storage. These business network cards are identities issued by using the Fabric-CA server to register enrollment certificates. The REST server is served in an isolated Docker image, and multiple instances of the REST server are allowed to configure a highly available instance of the persistent data store. It is worth noting that only the network administrator is authenticated to stop, restart, or remove the REST server instance without the application users losing access to the deployed business network over REST.

8. Conclusion and Future Work

Blockchain technology has many advantages in security, data protection, and the ability to bridge the disparate system manufacturers, CROs, and study sites. This technology has the benefit of centralization without having all of the data located in one place, making it less vulnerable to external or internal attacks. Few studies focus on a specific implementation approach that guarantees integrity and reliability by using blockchain technology in clinical trials, and there is a lack of practical use cases using blockchain in clinical trials. This paper proposes a clinical trial service platform on blockchain to ensure trial data integrity and secure trial-related services. To demonstrate the proposed approach's usability, a proof of concept is implemented on a permissioned-based network, namely, Hyperledger Fabric. A web-based application is also developed to ease the interaction with the blockchain network. The results demonstrate that smart contracts running on the Hyperledger Fabric network can be used to improve the transparency of data management in clinical trials. The proposed system can improve data integrity and accountability and transparency for the data exchange process and lower transaction costs. Besides, it could contribute to the core technologies in the safe management of trial data and the development of information security services. Furthermore, this study can be extended to a control system to maintain and manage reliable data in any other medical institution. It can be used as a critical technology to develop secure service and data management systems in smart healthcare.

Clinical research generates enormous amounts of trial data every day, which increases the pressures of regulatory agencies to overcome such significant data barriers. In particular, clinical trial management systems require high transaction throughput as well as low processing latency. Our future work will refine the prototype by adopting the blockchain implementation in the production environment.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government and Development of ICT Convergence Technology for Daegu-Gyeongbuk Regional Industry, and this research was supported by Energy Cloud R&D Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (2019M3F2A1073387).

References

[1] F. Farrokhyar, P. J. Karanicolas, A. Thoma et al., "Randomized controlled trials of surgical interventions," *Annals of Surgery*, vol. 251, no. 3, pp. 409–416, 2010.

[2] A. Bhatt, "Quality of clinical trials: a moving target," *Perspectives in Clinical Research*, vol. 2, no. 4, pp. 124–128, 2011.

[3] W. Gardner, C. W. Lidz, and K. C. Hartwig, "Authors' reports about research integrity problems in clinical trials," *Contemporary Clinical Trials*, vol. 26, no. 2, pp. 244–251, 2005.

[4] H. Rang, "Bad pharma: how drug companies mislead doctors and harm patients," *British Journal of Clinical Pharmacology*, vol. 75, no. 5, pp. 1377–1379, 2013.

[5] C. Weng, Y. Li, S. Berhe et al., "An Integrated Model for Patient Care and Clinical Trials (IMPACT) to support clinical research visit scheduling workflow for future learning health systems," *Journal of Biomedical Informatics*, vol. 46, no. 4, pp. 642–652, 2013.

[6] H. M. Colhoun, P. M. McKeigue, and G. D. Smith, "Problems of reporting genetic associations with complex outcomes," *Lancet*, vol. 361, no. 9360, pp. 865–872, 2003.

[7] E. C. Araujo de Carvalho, A. P. Batilana, W. Claudino et al., "Workflow in clinical trial sites & its association with near miss events for data quality: ethnographic, workflow & systems simulation," *PLoS One*, vol. 7, no. 6, article e39671, 2012.

[8] S. S. Ellenberg, "Protecting clinical trial participants and protecting data integrity: are we meeting the challenges?," *PLoS Med*, vol. 9, no. 6, article e1001234, 2012.

[9] "Blockchain for open science and knowledge creation, Bartling, Sönke, & et contributors to living document," 2017.

[10] Z. Zheng, S. Xie, H. Dai, X. Chen, and H. Wang, "An overview of blockchain technology: architecture, consensus, and future trends," in *Proceedings of the 2017 IEEE International Congress on Big Data (BigData Congress)*, Honolulu, HI, USA, June 2017.

[11] V. L. Lemieux, "Trusting records: is blockchain technology the answer?," *Records Management Journal*, vol. 26, no. 2, pp. 110–139, 2016.

[12] S. Underwood, *Blockchain beyond bitcoin*, pp. 15–17, 2016.

[13] A. Jeppsson and O. Olsson, *Blockchains as a Solution for Traceability and Transparency*, 2017.

[14] J. Golosova and A. Romanovs, "The advantages and disadvantages of the blockchain technology," in *2018 IEEE 6th Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE)*, Vilnius, Lithuania, November 2018.

[15] N. Szabo, "The idea of smart contracts," December 2019, http://fon.hum.uva.nl/rob/Courses/InformationInSpeech/.

[16] S. Nakamoto, *Bitcoin: A Peer-to-Peer Electronic Cash System*, Manubot, 2019.

[17] D. Macrinici, C. Cartofeanu, and S. Gao, "Smart contract applications within blockchain technology: a systematic mapping study," *Telematics and Informatics*, vol. 35, no. 8, pp. 2337–2354, 2018.

Hindawi
BioMed Research International
Volume 2024, Article ID 9848271, 1 page
https://doi.org/10.1155/2024/9848271

Retraction

Retracted: Identification and Validation of Potential Biomarkers and Pathways for Idiopathic Pulmonary Fibrosis by Comprehensive Bioinformatics Analysis

BioMed Research International

Received 12 March 2024; Accepted 12 March 2024; Published 20 March 2024

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope

(2) Discrepancies in the description of the research reported

(3) Discrepancies between the availability of data and the research described

(4) Inappropriate citations

(5) Incoherent, meaningless and/or irrelevant content included in the article

(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] W. Qian, X. Cai, Q. Qian, and X. Zhang, "Identification and Validation of Potential Biomarkers and Pathways for Idiopathic Pulmonary Fibrosis by Comprehensive Bioinformatics Analysis," *BioMed Research International*, vol. 2021, Article ID 5545312, 15 pages, 2021.

Hindawi
BioMed Research International
Volume 2021, Article ID 5545312, 15 pages
https://doi.org/10.1155/2021/5545312

Research Article

Identification and Validation of Potential Biomarkers and Pathways for Idiopathic Pulmonary Fibrosis by Comprehensive Bioinformatics Analysis

Weibin Qian,[1] Xinrui Cai,[2] Qiuhai Qian,[3] and Xinying Zhang[3]

[1]Department of Lung Disease, Affiliated Hospital of Shandong University of Traditional Chinese Medicine, Jinan, 250011 Shandong, China
[2]Department of Traditional Chinese Medicine, Shandong Academy of Occupational Health and Occupational Medicine, Shandong First Medical University & Shandong Academy of Medical Sciences, Jinan, 250062 Shandong, China
[3]Department of Endocrinology, Affiliated Hospital of Shandong University of Traditional Chinese Medicine, Jinan, 250011 Shandong, China

Correspondence should be addressed to Weibin Qian; doctorqwb1@126.com

Received 3 February 2021; Revised 27 April 2021; Accepted 10 June 2021; Published 5 July 2021

Academic Editor: Parameshachari B. D.

Objective. Idiopathic pulmonary fibrosis (IPF) is a chronic, progressive, irreversible, high-mortality lung disease, but its pathogenesis is still unclear. Our purpose was to explore potential genes and molecular mechanisms underlying IPF. *Methods*. IPF-related data were obtained from the GSE99621 dataset. Differentially expressed genes (DEGs) were identified between IPF and controls. Their biological functions were analyzed. The relationships between DEGs and microRNAs (miRNAs) were predicted. DEGs and pathways were validated in a microarray dataset. A protein-protein interaction (PPI) network was constructed based on these common DEGs. Western blot was used to validate hub genes in IPF cell models by western blot. *Results*. DEGs were identified for IPF than controls in the RNA-seq dataset. Functional enrichment analysis showed that these DEGs were mainly enriched in immune and inflammatory response, chemokine-mediated signaling pathway, cell adhesion, and other biological processes. In the miRNA-target network based on RNA-seq dataset, we found several miRNA targets among all DEGs, like RAB11FIP1, TGFBR3, and SPP1. We identified 304 upregulated genes and 282 downregulated genes in IPF compared to controls both in the microarray and RNA-seq datasets. These common DEGs were mainly involved in cell adhesion, extracellular matrix organization, oxidation-reduction process, and lung vasculature development. In the PPI network, 3 upregulated and 4 downregulated genes could be considered hub genes, which were confirmed in the IPF cell models. *Conclusion*. Our study identified several IPF-related DEGs that could become potential biomarkers for IPF. Large-scale multicentric studies are eagerly needed to confirm the utility of these biomarkers.

1. Introduction

Idiopathic pulmonary fibrosis (IPF) is a progressive, chronic, irreversible lung disease, characterized by an irreversible decline in lung function, progressive pulmonary scarring, and common interstitial pneumonia [1–3]. It affects more than 3 million people worldwide [4–7]. However, the prognosis of IPF remains poor, and the median survival time of patients is only 2-4 years [8, 9]. Thus, it emphasizes a need for a more complete understanding of the pathogenesis of IPF.

Long-term clinical work has shown that clarifying the pathogenesis of IPF helps to diagnose early disease, which is of great significance for the treatment of this disease, and has long-term clinical results to improve this fatal disease [9, 10]. However, it is still challenging to diagnose IPF in

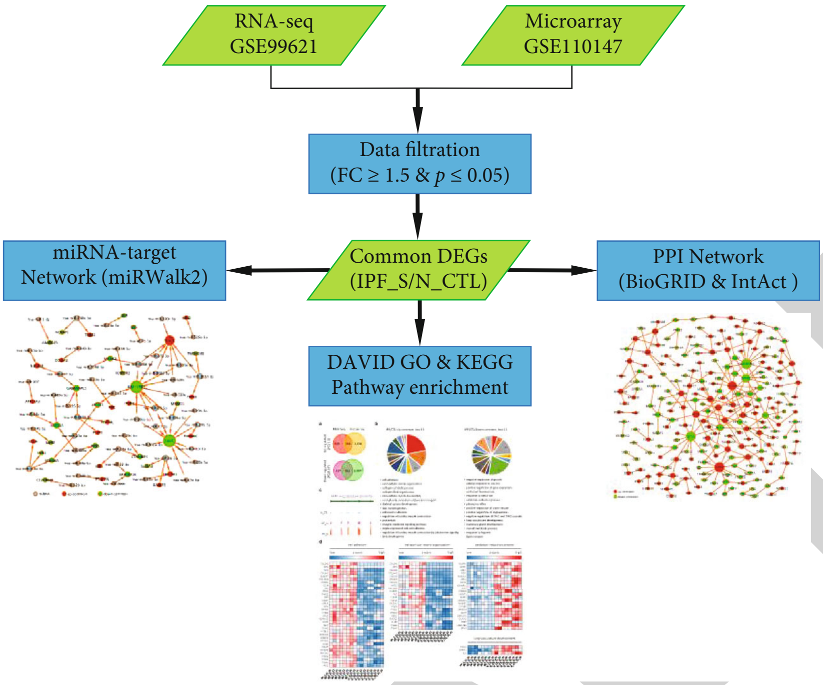


Figure 1: The flowchart of this study.

clinical work [11]. The clinical manifestations of IPF lack specificity, and the diagnosis needs to be combined with the detailed medical history of patients with multiple similar diseases [12, 13]. As the understanding of the disease deepens, biomarkers play an increasingly important role in the research and development of diseases [14, 15]. However, it is still difficult to reliably predict the course of IPF and the response to therapy for an individual patient [11]. There is a long way until biomarkers complete or substitute for the clinical and functional parameters currently available for IPF [16]. Only a very small number of DEGs have been found, and they are not consistent across all these studies [17]. Thus, further development into available markers and therapeutic targets is limited due to these inconsistent results [18]. Small sample sizes, different platforms, and different statistical methods are limiting factors that lead to inconsistent results [19]. To resolve this limitation, in this study, we comprehensively analyzed RNA-seq and microarray expression profiles of IPF from different platforms and validated hub genes in IPF cell models, which could lay the foundation for clinical research and IPF treatment.

miRNAs, a class of noncoding small RNAs, are involved in RNA silencing, posttranscriptional regulation, and other biological processes [20]. It has been confirmed that miRNAs play a critical role in the occurrence and development of various diseases, including IPF [21–23]. As an example, miR-92, miR-210, and miR-let-7d have been confirmed to be associated with IPF [24–26]. Therefore, in this study, differentially expressed genes (DEGs) in IPF were identified and miRNA-mediated regulatory network among all DEGs was constructed, which might shed novel light on molecular mechanisms of IPF progression.

2. Materials and Methods

2.1. Data Acquisition and Processing. The Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo) is a public functional genomics data repository. Based on the filter criteria of keywords: IPF, organism: homo sapiens, and experiment type: expression profiling by high-throughput sequencing or expression profiling by array, two datasets GSE99621 and GSE110147 were included for this study. RNA-seq data of IPF were obtained from the GSE99621 dataset. This dataset contained 8 affected areas of the lung (IPF_S), 10 unaffected areas of the lung (IPF_N), and 8 healthy controls (N_CTL) on the GPL16791 platform [27]. The adapters were removed with Trimmomatic-0.38 and the localized Perl script was used to remove $5'$ and $3'$ low-quality bases ($Q < 20$), retaining sequences with $Q > 20$ bases above 90% and total length > 35 bp. The clean reads were mapped to the protein-coding gene sequence of Homo sapiens (assembly GRCh38.p12) using the HISAT2. Then, bedtools was utilized to calculate the number of reads and the RPKM expression value by a localized Perl script. After that, the GeneCluster3.0 was used for the systematic hierarchical clustering of samples. The principal component analysis (PCA) was then conducted. Microarray data of IPF were also retrieved from the GSE110147 dataset on the GPL6244 platform, including 22 IPF tissues and 11 normal lung tissues [28]. The flowchart of this study is shown in Figure 1.

2.2. Differentially Expressed Analysis. Differential expression analysis between the IPF_S, IPF_N, and N_CTL samples was performed using the limma package in R (http://bioconductor.org/packages/release/bioc/html/limma.html) [29]. Genes with

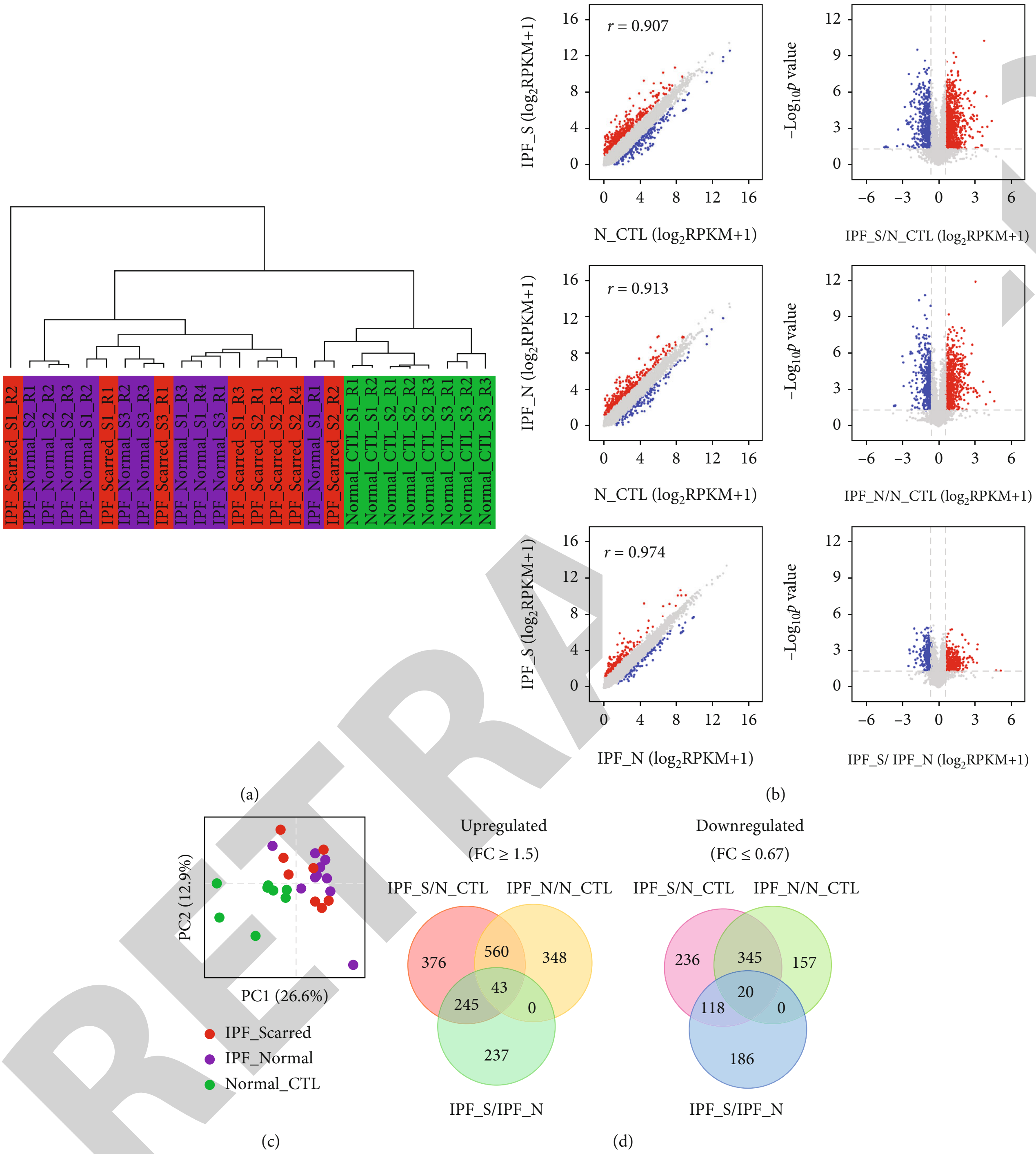


Figure 2: Continued.

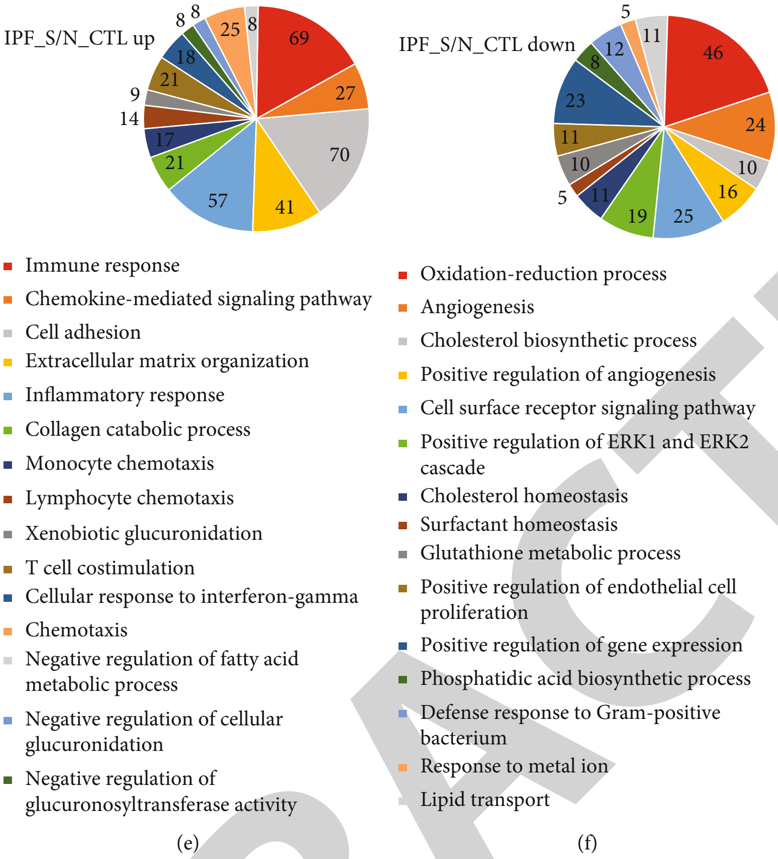


Figure 2: Identification of DEGs and relevant signaling pathways for IPF using the GSE99621 RNA-seq dataset. (a) Hierarchical clustering analysis of 8 affected areas of the lung (IPF_S), 10 unaffected areas of the lung (IPF_N), and 8 healthy controls (N_CTL). (b) Volcano map and scatter plot showing the DEGs between IPF_S and N_CTL and between IPF_N and N_CTL. Red represents upregulated genes, and blue represents downregulated genes. (c) PCA analysis of 8 IPF_S, 10 IPF_N, and 8 N_CTL. (d) Venn diagram showing up- or downregulated genes between IPF_S and N_CTL and between IPF_N and N_CTL. (e) Functional enrichment analysis of upregulated genes. (f) Functional enrichment analysis of downregulated genes.

adjusted p value ≤ 0.05 and fold change ≥ 1.5 were screened as upregulated genes. Meanwhile, those with adjusted p value ≤ 0.05 and fold change ≤ 1.5 were identified as downregulated genes.

2.3. Functional Enrichment Analysis. Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses of DEGs were carried out by a functional annotation analysis tool online, Database for Annotation, Visualization and Integrated Discovery (DAVID; http://david.abcc.ncifcrf.gov/) [30]. In the biological processes, the gene expression heat maps most relevant to IPF were depicted. Terms with false discovery rate (FDR) < 0.05 were significantly enriched.

2.4. The miRNA-Target Network. The miRWalk 2.0 database was used to predict the relationships between DEGs and miRNAs. The miRWalk (http://mirwalk.uni-hd.de/) is a publicly available comprehensive resource that hosts predictive and experimentally validated miRNA-target interaction pairs. This database allows for possible miRNA binding site predictions within the complete sequence of all known genes in the three genomes (human, mouse and rat), including ten different prediction datasets [31]. We took the interactions between the three prediction sets of the TargetScan, miRDB and miRTarBase databases. Furthermore, a miRNA-target network was visualized using Cytoscape. Cytoscape is an open software platform for visualization and data integration of molecular interaction networks [32].

2.5. Common DEGs Both in RNA-seq and Microarray Datasets. Common DEGs were intersected between RNA-seq and microarray datasets. Furthermore, biological processes and pathways of these common DEGs were analyzed by GO and KEGG enrichment analyses. The peak map of DEGs was then built up.

2.6. Construction of a PPI Network. It is helpful to clarify the key mechanisms of disease development and reveal key cellular functions and biological processes by studying the interactions between transcripts or proteins. The data of BioGRID (http://www.thebiogrid.org) [33] and IntAct (http://www.ebi.ac.uk/intact) [34], two protein interaction databases, were integrated to find the interactions between common DEGs. Finally, a PPI network was visualized with Cytoscape software.

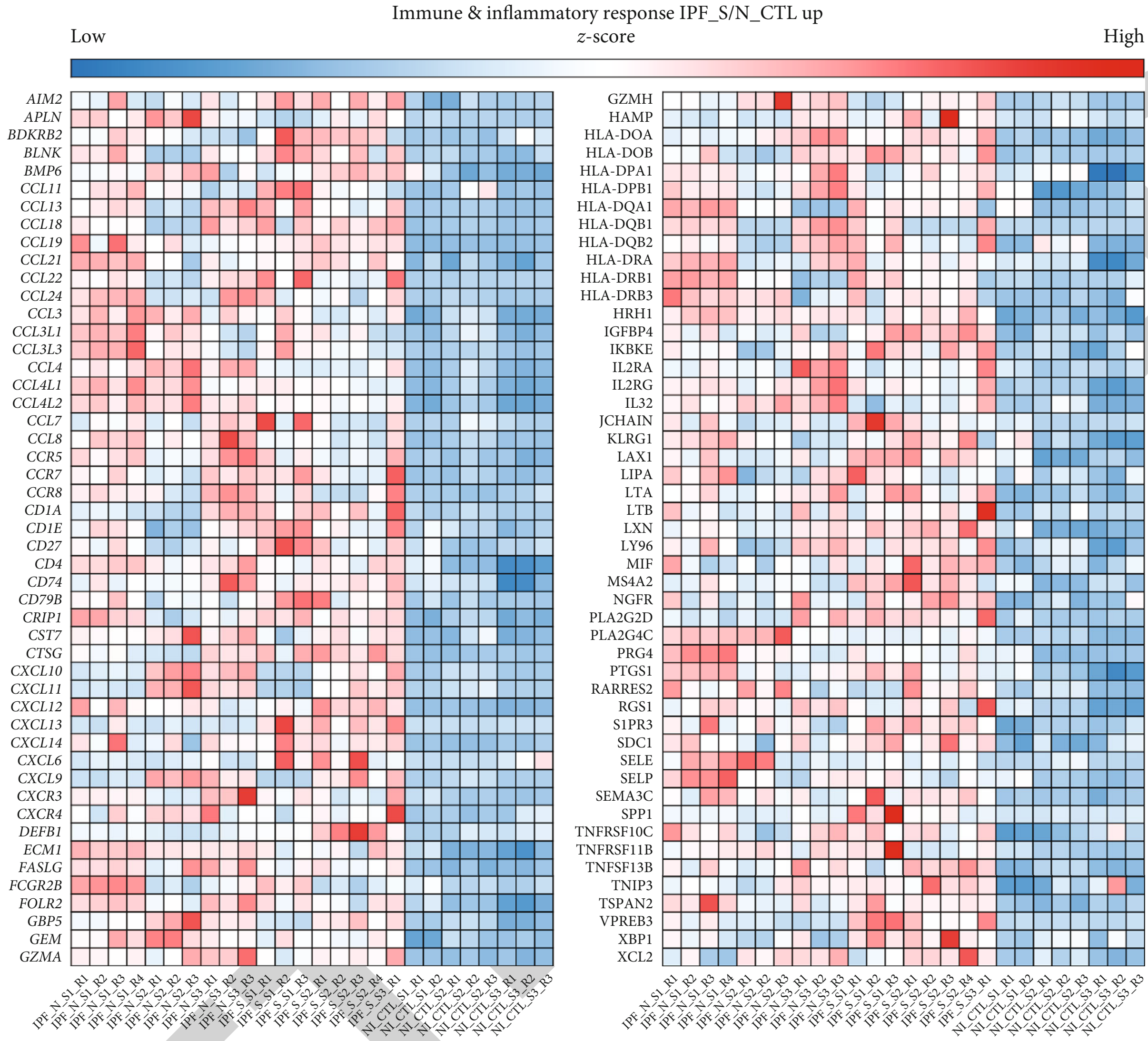


FIGURE 3: The difference in expression pattern of immune and inflammatory response-related DEGs in IPF_N, IPF_S, and N_CTL.

2.7. Cell Culture. Human normal lung and bronchus epithelial cell line BEAS-2B (CRL-9609, ATCC, USA) were grown in BEGM kit from Lonza (CC-3170; Basel, Switzerland) at 37°C in 5% CO_2. BEAS-2B cells were treated with 0.5 ng/ml TGF-β1 for 48 h to construct an IPF cell model. The IPF model was validated by examining α-SMA, fibronectin, and Col I expression levels.

2.8. Western Blot. TGF-β1-induced BEAS-2B cells and controls were lysed by RIPA lysates (Beyotime, Shanghai, China). After the lysates were centrifuged at 12,000 × g for 20 min, the supernatant was harvested. The extracted protein was resolved by 10% SDS-PAGE and transferred onto PVDF membranes. Following blocking, the membranes were incubated with α-SMA (1 : 1000; ab108424, Abcam, USA), Fibronectin (1 : 1000; ab32419, Abcam), Col I (1 : 1000; ab255809, Abcam), GABARAPL1 (1 : 1000; ab229729, Abcam), GPX8 (1 : 1000; ab183664, Abcam), SGTA (1 : 1000; ab96571, Abcam), VCAM1 (1 : 1000; ab174279, Abcam), ARRB1 (1 : 1000; ab32099, Abcam), and GAPDH (1 : 1000; ab8245, Abcam), followed by secondary antibodies. The optical density of the bands was quantified using ImageJ software.

2.9. Statistical Analysis. All statistical analyses were carried out by R packages and GraphPad prism software. Each experiment was independently repeated at least three times. Data were presented as the mean ± standard deviation. Comparisons between two groups were presented by Student's t -test. p value < 0.05 indicated statistical significance.

3. Results

3.1. Upregulated and Downregulated Genes in IPF and Their Biological Functions. RNA-seq data of 8 IPF_S, 10 IPF_N, and 8 N_CTL samples were obtained from the GSE99621 dataset. Our hierarchical clustering analysis results showed

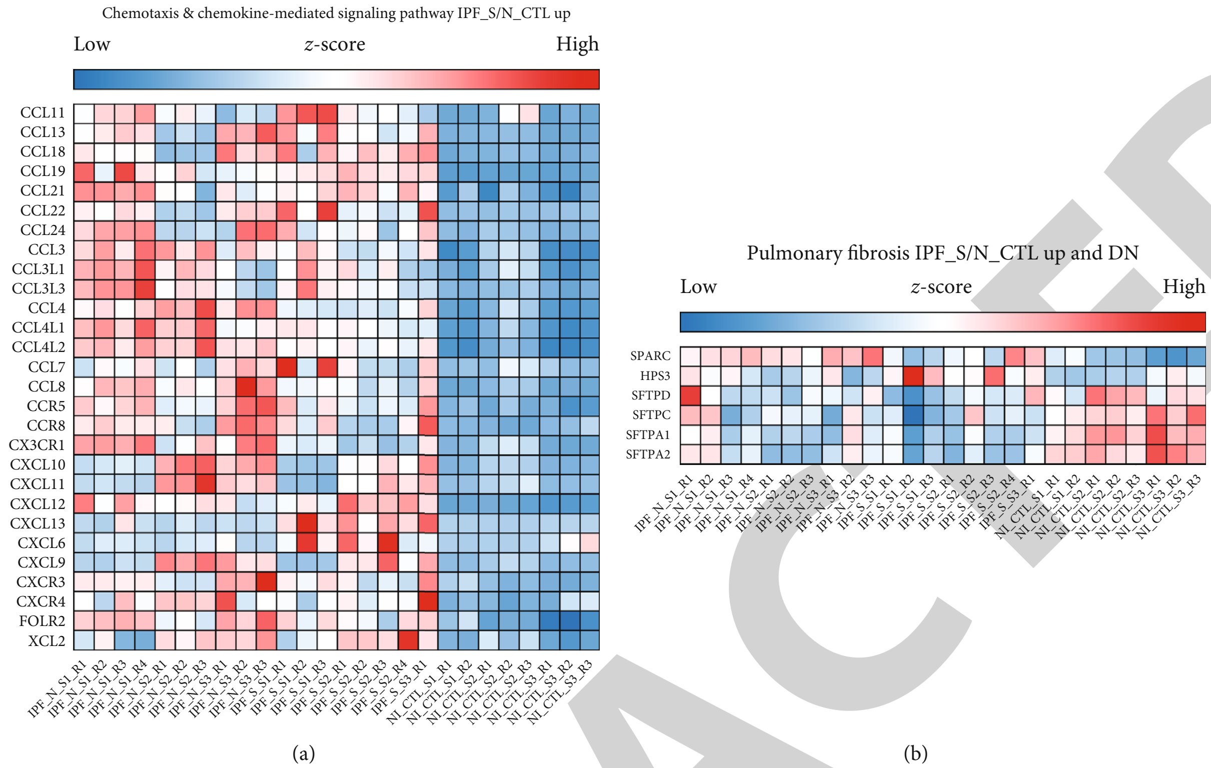


Figure 4: The difference in expression pattern of chemotaxis and chemokine-mediated signaling pathway- or pulmonary fibrosis-related DEGs in IPF_N, IPF_S, and N_CTL. (a) Chemotaxis and chemokine-mediated signaling pathway. (b) Pulmonary fibrosis.

that N_CTL samples were distinctly distinguished from IPF_S and IPF_N samples (Figure 2(a)). Several IPF_S and IPF_N samples were grouped into one category. In Figure 2(b), at the gene expression levels, there were significant correlations between the IPF_S, IPF_N, and N_CTL samples. Furthermore, PCA confirmed that N_CTL samples were different from the IPF_S and IPF_N samples (Figure 2(c)). With the cutoff of adjusted $p < 0.05$ and fold change ≥ 1.5, upregulated genes were screened between the IPF_S, IPF_N, and N_CTL samples (Figure 2(d)). We also identified downregulated genes with adjusted $p < 0.05$ and fold change ≤ 1.5 between the IPF_S, IPF_N, and N_CTL samples. Totally, 1224 genes were upregulated in the IPF_S than N_CTL samples, while 719 genes were downregulated in IPF_S compared to N_CTL samples. We explored potential biological functions of abnormally expressed genes between the IPF_S and N_CTL samples. Our data suggested that these upregulated genes were distinctly enriched in immune- or inflammatory-related pathways such as immune response, chemokine-mediated signaling pathway, inflammatory response, cell adhesion, extracellular matrix organization, monocyte chemotaxis, and lymphocyte chemotaxis (Figure 2(e)). In addition, these downregulated genes were mainly involved in IPF-related pathways such as oxidation-reduction process, angiogenesis, and cholesterol biosynthetic process (Figure 2(f)).

3.2. IPF-Related Upregulated Genes in Immune and Inflammatory Responses. Each stage of IPF is accompanied by innate or adaptive immune response [35]. Herein, we found that upregulated genes were mainly enriched in immune and inflammatory pathways. We further focused on which genes were involved in these pathways. As a result, chemokine (C-C motifs such as CCL-11, 13, 18, 21, 22, 24, 3, 3L1, 3L3, 4, 4L1, 4L2, 7, and 8 and C-X-C motifs such as CXCL10, 11, 12, 13, 14, 6, and 9) ligand family members and human leucocyte antigen (such as HLA-DOA, DOB, DPA1, DPB1, DQA1, DQB1, DQB2, DRA, DRB1, and DRB3) genes were significantly enriched in immune and inflammatory responses (Figure 3).

3.3. IPF-Related Upregulated Genes in Chemotaxis and Chemokine-Mediated Signaling Pathway. There is evidence that severity of IPF relies on chemotaxis [36]. Figure 4(a) showed all the upregulated genes enriched in chemotaxis and chemokine-mediated signaling pathway. Chemokine (C-C motifs such as CCL-11, 13, 18, 19, 21, 22, 24, 3, 3L1, 3L3, 4, 4L1, 4L2, 7, and 8 and C-X-C motifs such as CXCL10, 11, 12, 13, 6, and 9) ligand family members were distinctly enriched in chemotaxis and chemokine-mediated signaling pathway.

3.4. IPF-Related Genes in Pulmonary Fibrosis Pathway. We focused on the up- and downregulated genes enriched by pulmonary fibrosis pathway. As shown in Figure 4(b), SPARC, HPS3, SFTPD, SFTPC, SFTPA1, and SFTPA2 were involved in the pulmonary fibrosis pathway, indicating that

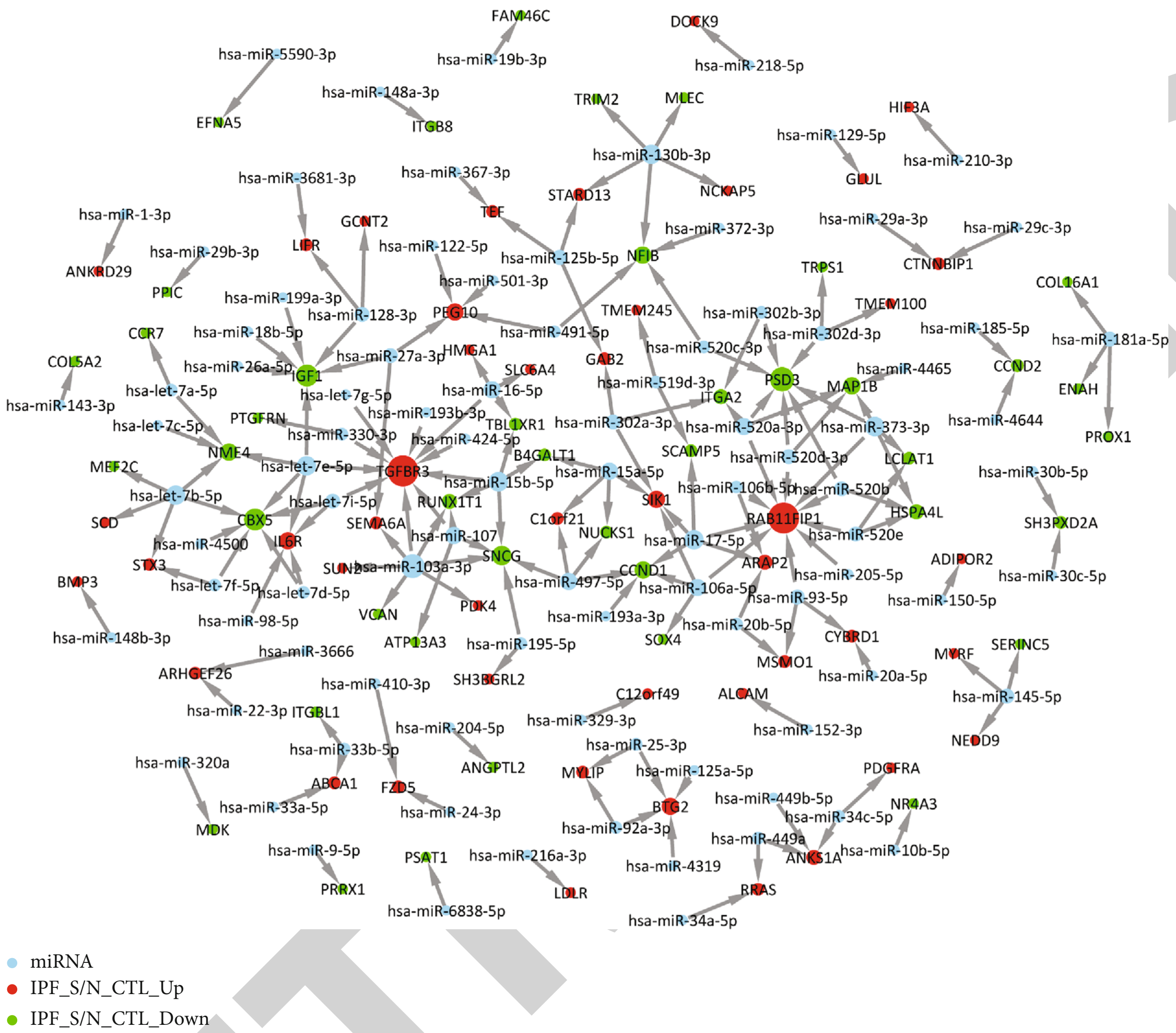


FIGURE 5: miRNA-target network for IPF. The greater the degree, the larger the node. Red represents upregulated genes, green represents downregulated genes, and blue indicates miRNAs.

abnormal expression of these genes could participate in the pulmonary fibrosis pathway.

3.5. Prediction of Up- and Downregulated Genes Regulated by miRNAs. Based on the miRWalk 2.0 database, we predicted the relationships between IPF-related DEGs and upstream miRNAs. The miRNA-target network for IPF was constructed (Figure 5). Our data showed that a DEG was often regulated by multiple miRNAs. For example, TGFBR3 was a target of hsa-let-7e-5p, hsa-let-7g-5p, hsa-let-7i-5p, hsa-miR-103a-3p, hsa-miR-107, hsa-miR-15b-5p, and hsa-miR-16-5p. RAB11FIP1 was regulated by hsa-miR-106a-5p, hsa-miR-106b-5p, hsa-miR-17-5p, hsa-miR-205-5p, hsa-miR-373-3p, hsa-miR-520a-3p, hsa-miR-520b, hsa-miR-520d-3p, hsa-miR-520e, and hsa-miR-93-5p.

3.6. Validation of DEGs and Their Biological Functions in IPF. To further validate DEGs in IPF, we comprehensively analyzed DEGs of IPF both in the microarray and RNA-seq datasets. As shown in Figure 6(a), 304 genes were upregulated in IPF_S compared to N_CTL both in the microarray and RNA-seq datasets. Furthermore, we found 282 downregulated genes in IPF_S compared to N_CTL both in the microarray and RNA-seq datasets. We further validated the biological functions enriched by these DEGs. In Figure 6(b), we listed the most significantly biological processes or pathways enriched by up- or downregulated genes, respectively. Our data confirmed that cell adhesion, extracellular matrix organization, collagen catabolic organization, collagen fibril organization, and extracellular matrix disassembly were significantly enriched by these upregulated genes. Moreover, we found that downregulated genes were most enriched in oxidation-reduction process and lung vasculature development. We gave an example of MMP7 expression in the IPF_S, IPF_N, and N_CTL samples (Figure 6(c)). We found the differences in the expression pattern of DEGs between the

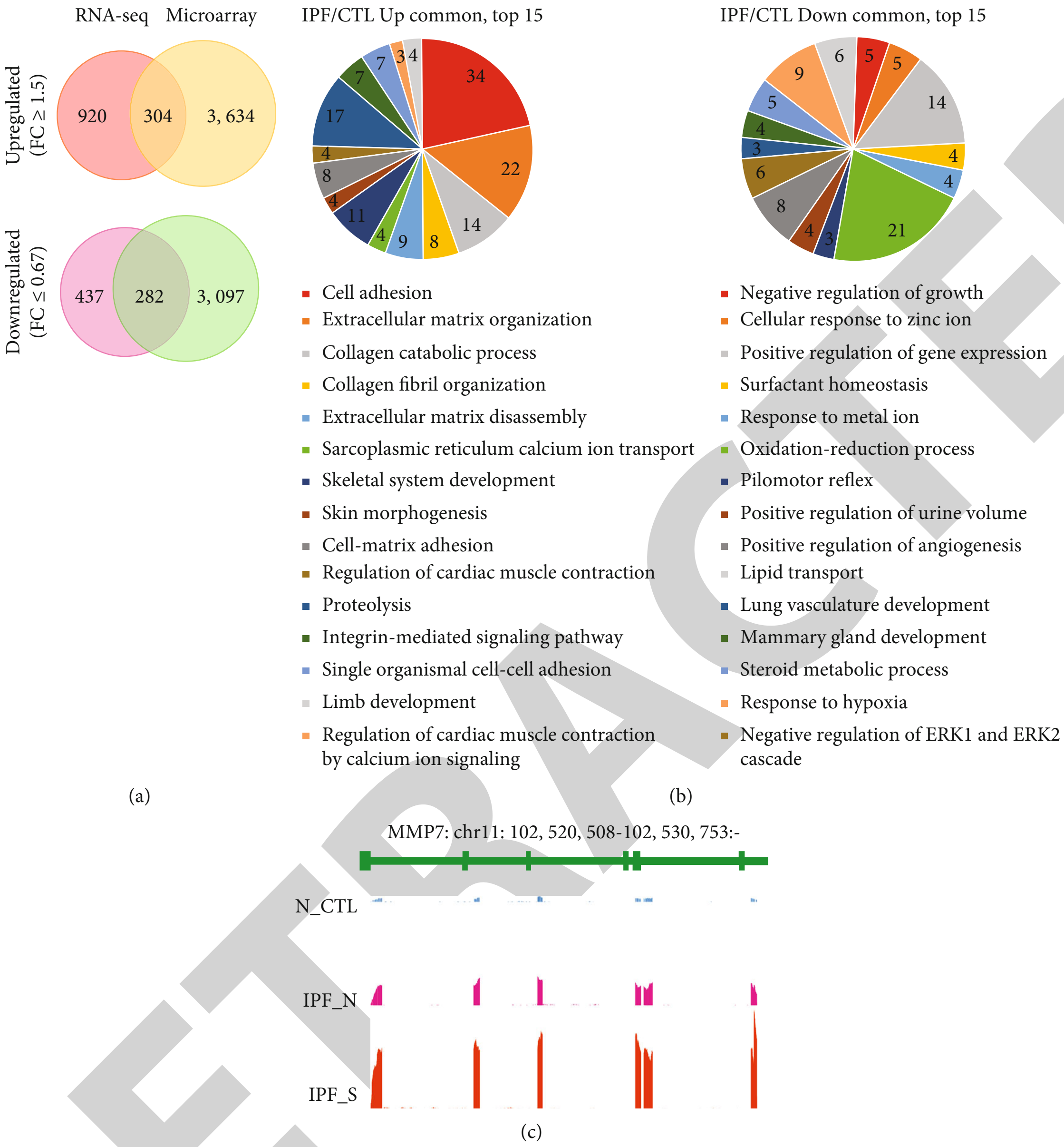
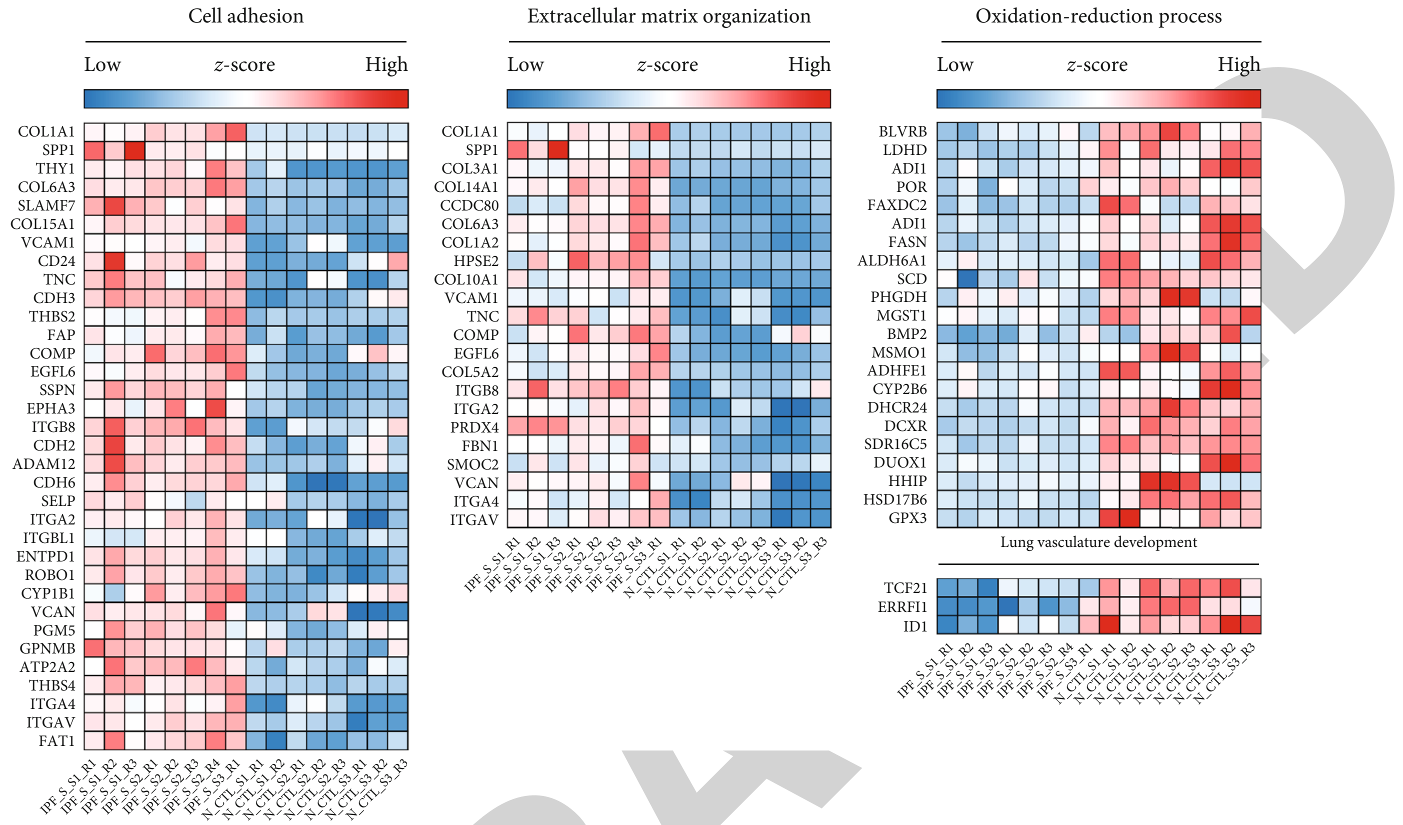


Figure 6: Continued.



(d)

Figure 6: Identification of common DEGs and relevant biological processes or pathways for IPF both in the microarray and RNA-seq datasets. (a) Venn diagram showing the common DEGs for IPF_S both in the microarray and RNA-seq datasets. (b) Functional enrichment analysis of common up- or downregulated genes in IPF_S compared to N_CTL. (c) Gene peak map of MMP7. (d) The difference in expression pattern of cell adhesion-, extracellular matrix organization-, oxidation-reduction process-, and lung vasculature development-related DEGs in IPF_S and N_CTL.

IPF_S and N_CTL samples in cell adhesion, extracellular matrix organization, oxidation-reduction process, and lung vasculature development (Figure 6(d)).

3.7. Construction of a PPI Network Based on Common DEGs in IPF. The PPI network was constructed to investigate the interactions between common DEGs both in the microarray and RNA-seq datasets (Figure 7). In the network, there were 227 nodes, including 122 upregulated and 105 downregulated genes. Nodes with degree ≥ 7 were considered hub genes, including GABARAPL1, GPX8, SGTA, VCAM1, ARRB1, SPP1, and HLA-B (Table 1).

3.8. miRNA-Target Network Based on Common DEGs in IPF. We further predicted the miRNAs of common DEGs both in the microarray and RNA-seq datasets. A miRNA-target network was established (Figure 8). We found that LDLR and RAB11FIP1 were regulated by most miRNAs. Both were upregulated in IPF_S compared to N_CTL.

3.9. Validation of Hub Genes in IPF Cell Models. TGF-β1-induced BEAS-2B cells were used to construct IPF cell models (Figure 9(a)). After verification, α-SMA, fibronectin, and Col I proteins were all highly expressed in TGF-β1-induced BEAS-2B cells than controls, suggesting that these IPF cell models were successfully constructed ($p < 0.0001$; Figures 9(b) and 9(c)). The hub genes were validated in TGF-β1-induced BEAS-2B cells by western blot. Our data confirmed that GABARAPL1, SGTA, and ARRB1 exhibited lower expression levels in IPF cells compared to controls ($p < 0.0001$; Figures 9(d) and 9(e)). GPX8 and VCAM1 were both downregulated in IPF cells than controls.

4. Discussion

In this study, we identified IPF-related DEGs (such as GABARAPL1, SGTA, ARRB1, GPX8, and VCAM1) and analyzed potential pathways (such as immune and inflammatory pathways) by comprehensively analyzing IPF-related RNA-seq and microarray datasets. Combining the PPI network, miRNA-target network, and functional enrichment analysis, we screened out potential biomarkers and their related regulatory mechanisms in IPF. These biomarkers might provide novel ideas and clues for further experimental research.

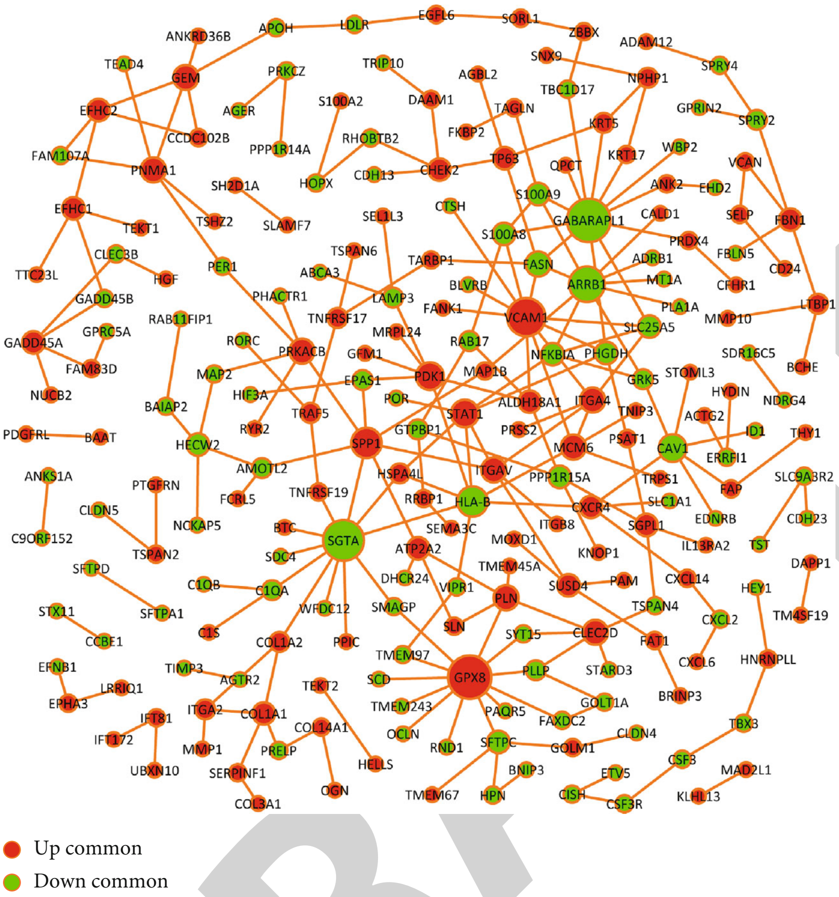


Figure 7: The construction of protein-protein interaction network based on common DEGs both in the microarray and RNA-seq datasets. Red represents upregulated genes, and green represents downregulated genes.

Table 1: Hub genes in the PPI network.

Node	Degree	Up-/downregulation
GABARAPL1	12	Down
GPX8	12	Up
SGTA	11	Down
VCAM1	10	Up
ARRB1	9	Down
SPP1	7	Up
HLA-B	7	Down

In this study, we analyzed DEGs between IPF_S and N_CTL. Functional enrichment analysis showed that upregulated genes in IPF_S were mainly enriched in immune response, chemokine-mediated signaling pathway, and cell adhesion and the like. Numerous molecules involved in immune response have been proposed as potential biomarkers for IPF [37]. For example, pyroptosis, a proinflammatory form of programmed cell death, mainly mediates the activation of caspase-1 through inflammasomes. A recent study has found that immunosuppressive molecule PD-L1 may trigger pyroptosis of pulmonary arterial smooth muscle cells, thereby accelerating pulmonary vascular fibrosis [38]. In addition, downregulated genes were mainly involved in oxidation-reduction process, angiogenesis, and cholesterol biosynthetic process and so on. It has been confirmed that reducing protein oxidation could reverse lung fibrosis [39]. Among all biological processes and pathways, we found that chemokine (C-C motif and C-X-C motif) ligand family genes were obviously associated with these significant biological processes, indicating that chemokine ligand family genes could play a key role in the processes of IPF, which was consistent with a previous study [40]. Also, a prospective case control study has found that chemokine ligand family member CCL18 is associated with IPF [41].

It has been well recognized that small sample sizes, different platforms, and different statistical methods could lead to inconsistent results [17, 42]. Therefore, in this study, we comprehensively analyzed DEGs between IPF_S and N_CTL both in the RNA-seq and microarray datasets. Our results showed that there are 304 upregulated genes and 282 downregulated genes in IPF_S compared to N_CTL both in the microarray and RNA-seq datasets. Functional enrichment analysis results revealed that these DEGs were mainly enriched in cell adhesion, extracellular matrix organization, oxidation-reduction process, and lung vasculature development. The PPI network showed that 3 upregulated including GPX8, VCAM1, and SPP1 and 4 downregulated genes

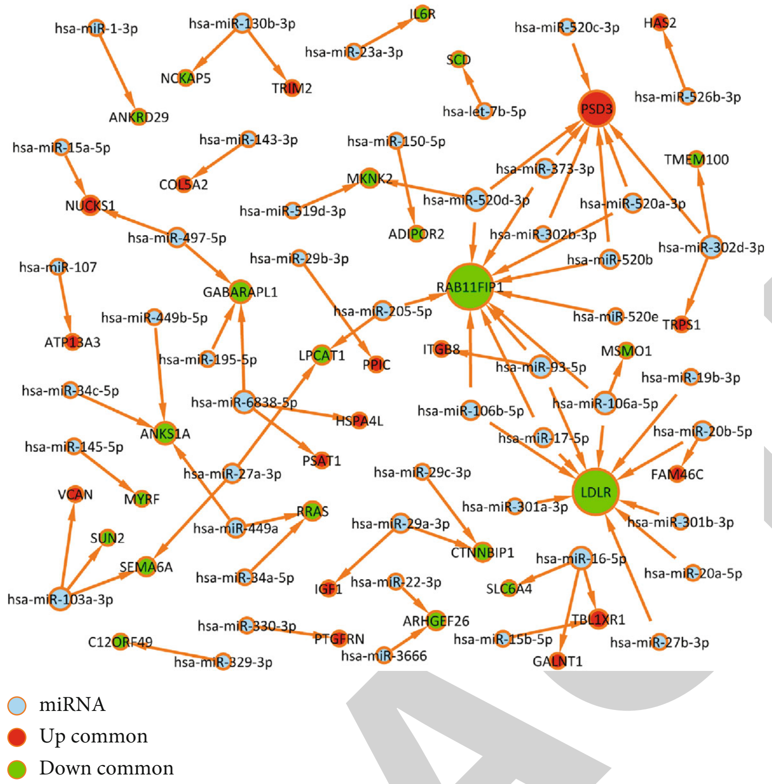


Figure 8: miRNA-target network based on common DEGs both in the microarray and RNA-seq datasets. The greater the degree, the larger the node. Red represents upregulated genes, green represents downregulated genes, and blue indicates miRNAs.

including GABARAPL1, SGTA, ARRB1, and HLA-B could be considered hub genes for IPF. Among them, VCAM1 and TGFBR3 have been reported to be involved in the development of IPF. It has been confirmed that VCAM1 may mediate the adhesion of lymphocytes, monocytes, eosinophils, and basophils to vascular endothelium. Furthermore, it plays a key role in leukocyte-endothelial cell signal transduction. It is currently widely believed that VCAM1 is involved in the pathogenesis of atherosclerosis and can serve as a potential therapeutic target [43, 44]. More importantly, Agassandian et al. and other studies have shown that VCAM-1 can induce TGF-β1 expression upregulation and increase fibroblast expression in IPF [45]. Furthermore, GABARAPL1 may be involved in mediating the proliferation of endothelial progenitor cells, migration angiogenesis, and autophagy [46]. Methylated SGTA has been implicated in synovial fibroblast proliferation in patients with rheumatoid arthritis [47]. Based on the above findings, these hub genes might have potential effects in the pathogenesis of IPF, which require further research.

It has been confirmed that miRNA-mRNA interactions play a critical role in the development of IPF [48–50]. Here, we predicted the potential miRNA targets of all DEGs using the miRWalk 2.0 database. We found that DEGs were potential targets of many miRNAs, especially RAB11FIP1 and TGFBR3, indicating that the altered expression of these DEGs could be induced by miRNAs at the posttranscriptional level. For example, Rab11FIP1, a member of the large Rab GTPase family, has a regulatory role in the formation, targeting, and fusion of intracellular transport vesicles [51, 52]. As described in previous studies, Rab11FIP1 may play a vital role in several cancers such as breast cancer and cervical cancer [51, 52]. Otherwise, Hwang et al. believed that Rab coupling protein could activate epithelial-to-mesenchymal transition [53]. Interestingly, it has been proved that epithelial-to-mesenchymal transition is a key step in the development of IPF [54, 55]. Combining previous studies, Rab11FIP1 has the potential to become a potential biomarker for IPF.

Collectively, this study provided several novel draggable-target molecules for IPF by bioinformatics. The reliability of results for biological investigations was verified in IPF cell models. The consistent results between bioinformatics and biological investigation suggested convincing evidence that hub genes including GABARAPL1, SGTA, ARRB1, GPX8, and VCAM1 were abnormally expressed in IPF and could be utilized as a promising novel target for IPF treatment. However, there are several limitations in our study. First, although we validated hub genes in IPF cell models by western blot, their functions in IPF should be clarified. Second, although several pathways were identified for IPF, molecular experiments should be presented to prove more reliable evidence for the phenotypes and pathways underlying IPF. In conclusion, our study identified several IPF-related DEGs that could become potential biomarkers for IPF. Large-scale multicentric studies are eagerly needed to confirm the utility of these biomarkers in our future studies.

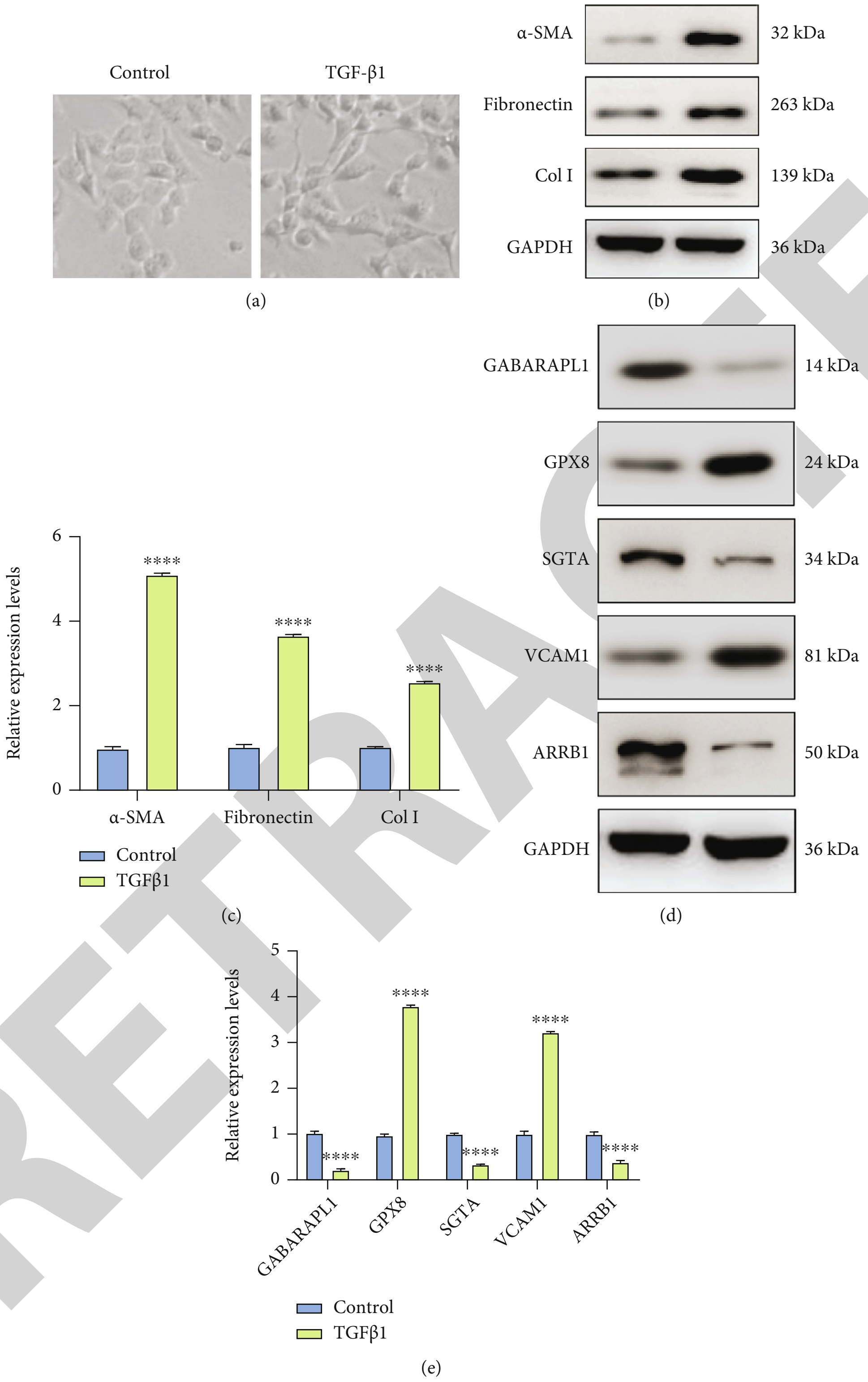


Figure 9: Validation of hub genes in IPF cell models by western blot. (a) Construction of TGF-β1-induced IPF cell models. (b, c) Assessment of the expression of α-SMA, fibronectin, and Col I proteins in IPF cells. (d, e) Validation of the expression of GABARAPL1, SGTA, ARRB1, GPX8, and VCAM1 in IPF cells. $^{****}p < 0.0001$.

5. Conclusion

In this study, we comprehensively analyzed IPF-related DEGs and potential signaling pathways using the RNA-seq and microarray datasets. By combining the PPI network, miRNA-target network, and functional enrichment analysis, we identified potential biomarkers including GABARAPL1, SGTA, ARRB1, GPX8, and VCAM1 for IPF. Among them, GABARAPL1, SGTA, and ARRB1 exhibited lower expression levels in IPF while GPX8 and VCAM1 were both downregulated in IPF. These biomarkers might provide novel insights for further experimental research.

Abbreviations

IPF:	Idiopathic pulmonary fibrosis
DEGs:	Differentially expressed genes
PPI network:	Protein interaction network
PCA:	Principal component analysis
DAVID:	Database for Annotation, Visualization and Integrated Discovery
GO:	Gene Ontology
KEGG:	Kyoto Encyclopedia of Genes and Genomes.

Data Availability

The datasets analyzed during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest

The authors declare no conflicts of interest.

Authors' Contributions

Weibin Qian and Xinrui Cai contributed equally to this work.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (81904143, 81704071); Taishan Scholars Youth Expert Program of Shandong Province in China (tsqn201812146); Young Elite Scientists Sponsorship Program by CAST (CACM-2018-QNRC2-B01, CACM-2019-QNRC2-C01), Outstanding young talents of Qilu health, China Postdoctoral Science Foundation Grant (2019T120608, 2019M652462); Postdoctoral Innovation Project of Shandong Province (201901014); Innovation Project of Shandong Academy of Medical Sciences; Academic Promotion Programme of Shandong First Medical University (2019QL001); Key Research and Development Plan of Shandong province (2018GSF119027); Inheritance Studio of Famous experts in traditional Chinese Medicine in Shandong Province of Qiuhai Qian, the Project of Scientific and Technological Development Program of Jinan (201805081, 201805009).

References

[1] T. Parimon, C. Yao, B. R. Stripp, P. W. Noble, and P. Chen, "Alveolar epithelial type II cells as drivers of lung fibrosis in idiopathic pulmonary fibrosis," *International Journal of Molecular Sciences*, vol. 21, no. 7, p. 2269, 2020.

[2] A. U. Wells, "Pamrevlumab in idiopathic pulmonary fibrosis," *The Lancet Respiratory Medicine*, vol. 8, no. 1, pp. 2-3, 2020.

[3] H. Wu, Y. Yu, H. Huang et al., "Progressive pulmonary fibrosis is caused by elevated mechanical tension on alveolar stem cells," *Cell*, vol. 180, no. 1, pp. 107–121.e117, 2020.

[4] H. Robbie, C. Daccord, F. Chua, and A. Devaraj, "Evaluating disease severity in idiopathic pulmonary fibrosis," *European respiratory review : an official journal of the European Respiratory Society*, vol. 26, no. 145, p. 170051, 2017.

[5] Y. M. Liu, K. Nepali, and J. P. Liou, "Idiopathic pulmonary fibrosis: current status, recent progress, and emerging targets," *Journal of Medicinal Chemistry*, vol. 60, no. 2, pp. 527–553, 2017.

[6] G. Sgalla, A. Biffi, and L. Richeldi, "Idiopathic pulmonary fibrosis: diagnosis, epidemiology and natural history," *Respirology*, vol. 21, no. 3, pp. 427–437, 2016.

[7] G. Sgalla, B. Iovene, M. Calvello, M. Ori, F. Varone, and L. Richeldi, "Idiopathic pulmonary fibrosis: pathogenesis and management," *Respiratory Research*, vol. 19, no. 1, p. 32, 2018.

[8] G. Raghu, B. Ley, K. K. Brown et al., "Risk factors for disease progression in idiopathic pulmonary fibrosis," *Thorax*, vol. 75, no. 1, pp. 78–80, 2020.

[9] L. Richeldi, E. R. Fernández Pérez, U. Costabel et al., "Pamrevlumab, an anti-connective tissue growth factor therapy, for idiopathic pulmonary fibrosis (PRAISE): a phase 2, randomised, double-blind, placebo- controlled trial," *The Lancet Respiratory Medicine*, vol. 8, no. 1, pp. 25–33, 2020.

[10] W. A. Wuyts, M. Wijsenbeek, B. Bondue et al., "Idiopathic pulmonary fibrosis: best practice in monitoring and managing a relentless fibrotic disease," *Respiration*, vol. 99, no. 1, pp. 73–82, 2020.

[11] T. Khan, S. Dasgupta, N. Ghosh, and K. Chaudhury, "Proteomics in idiopathic pulmonary fibrosis: the quest for biomarkers," *Mol Omics*, vol. 17, no. 1, pp. 43–58, 2021.

[12] D. J. Lederer and F. J. Martinez, "Idiopathic pulmonary fibrosis," *The New England Journal of Medicine*, vol. 378, no. 19, pp. 1811–1823, 2018.

[13] D. A. Lynch, N. Sverzellati, W. D. Travis et al., "Diagnostic criteria for idiopathic pulmonary fibrosis: a Fleischner Society White Paper," *Respiratory Medicine*, vol. 6, no. 2, pp. 138–153, 2018.

[14] M. R. Hadjicharalambous and M. A. Lindsay, "Idiopathic pulmonary fibrosis: pathogenesis and the emerging role of long non-coding RNAs," *International Journal of Molecular Sciences*, vol. 21, no. 2, p. 524, 2020.

[15] Y. Inoue, R. J. Kaner, J. Guiot et al., "Diagnostic and prognostic biomarkers for chronic fibrosing interstitial lung diseases with a progressive phenotype," *Chest*, vol. 158, no. 2, pp. 646–659, 2020.

[16] F. Drakopanagiotakis, L. Wujak, M. Wygrecka, and P. Markart, "Biomarkers in idiopathic pulmonary fibrosis," *Matrix Biology*, vol. 68-69, pp. 404–421, 2018.

[17] H. Wang, M. Wang, K. Xiao et al., "Bioinformatics analysis on differentially expressed genes of alveolar macrophage in IPF," *Experimental Lung Research*, vol. 45, no. 9-10, pp. 288–296, 2019.

[18] S. Matsuda, J. D. Kim, F. Sugiyama et al., "Transcriptomic evaluation of pulmonary fibrosis-related genes: utilization of transgenic mice with modifying p38 signal in the lungs," *International Journal of Molecular Sciences*, vol. 21, no. 18, p. 6746, 2020.

[19] H. Wang, Q. Xie, W. Ou-Yang, and M. Zhang, "Integrative analyses of genes associated with idiopathic pulmonary fibrosis," *Journal of Cellular Biochemistry*, vol. 120, no. 5, pp. 8648–8660, 2019.

[20] D. H. Yu, X. L. Ruan, J. Y. Huang et al., "Analysis of the interaction network of hub miRNAs-hub genes, being involved in idiopathic pulmonary fibers and its emerging role in non-small cell lung cancer," *Frontiers in Genetics*, vol. 11, p. 302, 2020.

[21] Y. Chen, Q. Zhang, Y. Zhou, Z. Yang, and M. Tan, "Inhibition of miR-182-5p attenuates pulmonary fibrosis via TGF-β/Smad pathway," *Human & Experimental Toxicology*, vol. 39, no. 5, pp. 683–695, 2020.

[22] S. J. Cho, M. Lee, H. W. Stout-Delgado, and J. S. Moon, "DROSHA-dependent miRNA and AIM2 inflammasome activation in idiopathic pulmonary fibrosis," *International Journal of Molecular Sciences*, vol. 21, no. 5, p. 1668, 2020.

[23] J. Guiot, M. Cambier, A. Boeckx et al., "Macrophage-derived exosomes attenuate fibrosis in airway epithelial cells through delivery of antifibrotic miR-142-3p," *Thorax*, vol. 75, no. 10, pp. 870–881, 2020.

[24] P. Perge, A. Decmann, R. Pezzani et al., "Analysis of circulating extracellular vesicle-associated microRNAs in cortisol-producing adrenocortical tumors," *Endocrine*, vol. 59, no. 2, pp. 280–287, 2018.

[25] R. L. Montgomery, G. Yu, P. A. Latimer et al., "MicroRNA mimicry blocks pulmonary fibrosis," *EMBO Molecular Medicine*, vol. 6, no. 10, pp. 1347–1356, 2014.

[26] J. Hausser and M. Zavolan, "Identification and consequences of miRNA-target interactions – beyond repression of gene expression," *Nature Reviews. Genetics*, vol. 15, no. 9, pp. 599–612, 2014.

[27] I. G. Luzina, M. V. Salcedo, M. L. Rojas-Pena et al., "Transcriptomic evidence of immune activation in macroscopically normal- appearing and scarred lung tissues in idiopathic pulmonary fibrosis," *Cellular Immunology*, vol. 325, pp. 1–13, 2018.

[28] M. J. Cecchini, K. Hosein, C. J. Howlett, M. Joseph, and M. Mura, "Comprehensive gene expression profiling identifies distinct and overlapping transcriptional profiles in non-specific interstitial pneumonia and idiopathic pulmonary fibrosis," *Respiratory Research*, vol. 19, no. 1, p. 153, 2018.

[29] M. E. Ritchie, B. Phipson, Y. H. Di Wu, C. W. Law, W. Shi, and G. K. Smyth, "limma powers differential expression analyses for RNA-sequencing and microarray studies," *Nucleic Acids Research*, vol. 43, no. 7, p. e47, 2015.

[30] X. Jiao, B. T. Sherman, D. W. Huang et al., "DAVID-WS: a stateful web service to facilitate gene/protein list analysis," *Bioinformatics*, vol. 28, no. 13, pp. 1805-1806, 2012.

[31] H. Dweep, N. Gretz, and C. Sticht, "miRWalk database for miRNA-target interactions," *Methods in molecular biology*, vol. 1182, pp. 289–305, 2014.

[32] M. Li, D. Li, Y. Tang, F. Wu, and J. Wang, "CytoCluster: a Cytoscape plugin for cluster analysis and visualization of biological networks," *International Journal of Molecular Sciences*, vol. 18, no. 9, p. 1880, 2017.

[33] R. Oughtred, C. Stark, B. J. Breitkreutz et al., "The BioGRID interaction database: 2019 update," *Nucleic Acids Research*, vol. 47, no. D1, pp. D529–d541, 2019.

[34] S. Kerrien, B. Aranda, L. Breuza et al., "The IntAct molecular interaction database in 2012," *Nucleic Acids Research*, vol. 40, no. D1, pp. D841–D846, 2012.

[35] P. Heukels, C. C. Moor, J. H. von der Thüsen, M. S. Wijsenbeek, and M. Kool, "Inflammation and immunity in IPF pathogenesis and treatment," *Respiratory Medicine*, vol. 147, pp. 79–91, 2019.

[36] I. Y. Cheng, C. C. Liu, J. H. Lin et al., "Particulate matter increases the severity of bleomycin-induced pulmonary fibrosis through KC-mediated neutrophil chemotaxis," *International Journal of Molecular Sciences*, vol. 21, no. 1, p. 227, 2020.

[37] K. B. R. Belchamber and L. E. Donnelly, "Targeting defective pulmonary innate immunity - a new therapeutic option?," *Pharmacology & Therapeutics*, vol. 209, p. 107500, 2020.

[38] M. Zhang, W. Xin, Y. Yu et al., "Programmed death-ligand 1 triggers PASMCs pyroptosis and pulmonary vascular fibrosis in pulmonary hypertension," *Journal of Molecular and Cellular Cardiology*, vol. 138, pp. 23–33, 2020.

[39] V. Anathy, K. G. Lahue, D. G. Chapman et al., "Reducing protein oxidation reverses lung fibrosis," *Nature Medicine*, vol. 24, no. 8, pp. 1128–1135, 2018.

[40] W. Gong, P. Guo, L. Liu, Q. Guan, and Z. Yuan, "Integrative analysis of transcriptome-wide association study and mRNA expression profiles identifies candidate genes associated with idiopathic pulmonary fibrosis," *Frontiers in Genetics*, vol. 11, p. 604324, 2020.

[41] G. Raghu, L. Richeldi, A. Jagerschmidt et al., "Idiopathic pulmonary fibrosis: prospective, case-controlled study of natural history and circulating biomarkers," *Chest*, vol. 154, no. 6, pp. 1359–1370, 2018.

[42] M. Vukmirovic, J. D. Herazo-Maya, J. Blackmon et al., "Identification and validation of differentially expressed transcripts by RNA-sequencing of formalin-fixed, paraffin-embedded (FFPE) lung tissue from patients with idiopathic pulmonary fibrosis," *BMC Pulmonary Medicine*, vol. 17, no. 1, p. 15, 2017.

[43] X. Li, W. Chen, P. Li et al., "Follicular stimulating hormone accelerates atherogenesis by increasing endothelial VCAM-1 expression," *Theranostics*, vol. 7, no. 19, pp. 4671–4688, 2017.

[44] S. K. Kunutsor, S. J. L. Bakker, and R. P. F. Dullaart, "Soluble vascular cell adhesion molecules may be protective of future cardiovascular disease risk: findings from the PREVEND prospective cohort study," *Journal of Atherosclerosis and Thrombosis*, vol. 24, no. 8, pp. 804–818, 2017.

[45] M. Agassandian, J. R. Tedrow, J. Sembrat et al., "VCAM-1 is a TGF-β1 inducible gene upregulated in idiopathic pulmonary fibrosis," *Cellular Signalling*, vol. 27, no. 12, pp. 2467–2473, 2015.

[46] J. Mo, D. Zhang, and R. Yang, "MicroRNA-195 regulates proliferation, migration, angiogenesis and autophagy of endothelial progenitor cells by targeting GABARAPL1," *Bioscience Reports*, vol. 36, no. 5, 2016.

[47] S. H. Park, S. K. Kim, J. Y. Choe et al., "Hypermethylation of EBF3 and IRX1 genes in synovial fibroblasts of patients with rheumatoid arthritis," *Molecules and Cells*, vol. 35, no. 4, pp. 298–304, 2013.

[48] C. C. Sheu, W. A. Chang, M. J. Tsai, S. H. Liao, I. W. Chong, and P. L. Kuo, "Bioinformatic analysis of next-generation sequencing data to identify dysregulated genes in fibroblasts

Hindawi
BioMed Research International
Volume 2024, Article ID 9846925, 1 page
https://doi.org/10.1155/2024/9846925

Retraction

Retracted: Distinct Molecular Subtypes of Diffuse Large B Cell Lymphoma Patients Treated with Rituximab-CHOP Are Associated with Different Clinical Outcomes and Molecular Mechanisms

BioMed Research International

Received 12 March 2024; Accepted 12 March 2024; Published 20 March 2024

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope

(2) Discrepancies in the description of the research reported

(3) Discrepancies between the availability of data and the research described

(4) Inappropriate citations

(5) Incoherent, meaningless and/or irrelevant content included in the article

(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] H. Yu, S. Peng, S. Han, X. Chen, Q. Lyu, and T. Lei, "Distinct Molecular Subtypes of Diffuse Large B Cell Lymphoma Patients Treated with Rituximab-CHOP Are Associated with Different Clinical Outcomes and Molecular Mechanisms," *BioMed Research International*, vol. 2021, Article ID 5514726, 13 pages, 2021.

Hindawi
BioMed Research International
Volume 2021, Article ID 5514726, 13 pages
https://doi.org/10.1155/2021/5514726

Research Article

Distinct Molecular Subtypes of Diffuse Large B Cell Lymphoma Patients Treated with Rituximab-CHOP Are Associated with Different Clinical Outcomes and Molecular Mechanisms

Haifeng Yu,[1] Shuailing Peng,[1] Shuiyun Han,[1] Xi Chen,[1] Qinghua Lyu,[2] and Tao Lei[1]

[1] *Department of Lymphatic Medical Oncology, The Cancer Hospital of the University of Chinese Academy of Sciences (Zhejiang Cancer Hospital), Institute of Basic Medicine and Cancer (IBMC), Chinese Academy of Sciences, Hangzhou, 310022 Zhejiang, China*
[2] *Cancer Institute (Key Laboratory of Cancer Prevention & Intervention, National Ministry of Education, Provincial Key Laboratory of Molecular Biology in Medical Sciences), The Second Affiliated Hospital Zhejiang University School of Medicine, Hangzhou, 310009 Zhejiang, China*

Correspondence should be addressed to Tao Lei; leitao070713@163.com

Received 17 February 2021; Revised 27 April 2021; Accepted 1 June 2021; Published 21 June 2021

Academic Editor: Parameshachari B. D.

Objective. Our purpose was to characterize distinct molecular subtypes of diffuse large B cell lymphoma (DLBCL) patients treated with rituximab-CHOP (R-CHOP). *Methods*. Two gene expression datasets of R-CHOP-treated DLBCL patients were downloaded from GSE10846 ($n = 233$, training set) and GSE31312 ($n = 470$, validation set) datasets. Cluster analysis was presented via the ConsensusClusterPlus package in R. Using the limma package, differential expression analysis was utilized to identify feature genes. Kaplan-Meier survival analysis was presented to compare the differences in the prognosis between distinct molecular subtypes. Correlation between molecular subtypes and clinical features was analyzed. Based on the sets of highly expressed genes, biological functions were explored by gene set enrichment analysis (GSEA). Several feature genes were validated in the molecular subtypes via qRT-PCR and western blot. *Results*. DLBCL samples were clustered into two molecular subtypes. Samples in subtype I displayed poorer overall survival time in the training set ($p < 0.0001$). Consistently, patients in subtype I had shorter overall survival ($p = 0.0041$) and progression-free survival time ($p < 0.0001$) than those in subtype II. Older age, higher stage, and higher international prognostic index (IPI) were found in subtype I. In subtype I, T cell activation, lymphocyte activation, and immune response were distinctly enriched, while cell adhesion, migration, and motility were significantly enriched in subtype II. T cell exhaustion-related genes including TIM3 ($p < 0.001$), PD-L1 ($p < 0.0001$), LAG3 ($p < 0.0001$), CD160 ($p < 0.001$), and CD244 ($p < 0.001$) were significantly highly expressed in subtype I than subtype II. *Conclusion*. Two molecular subtypes were constructed in DLBCL, which were characterized by different clinical outcomes and molecular mechanisms. Our findings may offer a novel insight into risk stratification and prognosis prediction for DLBCL patients.

1. Introduction

Diffuse large B cell lymphoma (DLBCL) is the most common aggressive non-Hodgkin's lymphoma globally [1]. According to different cell origins, DLBCL is divided into three subtypes including germinal center B cell-like (GCB; 41%) and activated B cell-like (ABC; 35%) subtypes and others based on gene expression profile, which has become the standard method of prognosis in clinical practice [2]. Nevertheless, its prognostic effectiveness of this classification has not been uniformly proven due to the heterogeneity of classification structures [3]. The international prognostic index (IPI) is an effective clinical tool for predicting risk stratification and prognosis. However, it cannot guide personalized therapy [4]. The rituximab, cyclophosphamide, doxorubicin, and prednisone (R-CHOP) therapy is currently proven to be one of the most effective treatment regimen for most DLBCL subtypes. However, approximately 40% of patients will experience fatal recurrence or progression [5]. The 5-year overall survival rate of patients receiving first-line treatment is 60-

70% [6]. Through the first-line treatment of R-CHOP, most patients can be completely relieved. However, due to obscure reasons, some individuals in remission will develop relapse [7]. Hence, it is of importance to develop a novel prognostic stratification tool to accurately predict the prognosis of patients treated with R-CHOP and identify those who will experience immunosuppressive chemotherapy resistance.

Next-generation sequencing technology offers novel insights for individualized therapy of DLBCL patients. Moreover, some promising targets have been detected for the prevention and treatment of relapsed/refractory DLBCL [8]. For instance, coexpression of CD5 and CD43 may predict a poor prognosis of DLBCL patients [9]. A high NEAT1_1 level is positively correlated with stage, IPI, extranodal involvement, drug response, and poor prognosis [10]. Furthermore, LAMP1 expression is in relationship with IPI, overall survival, and progression-free survival for DLBCL [11]. Despite these biomarkers predicting the clinical outcomes of DLBCL, none of them have been translated into clinical practice. Thus, this study is aimed at screening and validating potential biomarkers for predicting survival outcomes of DLBCL patients and serving as therapeutic targets.

There is high heterogeneity in immune cells surrounding malignant B cells, which is related to the prognosis of DLBCL patients [12–14]. Chronic inflammation in DLBCL can suppress the differentiation and proliferation of T cells, caused by the continuous expression of inhibitory receptors, such as LAG3 and TIM3 [15]. By the suppression of the immune response, tumor cells are protected from immune surveillance [15]. It has been strikingly highlighted that the heterogeneity of DLBCL is reflected in molecular subtypes at the transcriptional level, which can provide insights into pathogenesis and candidate therapeutic targets for DLBCL [16]. Thus, it is of importance to characterize molecular subtypes of R-CHOP-treated DLBCL patients. Based on gene expression profiles of DLBCL, we aimed to characterize molecular subtypes with distinct prognoses and clinical features, which might improve the treatment strategy of DLBCL and prolong high-risk patients' survival time.

2. Materials and Methods

2.1. Data Collection and Preprocessing. From the Gene Expression Omnibus (GEO) repository (https://www.ncbi.nlm.nih.gov/gds/), we searched the gene expression profiles of DLBCL samples according to the following criteria: (1) patients were treated with R-CHOP, (2) patients had complete follow-up information, and (3) the number of patients was over 200. As a result, GSE10846 and GSE31312 datasets were included in this study. The GSE10846 dataset including 233 DLBCL patients treated with R-CHOP based on the GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array platform was used as the training set [17, 18]. The GSE31312 dataset including 470 DLBCL patients on the platform of GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array was utilized as the validation set. When a gene corresponded to multiple probes, we took the average value as the expression value of the gene. Clinical information including age, stage, sex, IPI, overall survival time, and disease-free survival time was also extracted from the two datasets. Among them, the IPI scoring standard is based on the patient's age, general condition score, clinical stage, involvement of extranodal sites, and lactate dehydrogenase.

2.2. Consensus Cluster Analysis. Consensus cluster analysis was performed to determine the optimal clustering number (k) by the ConsensusClusterPlus package in R [19]. The stability of the cluster was evaluated through resampling-based methods.

2.3. Differential Expression Analysis. Differential expression analysis was presented between the two molecular subtypes on raw data or Microarray Analysis (limma) package in R [20]. Genes with a false discovery rate (FDR) < 0.05 were set as feature genes. On the basis of these feature genes, the samples in the validation set were used for cluster analysis via the nearest template prediction (NTP) algorithm [21].

2.4. Kaplan-Meier Survival Analysis. Kaplan-Meier overall and progression-free survival was presented for patients between the two molecular subtypes via Survival package in R, which was assessed by the log-rank test. The definition of overall survival refers to the time from the beginning of randomization to the death of a patient from any cause. Progression-free survival is defined as the time from diagnosis to any cause leading to progression, recurrence, or death. Multivariate Cox regression analysis was presented to assess whether the molecular subtypes could independently predict overall and progression-free survival time following adjusting gene expression profile classification.

2.5. Gene Set Enrichment Analysis (GSEA). Gene Ontology (GO) biological process enrichment analysis was separately presented on the basis of two sets of highly expressed genes in the two molecular subtypes using the GSEA software (http://software.broadinstitute.org/gsea/index.jsp) [22, 23]. The number of permutations was set as 1000. Adjusted p value < 0.05 was set as the threshold value.

2.6. Correlation Analysis. We further analyzed the relationships between the molecular subtypes and other clinical factors including age, stage, sex, and IPI. The differences between the two subtypes were assessed via the Wilcoxon rank-sum test.

2.7. Patient Samples. Totally, 30 DLBCL patients and matched 30 healthy individuals were recruited in The Cancer Hospital of the University of Chinese Academy of Sciences between 2019 and 2020 in this study. All patients were diagnosed by experienced pathologists. Formalin-fixed paraffin-embedded (FFPE) biopsy specimens were used for qRT-PCR and western blots. Each patient signed an informed consent form. This research project gained the approval of The Cancer Hospital of The University of Chinese Academy of Sciences ethics committee (2019-037).

2.8. Quantitative Real-Time PCR (qRT-PCR). After the tissue was lysed by TRIzol (15596-018; Invitrogen, Carlsbad, California, USA), the sample was transferred to a 1.5 ml EP tube. 200 μl chloroform (100006818; Sinopharm Chemical

Reagent Co., Ltd, Shanghai, China) was added to the EP tube and left at room temperature for 5 min. After 12000 rpm centrifugation for 15 min at 4°C, the upper aqueous phase was transferred to a new 1.5 ml EP tube; 400 μl isopropanol (A507048; Sangon Biotech, Shanghai) was added and allowed to stand at room temperature for 10 min. Following centrifugation at 12000 rpm for 10 min at 4°C, the supernatant was discarded. A spectrophotometer (752; Shanghai Sunny Heng Scientific Instrument Co., Ltd.) was used to determine RNA concentration. The total RNA was reverse transcribed into cDNA via the RNA cDNA first strand synthesis kit (AT341; TransGen Biotech, Beijing, China). Fluorescence quantitative PCR detection was presented by the ABI StepOnePlus type fluorescence quantitative PCR instrument. GAPDH was used as an internal control. Table 1 lists the information of primers.

2.9. Western Blot. The tissue samples were lysed in RIPA lysis buffer (P0013B; Beyotime, Beijing, China) at 4°C for 30 min. The protein concentration was determined with the BCA kit (P0010; Beyotime). The absorbance at OD568 was measured with a microplate reader (EPOCH2; BioTek, Vermont, USA). The sample was separated by SDS-PAGE gel and transferred to a PVDF membrane. The PVDF membrane was sealed with TBST blocking solution containing 5% skimmed milk powder at room temperature for 30 min. The membrane was incubated with the primary antibodies against CD244 (1 : 1000; ab196745; Abcam, USA), TIM3 (1 : 1000; ab185703), CD160 (1/200; ab202845), LAG3 (1 : 1000; ab237718), and GAPDH (1 : 20000; #5174; cst, USA) overnight at 4°C. After washing the PVDF membrane using TBST 5-6 times, the PVDF membrane was soaked with anti-mouse (1 : 5000; #7076) or anti-rabbit (1 : 5000; #7074) IgG (HRP) secondary antibodies with TBST for 2 h at 37°C. The enhancement solution in the ECL reagent was mixed with the stable peroxidase solution in a ratio of 1 : 1. Then, the mixture was added to the PVDF membrane and protein band was visualized.

2.10. Construction of a Gene Signature. Univariate Cox regression analysis was presented for screening survival-related genes with $p < 0.05$. The top 40 genes were selected for multivariate Cox regression analysis with a stepwise method. Finally, a gene signature was constructed. Patients in the training set were separated into the high and low score groups. Kaplan-Meier of overall survival was presented. The prognostic value was validated in the validation set. Furthermore, the predictive independency of this signature was evaluated in different subtypes.

2.11. Statistical Analysis. All statistical analyses were conducted using R language (https://www.r-project.org/) and GraphPad Prism 8 (GraphPad Software Inc., La Jolla, CA). p value < 0.05 was set as the evaluation criteria.

3. Results

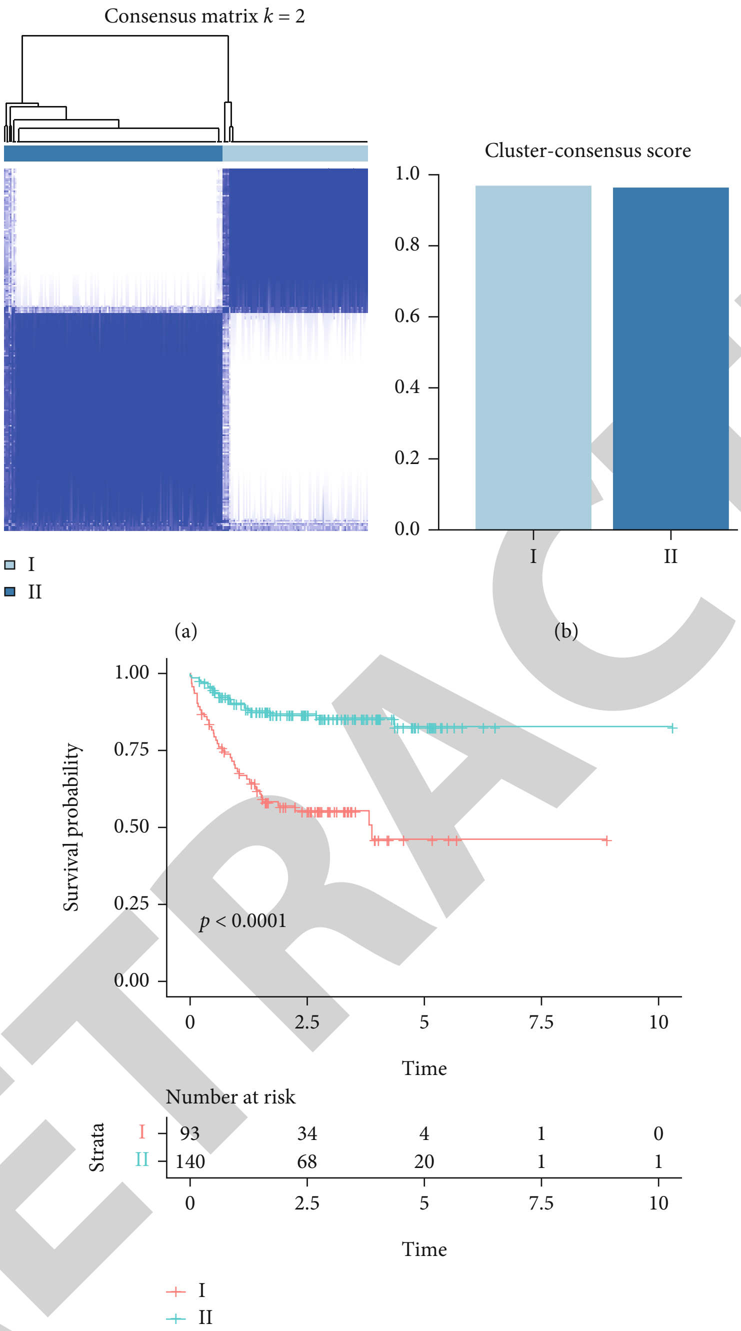
3.1. Development of Two Distinct Molecular Subtypes for R-CHOP-Treated DLBCL Patients. Totally, 233 R-CHOP-treated DLBCL samples were included in the training set, which were clustered by the ConsensusClusterPlus package in R. When $k = 2$, two molecular subtypes were conducted (Figure 1(a)). Following resampling, the cluster-consensus scores of the two subtypes were both higher than 0.8, indicating that the cluster analysis had high stability (Figure 1(b)). The overall survival results showed that the patients' prognosis was significantly different between the two subtypes (Figure 1(c)). Patients in the subtype I group exhibited a worse prognosis than those in the subtype II group ($p < 0.0001$; Figure 1(c)). This indicated that there was a distinct difference in clinical outcomes between the two subtypes.

Table 1: Primer sequences for quantitative real-time PCR.

Target	Primer sequence (5′-3′)	Product (bp)
h-GAPDH-F	ACAACTTTGGTATCGTGGAAGG	101
h-GAPDH-R	GCCATCACGCCACAGTTTC	
h-TIM3-F	TTGGACATCCAGATACTGGCT	86
h-TIM3-R	CACTGTCTGCTAGAGTCACATTC	
h-PDL1-F	TGGCATTTGCTGAACGCATTT	120
h-PDL1-R	TGCAGCCAGGTCTAATTGTTTT	
h-LAG3-F	GCCTCCGACTGGGTCATTTT	131
h-LAG3-R	CTTTCCGCTAAGTGGTGATGG	
h-CD160-F	GCTGAGGGGTTTGTAGTGTTT	154
h-CD160-R	GTGTGACTTGGCTTATGGTGA	
h-CD244-F	TCGTGATTCTAAGCGCACTGT	237
h-CD244-R	CAGGTTCTTGTGACGTGGGAG	

3.2. Molecular Subtypes Are Associated with Distinct Clinical Outcomes. Based on the gene expression profiles, feature genes with FDR < 0.05 were screened between subtypes I and II in the training set. According to these feature genes, 383 samples in the validation dataset were divided into two subtypes using the NTP algorithm. As a result, these samples were significantly clustered into subtypes I and II. We further analyzed the differences in overall and progression-free survival time between the two molecular subtypes. The results showed that R-CHOP-treated DLBCL patients in subtype I significantly exhibited shorter overall survival time than those in subtype II ($p = 0.0041$; Figure 2(a)), which were consistent with the results from the training set. Furthermore, we found that patients in subtype I usually experienced poorer progression-free survival time in comparison to those in subtype II ($p < 0.0001$; Figure 2(b)). Also, we performed multivariate Cox regression analysis. Consistently, there were also distinct differences in overall survival and progression-free survival time between the two subtypes after adjusting gene expression profile classifications (Table 2). Thus, the two molecular subtypes could be associated with distinct clinical outcomes.

3.3. Different Clinicopathological Features of Molecular Subtypes. We analyzed the differences in clinicopathological



(a)

(b)

(c)

Figure 1: Development of two distinct molecular subtypes for R-CHOP-treated DLBCL patients. (a) A consensus matrix heat map when $k = 2$. The row and column of the matrix represent different samples. The values of the consensus matrix range from 0 (cannot be clustered together) to 1 (always clustered together) in white to dark blue. The consensus matrix is arranged in accordance with the consistency classification (the tree diagram above the heat map). (b) Cluster-consensus scores of the two molecular subtypes when $k = 2$. The higher the score, the higher the stability. (c) Kaplan-Meier overall survival analysis for patients in the two molecular subtypes. Red represents the subtype I group, and blue represents the subtype II group.

features between molecular subtypes via the Wilcoxon rank-sum test. In Figure 3(a), age in subtype I was significantly higher than that in subtype II ($p < 0.01$). Compared to subtype II, there were fewer DLBCL samples at stage 1 in subtype I ($p < 0.05$; Figure 3(b)). In subtype I, more clinical samples were at stage 4 in comparison to subtype II ($p < 0.01$). Thus, these molecular subtypes were highly correlated to the degree of malignancy of DLBCL. As shown in Figure 3(c), there was no statistical difference in sex between the two molecular subtypes ($p > 0.05$). Regarding the international prognostic index (IPI), the percentages of IPI > 3 samples were significantly higher in subtype I than subtype II ($p < 0.05$) in Figure 3(d).

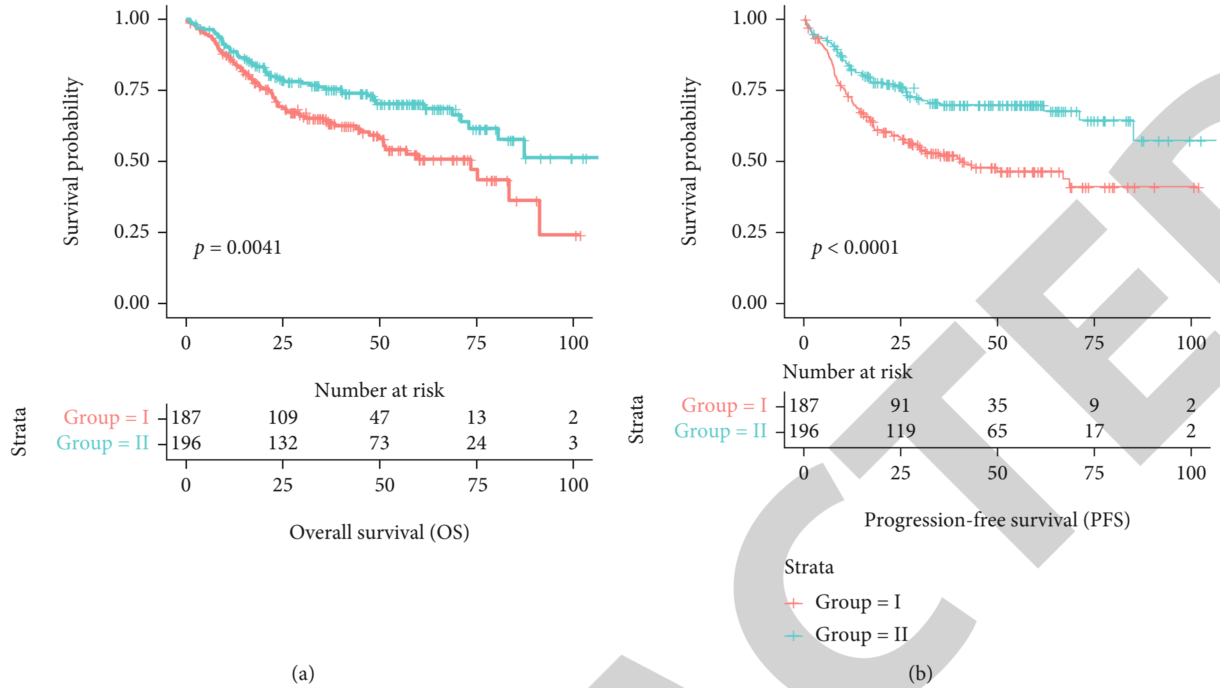


Figure 2: Molecular subtypes are associated with distinct clinical outcomes. (a) Overall survival analysis of R-CHOP-treated DLBCL patients in subtypes I and II. (b) Progression-free survival analysis of patients in subtypes I and II. Red represents subtype I, and blue indicates subtype II.

Table 2: Independence of the two subtypes in overall and progression-free survival prediction.

Survival	Variables	coef	exp(coef)	se(coef)	z	p
Overall survival	Group: II	-1.1084	0.3301	0.3426	-3.2348	$6.09E-04$
	Subtype: GCB	-0.7356	0.4792	0.3571	-2.0599	$3.94E-02$
	Subtype: UC	-0.6358	0.5295	0.4019	-1.5820	$1.14E-01$
Progression-free survival	Group: II	-1.3205	0.2670	0.3258	-4.0535	$2.52E-05$
	Subtype: GCB	-0.4108	0.6631	0.3380	-1.2154	$2.24E-01$
	Subtype: UC	-0.4443	0.6412	0.3687	-1.2053	$2.28E-01$

Abbreviations: coef: coefficients; exp: exponential; se: standard error; z: z-score; p: p value; GCB: germinal center B cell; UC: unclassified.

3.4. GSEA and Identification of T Cell Exhaustion-Related Genes. To further probe into underlying biological processes of these genes in the two subtypes, GSEA was carried out. Highly expressed genes in subtype I were mainly enriched in lymphocyte activation, T cell activation, regulation of leukocyte activation, cellular response to interferon-gamma, antigen processing and presentation of exogenous peptide antigen via MHC class I, regulation of lymphocyte activation, antigen processing and presentation of peptide antigen via MHC class I, response to lipopolysaccharide, activation of immune response, and lymphocyte proliferation (Figure 4(a)). Highly expressed genes in subtype II were distinctly enriched in extracellular matrix organization, extracellular structure organization, homophilic cell adhesion, cell-cell adhesion, extracellular matrix disassembly, collagen catabolic process, regulation of cellular component movement, regulation of cell migration, regulation of cell motility, and multicellular organismal catabolic process (Figure 4(a)). In Figure 4(b), T cell exhaustion-related genes including TIM3 ($p < 0.001$), PD-L1 ($p < 0.0001$), LAG3 ($p < 0.0001$), CD160 ($p < 0.001$), and CD244 ($p < 0.001$) had distinctly higher expression levels in subtype I than in subtype II.

3.5. Validation of T Cell Exhaustion-Related Genes in DLBCL. qRT-PCR was used to validate the mRNA expression of T cell exhaustion-related genes in 30 paired DLBCL and healthy specimens. Consistent with our bioinformatics results, TIM3 ($p < 0.001$; Figure 5(a)), PD-L1 ($p < 0.05$; Figure 5(b)), LAG3 ($p < 0.05$; Figure 5(c)), CD160 ($p < 0.05$; Figure 5(d)), and CD244 ($p < 0.05$; Figure 5(e)) displayed significantly higher mRNA expression levels in DLBCL than healthy specimens. Consistently, western blot results also showed that TIM3 ($p < 0.001$), PD-L1 ($p < 0.01$), LAG3 ($p < 0.001$), CD160 ($p < 0.001$), and CD244 ($p < 0.001$) proteins exhibited higher expression levels in DLBCL specimens in comparison to healthy specimens (Figures 6(a) and 6(b)).

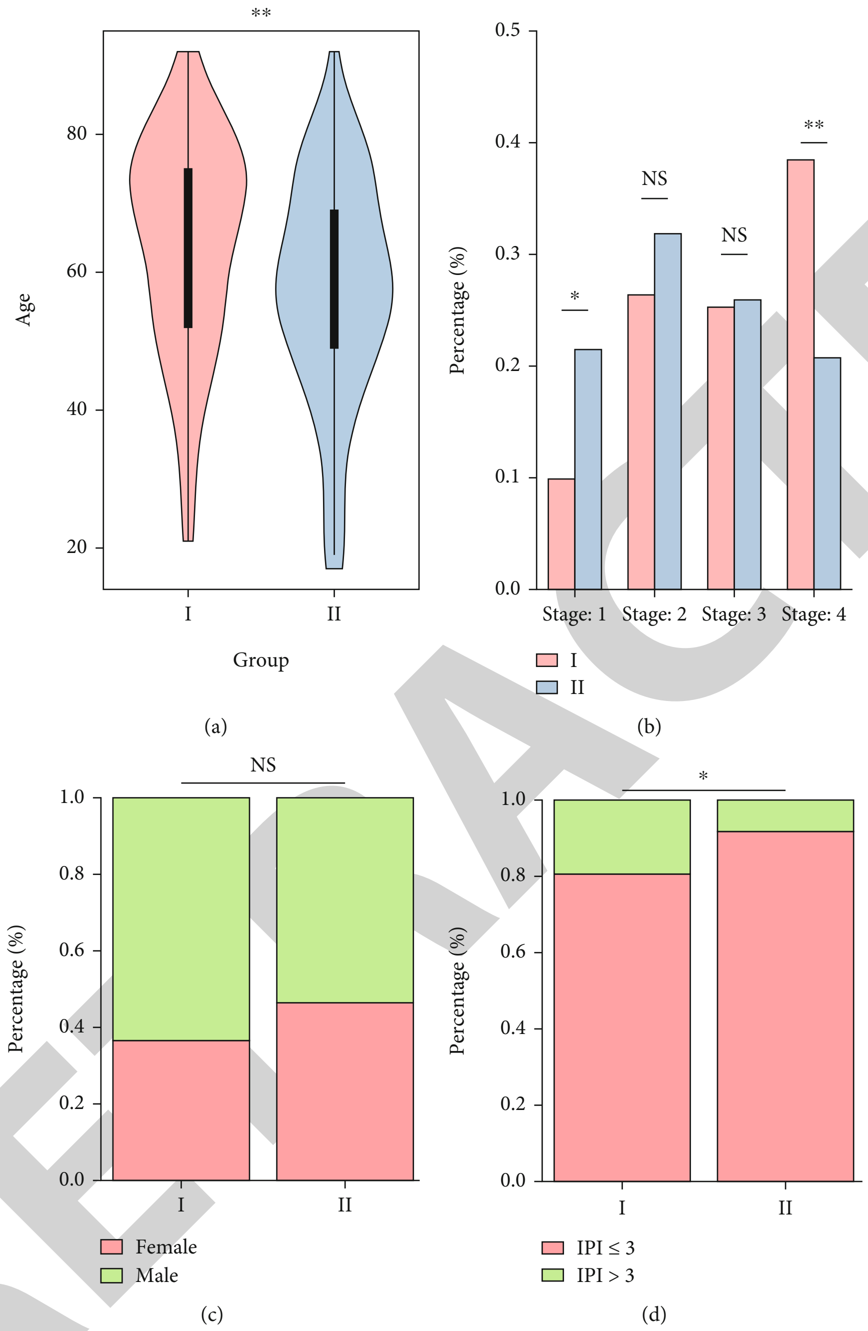
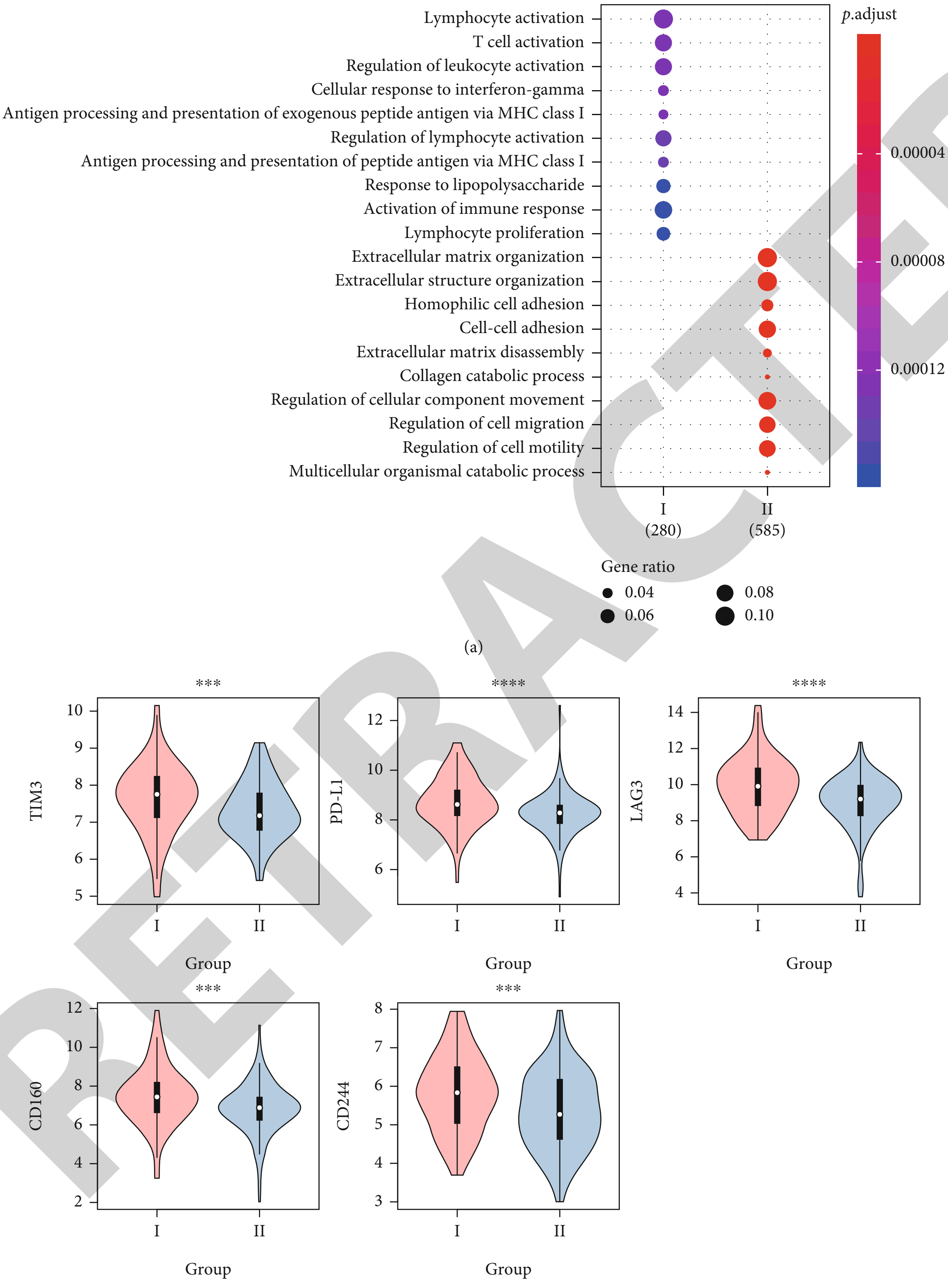


Figure 3: Correlation between different clinicopathological features and molecular subtypes. (a) Violin diagram depicting the differences in age between the two molecular subtypes. (b) Stage distributions between the two molecular subtypes. (c) Sex distributions of samples between the two molecular subtypes. (d) Differences in international prognostic index (IPI) between the two molecular subtypes. NS: $p > 0.05$; $^{*}p < 0.05$; $^{**}p < 0.01$.

3.6. Development of a Prognostic Gene Signature for DLBCL. By applying univariate Cox regression analysis, 375 survival-related genes were identified for DLBCL. We selected the top 40 genes for multivariate Cox regression analysis. As a result, a 10-gene signature was established, including TRPC4, TIMP1, PPP1R7, NPIPB11, NLK, NCOA1, LMO2, CPNE3, CD3EAP, and CD209 (Figure 7(a)). In the training set, patients with high-risk scores indicated undesirable outcomes than those with low-risk scores ($p < 0.0001$; Figure 7(b)). The predictive efficacy was confirmed in the validation set ($p < 0.0001$; Figure 7(c)). Furthermore, both in the germinal center (GC) and non-GC groups, high-risk scores were indicative of shorter overall and progression-free survival time (Figure 8(a)). For ABC or GCB subtype, high-risk scores predicted poorer survival outcomes (Figure 8(b)).

4. Discussion

R-CHOP can relieve about 60% to 70% of patients. However, the remaining patients may relapse within 2-3 years after



(a)

(b)

Figure 4: GSEA and identification of T cell exhaustion-related genes. (a) The top ten biological processes enriched by highly expressed genes in subtype I or II, respectively. Blue bubbles represent the annotation results of highly expressed genes in subtype I. Red bubbles express the annotation results of highly expressed genes in subtype II. The size of the bubble is proportional to the number of enriched genes. The darker the color, the smaller the adjusted p value. (b) Violin diagram showing the expression of T cell exhaustion-related genes including TIM3, PD-L1, LAG3, CD160, and CD244 between subtypes I (red) and II (blue). $^{***}p < 0.001$; $^{****}p < 0.0001$.

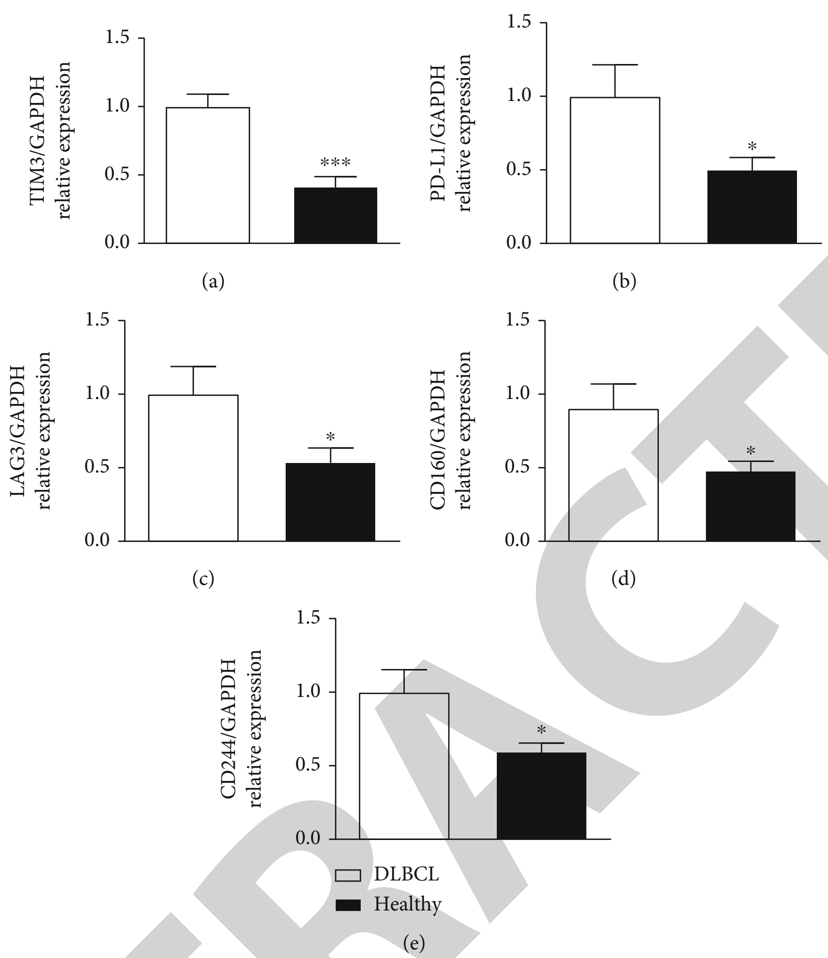


Figure 5: Validation of T cell exhaustion-related genes in DLBCL via qRT-PCR. The mRNA expression levels of TIM3 (a), PD-L1 (b), LAG3 (c), CD160 (d), and CD244 (e) were compared between DLBCL and control specimens. $^{*}p < 0.05$; $^{***}p < 0.001$.

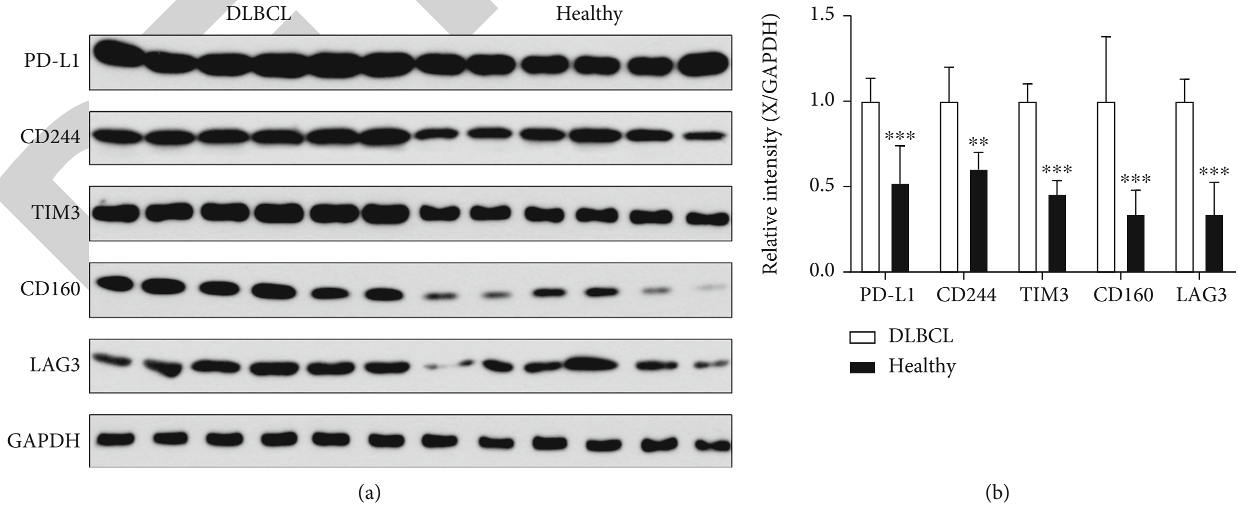
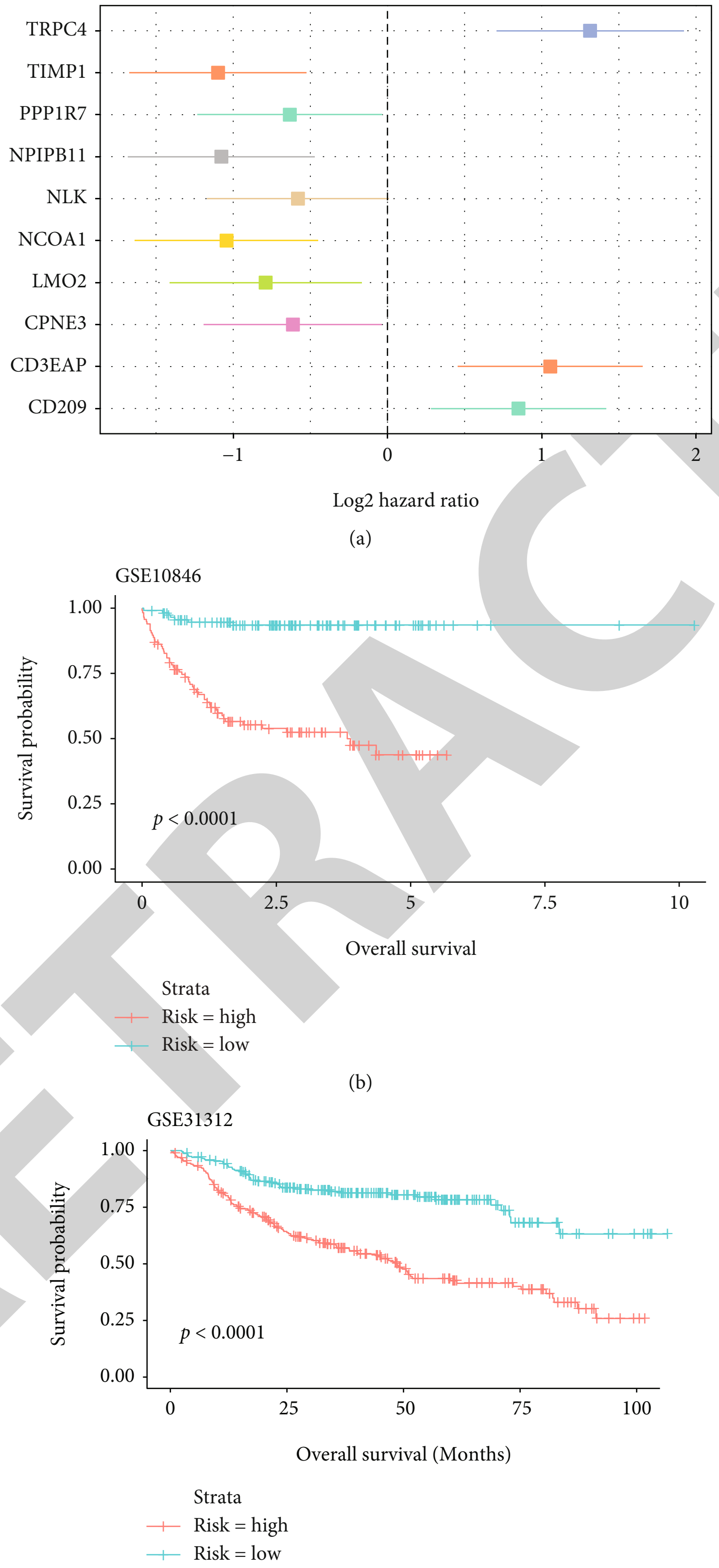


Figure 6: Western blot detecting the expression of T cell exhaustion-related genes in DLBCL. (a) Representative images of western blot results. (b) The protein expression levels of TIM3, PD-L1, LAG3, CD160, and CD244 were compared between DLBCL and control specimens. $^{**}p < 0.01$; $^{***}p < 0.001$.

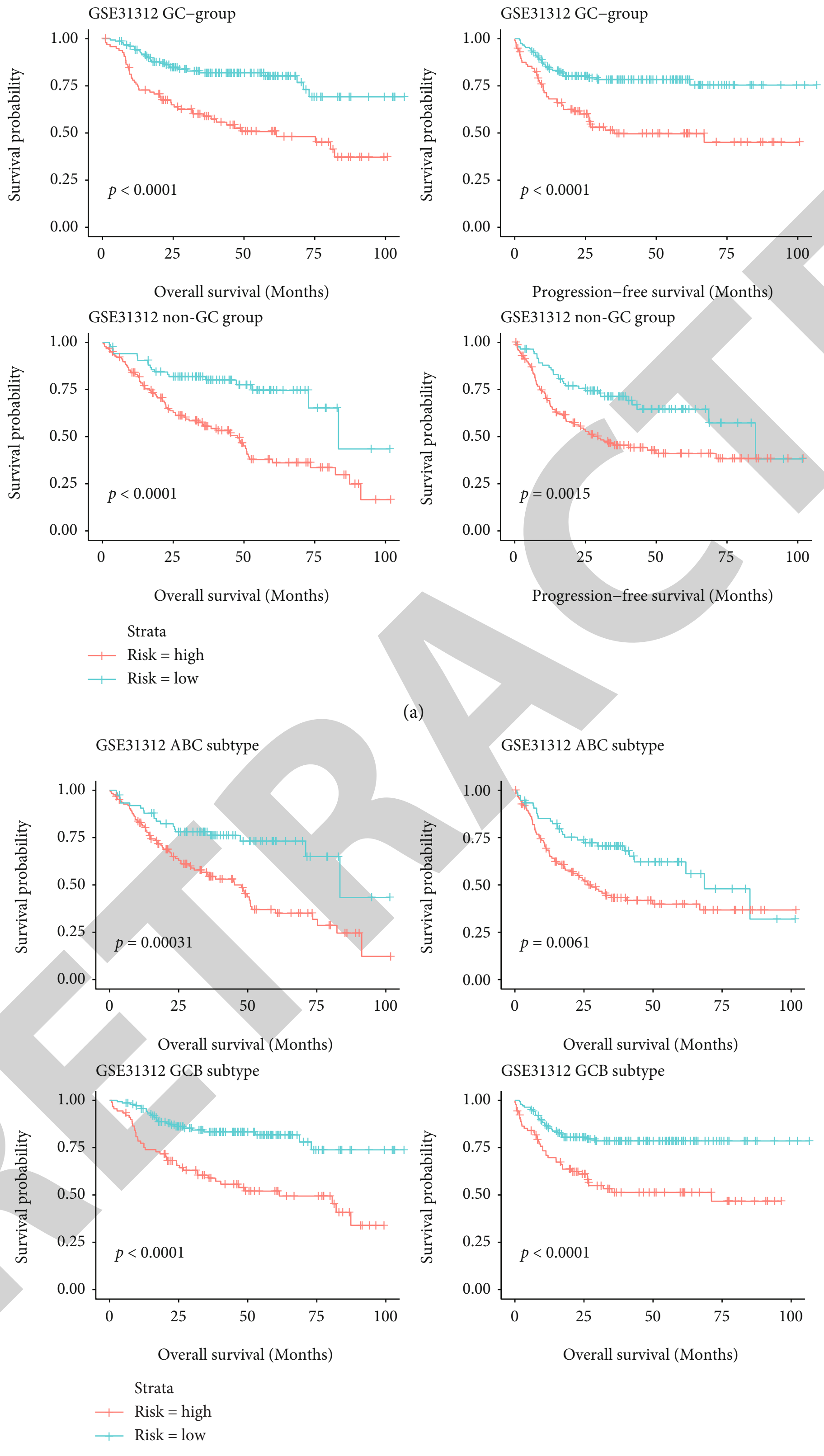


(a)

(b)

(c)

Figure 7: Establishment and validation of a 10-gene signature for DLBCL. (a) Forest plot for the hazard ratio of the 10 genes in this signature. (b, c) Kaplan-Meier curves of overall survival in the (b) training and (c) validation sets.



(a)

(b)

FIGURE 8: Subgroup survival analysis of the 10-gene signature for DLBCL. (a, b) Kaplan-Meier curves of overall and progression-free survival in the (a) GC and non-GC groups as well as (b) ABC and GCB subtypes.

treatment [2]. The choice of salvage therapy is very poor, with an adverse reaction rate of about 20%. How to predict the prognosis of patients in the early stage, so as to choose more effective treatment strategies for high-risk patients, is vital to saving lives [2].

Due to the high heterogeneity of DLBCL, it is necessary to identify specific molecular subtypes of DLBCL and identify biomarkers to predict the clinical outcomes of high-risk groups. Herein, DLBCL samples were mainly clustered into the two molecular subtypes (Figures 1(a) and 1(b)). In the training set, patients in subtype I exhibited shorter overall survival time in comparison to those in subtype II (Figure 1(c)). After validation, there were distinct differences in overall survival and progression-free time between the two molecular subtypes (Figures 2(a) and 2(b)). The prevalence of DLBCL in the elderly is increasing [24]. Age over 60 has been considered a risk factor for DLBCL [25]. Increasing age is in association with more complex biological behaviors. Our results showed that patients in subtype I had older age than those in subtype II (Figure 3(a)). Furthermore, compared to the patients in subtype II, more patients are in advanced stages in subtype I (Figure 3(b)). No significant difference in sex was found between the two subtypes (Figure 3(c)). IPI is widely applied for risk stratification of patients with DLBCL (Figure 3(d)). The 3-year overall survival rates of patients with IPI scores of 0-1, 2, 3, and 4-5 were 91%, 81%, 65%, and 59%, respectively [26]. For patients in subtype I, the percentage of IPI score > 3 was distinctly higher than those in subtype II. However, due to the addition of rituximab to the CHOP regimen, the ability of IPI to distinguish high and low risk has decreased. Efforts to characterize the prognosis of DLBCL using immunohistochemistry have identified a variety of genetic markers with prognostic significance. These novel prognostic markers are independent of IPI but have few impacts on their prognostic capacity, largely due to the inherent limitations of the application of these markers [27].

In spite of a deep understanding concerning related signal transduction pathways among high-risk DLBCL populations, randomized phase III trials of integration of targeted therapy and R-CHOP regrettably failed to ameliorate the prognosis of DLBCL patients [28, 29]. 280 highly expressed genes in subtype I were mainly enriched in T cell activation, lymphocyte activation, and immune response (Figure 4(a)). Moreover, cell adhesion, cell migration, and motility were significantly enriched by 585 highly expressed genes in subtype II (Figure 4(a)). In chronic infections and cancer, T cells are exposed to persistent antigens or inflammatory signals. This process is usually related to the deterioration of T cell function. Exhausted T cells lose their effector functions, express a variety of inhibitory receptors, change the expression and use of key transcription factors, develop metabolic disorders, and fail to transition to a quiescent state [30].

The heterogeneity of clinical outcomes can be partly attributed to genetic heterogeneity [31]. Therefore, we further analyzed the differentially expressed genes between the two subtypes to explain the reasons for the differences in clinical outcomes involving DLBCL patients receiving R-CHOP therapy. Among them, T cell exhaustion-related genes including TIM3, PD-L1, LAG3, CD160, and CD244 were significantly higher in subtype I than in subtype II (Figure 4(b)). TIM3 is a membrane of the T lymphocyte immunoglobulin mucin (TIM) family, which is expressed in T helper 1 (Th1) and cytotoxic T cells (Tc1). As a negative regulator, it induces the apoptosis of Th1 and Tc1 cells [32]. Compared with healthy controls, DLBCL patients have higher expression of TIM3. Its expression was positively related to the DLBCL stage [33]. TIM3 expression can reflect the treatment efficiency of patients with chemotherapy [33]. LAG3 is a member of the immunoglobulin superfamily, which acts as a negative regulator of T cell homeostasis [34]. LAG3 is coexpressed with TIM3 and PD-L1 in DLBCL [34]. CD160 is an Ig-like glycoprotein expressed by natural killer cells and T cell subset [35]. Upregulation of PD-L1 can increase the immune escape of cancer cells in DLBCL [36]. CD244 is a family member of the signal lymphocyte activation molecule of immune cell receptors [37]. Our qRT-PCR and western blot confirmed the higher expression of TIM3, PD-L1, LAG3, CD160, and CD244 in DLBCL (Figures 5 and 6). Finally, we developed a 10-gene signature for predicting the prognosis of DLBCL patients (Figures 7 and 8). Thus, our research might provide potential information for precise drug treatment strategies and prognosis prediction for DLBCL.

Taken together, we constructed and confirmed two molecular subtypes of DLBCL with distinct prognosis features. The molecular subtype classifications may be adapted to the real world. In our future studies, we will validate the classifications in our DLBCL cohort. Moreover, several feature genes were identified, which might become promising therapeutic targets for future immunotherapy. Despite these feature genes being validated via qRT-PCR and western blot, their functions should be verified in a larger cohort of DLBCL.

5. Conclusion

In this study, we constructed and externally verified two novel molecular subtypes with distinct prognosis and clinical characteristics for DLBCL by consensus cluster analysis, which could be used for risk stratification and prognosis prediction in clinical practice.

Abbreviations

DLBCL:	Diffuse large B cell lymphoma
GSEA:	Gene set enrichment analysis
IPI:	International prognostic index
R-CHOP:	Rituximab, cyclophosphamide, doxorubicin, and prednisone
GEO:	Gene Expression Omnibus
FDR:	False discovery rate
NTP:	Nearest template prediction
GO:	Gene Ontology
qRT-PCR:	Quantitative real-time PCR
coef:	Coefficients
exp:	Exponential
se:	Standard error
z:	*z*-score

p: p value
GCB: Germinal center B cell
UC: Unclassified.

Data Availability

The datasets analyzed during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This work was funded by the Natural Science Foundation of Zhejiang Province (LY19C080001 and LY19C080002).

References

[1] Y. Liu and S. K. Barta, "Diffuse large B-cell lymphoma: 2019 update on diagnosis, risk stratification, and treatment," *American Journal of Hematology*, vol. 94, no. 5, pp. 604–616, 2019.

[2] L. Li, J. Zhang, J. Chen et al., "B-cell receptor-mediated NFATc1 activation induces IL-10/STAT3/PD-L1 signaling in diffuse large B-cell lymphoma," *Blood*, vol. 132, no. 17, pp. 1805–1817, 2018.

[3] A. M. Staiger, M. Ziepert, H. Horn et al., "Clinical impact of the cell-of-origin classification and the MYC/BCL2 dual expresser status in diffuse large B-cell lymphoma treated within prospective clinical trials of the German High-Grade Non-Hodgkin's Lymphoma Study Group," *Journal of Clinical Oncology*, vol. 35, no. 22, pp. 2515–2526, 2017.

[4] J. L. Crombie and P. Armand, "Diffuse large B-cell lymphoma and high-grade B-cell lymphoma: genetic classification and its implications for prognosis and treatment," *Hematology/Oncology Clinics of North America*, vol. 33, no. 4, pp. 575–585, 2019.

[5] J. M. Yang, J. Y. Jang, Y. K. Jeon, and J. H. Paik, "Clinicopathologic implication of microRNA-197 in diffuse large B cell lymphoma," *Journal of Translational Medicine*, vol. 16, no. 1, p. 162, 2018.

[6] S. Li, K. H. Young, and L. J. Medeiros, "Diffuse large B-cell lymphoma," *Pathology*, vol. 50, no. 1, pp. 74–87, 2018.

[7] G. Liu, J. Luan, and Q. Li, "CD4(+)Foxp 3(-)IL-10(+) Tr1 cells promote relapse of diffuse large B cell lymphoma by enhancing the survival of malignant B cells and suppressing antitumor T cell immunity," *DNA and Cell Biology*, vol. 35, no. 12, pp. 845–852, 2016.

[8] J. W. Friedberg, "Relapsed/refractory diffuse large B-cell lymphoma," *Hematology. American Society of Hematology. Education Program*, vol. 2011, no. 1, pp. 498–505, 2011.

[9] X. B. Ma, Y. P. Zhong, Y. Zheng, J. Jiang, and Y. P. Wang, "Coexpression of CD5 and CD43 predicts worse prognosis in diffuse large B-cell lymphoma," *Cancer Medicine*, vol. 7, no. 9, pp. 4284–4295, 2018.

[10] L. Deng, L. Jiang, K. F. Tseng et al., "Aberrant NEAT1_1 expression may be a predictive marker of poor prognosis in diffuse large B cell lymphoma," *Cancer Biomarkers*, vol. 23, no. 2, pp. 157–164, 2018.

[11] Q. Dang, H. Zhou, J. Qian et al., "LAMP1 overexpression predicts for poor prognosis in diffuse large B-cell lymphoma," *Clinical Lymphoma, Myeloma & Leukemia*, vol. 18, no. 11, pp. 749–754, 2018.

[12] E. Bachy and G. Salles, "Treatment approach to newly diagnosed diffuse large B-cell lymphoma," *Seminars in Hematology*, vol. 52, no. 2, pp. 107–118, 2015.

[13] S. Dubois, P. J. Viailly, S. Mareschal et al., "Next-generation sequencing in diffuse large B-cell lymphoma highlights molecular divergence and therapeutic opportunities: a LYSA study," *Clinical Cancer Research*, vol. 22, no. 12, pp. 2919–2928, 2016.

[14] L. Pasqualucci, "Molecular pathogenesis of germinal center-derived B cell lymphomas," *Immunological Reviews*, vol. 288, no. 1, pp. 240–261, 2019.

[15] M. Autio, S.-K. Leivonen, O. Brück et al., "Immune cell constitution in the tumor microenvironment predicts the outcome in diffuse large B-cell lymphoma," *Haematologica*, vol. 106, no. 3, pp. 718–729, 2020.

[16] B. Chapuy, C. Stewart, A. J. Dunford et al., "Molecular subtypes of diffuse large B cell lymphoma are associated with distinct pathogenic mechanisms and outcomes," *Nature Medicine*, vol. 24, no. 5, pp. 679–690, 2018.

[17] T. M. Cardesa-Salzmann, L. Colomo, G. Gutierrez et al., "High microvessel density determines a poor outcome in patients with diffuse large B-cell lymphoma treated with rituximab plus chemotherapy," *Haematologica*, vol. 96, no. 7, pp. 996–1001, 2011.

[18] G. Lenz, G. Wright, S. S. Dave et al., "Stromal gene signatures in large-B-cell lymphomas," *The New England Journal of Medicine*, vol. 359, no. 22, pp. 2313–2323, 2008.

[19] M. D. Wilkerson and D. N. Hayes, "ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking," *Bioinformatics*, vol. 26, no. 12, pp. 1572-1573, 2010.

[20] M. E. Ritchie, B. Phipson, Y. H. Di Wu, C. W. Law, W. Shi, and G. K. Smyth, "limma powers differential expression analyses for RNA-sequencing and microarray studies," *Nucleic Acids Research*, vol. 43, no. 7, 2015.

[21] Y. Hoshida, "Nearest template prediction: a single-sample-based flexible class prediction with confidence assessment," *PLoS One*, vol. 5, no. 11, article e15543, 2010.

[22] A. Subramanian, P. Tamayo, V. K. Mootha et al., "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 43, pp. 15545–15550, 2005.

[23] V. K. Mootha, C. M. Lindgren, K. F. Eriksson et al., "PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes," *Nature Genetics*, vol. 34, no. 3, pp. 267–273, 2003.

[24] V. A. Morrison, P. Hamlin, P. Soubeyran et al., "Approach to therapy of diffuse large B-cell lymphoma in the elderly: the International Society of Geriatric Oncology (SIOG) expert position commentary," *Annals of Oncology*, vol. 26, no. 6, pp. 1058–1068, 2015.

[25] G. Hedström, O. Hagberg, M. Jerkeman, and G. Enblad, "The impact of age on survival of diffuse large B-cell lymphoma - a population-based study," *Acta Oncologica*, vol. 54, no. 6, pp. 916–923, 2015.

[26] M. Ziepert, D. Hasenclever, E. Kuhnt et al., "Standard international prognostic index remains a valid predictor of outcome for patients with aggressive CD20+ B-cell lymphoma in the rituximab era," *Journal of Clinical Oncology*, vol. 28, no. 14, pp. 2373–2380, 2010.

Hindawi
BioMed Research International
Volume 2024, Article ID 9854370, 1 page
https://doi.org/10.1155/2024/9854370

Retraction

Retracted: A Prognostic Model for Brain Glioma Patients Based on 9 Signature Glycolytic Genes

BioMed Research International

Received 12 March 2024; Accepted 12 March 2024; Published 20 March 2024

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope

(2) Discrepancies in the description of the research reported

(3) Discrepancies between the availability of data and the research described

(4) Inappropriate citations

(5) Incoherent, meaningless and/or irrelevant content included in the article

(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] X. Bingxiang, W. Panxing, F. Lu, Y. Xiuyou, and D. Chao, "A Prognostic Model for Brain Glioma Patients Based on 9 Signature Glycolytic Genes," *BioMed Research International*, vol. 2021, Article ID 6680066, 15 pages, 2021.

Hindawi
BioMed Research International
Volume 2021, Article ID 6680066, 15 pages
https://doi.org/10.1155/2021/6680066

Research Article

A Prognostic Model for Brain Glioma Patients Based on 9 Signature Glycolytic Genes

Xiao Bingxiang, Wu Panxing, Feng Lu, Yan Xiuyou, and Ding Chao

Department of Neurosurgery, Taizhou Central Hospital (Taizhou University Hospital), Taizhou 318000, China

Correspondence should be addressed to Xiao Bingxiang; xiaobx9376@tzzxyy.com

Received 20 November 2020; Revised 16 March 2021; Accepted 2 June 2021; Published 17 June 2021

Academic Editor: Parameshachari B D

Objective. To screen glycolytic genes linked to the glioma prognosis and construct the prognostic model. *Methods*. The relevant data of glioma were downloaded from TCGA and GTEx databases. GSEA of glycolysis-related pathways was carried out, and enriched differential genes were extracted. Screening out prognostic-related genes with conspicuous significance and construction of the prognostic model were conducted by multivariate Cox regression analysis and Lasso regression analysis. The model was evaluated, and cBioPortal was used to analyze the mutation of the model gene. The expression of the model gene in tumor and normal colon tissue was analyzed. The model was used to evaluate the prognosis of patients in different groups to verify the applicability of the model. *Results*. 339 differentially glycolytic-related genes were enriched in REACTOME_GLYCOLYSIS, GLYCOLYTIC_PROCESS, HALLMARK_GLYCOLYSIS, and other pathways. We obtained 9 key prognostic genes and constructed the prognostic evaluation model. The 3-year AUC values of the ROC curve display model are greater than 0.75, which indicates that the accuracy of the model is good. The relation of age and risk score to prognosis is shown by univariate and multivariate Cox analysis. The expression of SRD5A3, MDH2, and B3GAT3 genes was significantly upregulated in the tumor tissues, while the HDAC4 and G6PC2 genes were downregulated. The mutation rate of MDH2 and HDAC4 genes was the highest. This model could effectively distinguish the risk of poor prognosis of patients in any age stage. *Conclusion*. The prognostic assessment models based on glycolysis-related nine-gene signature could accurately predict the prognosis of patients with GBM.

1. Introduction

Cerebral glioma, the main lethal tumor in neurosurgery, is the primary malignant tumor of the brain [1, 2]. Surgery, chemotherapy, and radiation therapy are mainly used to treat malignant brain tumor, including glioma [3, 4]. There are also some significant progress in the basic research, and comprehensive therapy of glioma has been made in recent years. However, the prognosis of patients with glioma has not been significantly improved as glioma induces an immune infiltrating environment [5, 6]. In general, poor glioma prognosis shows a high rate of recurrence after treatment [7–9]. Thus, it is important to find novel targets for the treatment and prognosis of patients with glioma.

In normal cells, when the oxygen content is normal, pyruvate enters the tricarboxylic acid cycle; glucose changes to pyruvate and then to lactate in the absence of oxygen. However, glycolysis is the main energy source for the growth of tumor cells [10].

The glucose uptake and intracellular lactic acid accumulation of tumor cells will gradually increase even with normal oxygen content [11]. Tumor cells mainly convert glucose into lactic acid to get ATP by glycolysis [12, 13]. Regulating glycolysis-related genes to affect the activities of glycolysis rate-limiting enzyme and hypoxia-inducible factor can inhibit glycolysis process. Previous clinical studies have discovered that the characteristic dysregulated tumor cell metabolism can be found in a variety of cancers [14, 15]. Regulating the expression of glycolytic genes is expected to become a new method of cancer treatment. At present, studies have explored glycolysis gene targets related to the treatment of glioma and the prognosis of patients [16–18]. However, the accuracy of usage of glycolytic-related genes to predict the prognosis of patients with glioma remains unknown. Therefore, it is critical to analyze how these genes are related to the prognosis of patients with glioma.

In this study, we used gene set enrichment analysis (GSEA) to explore the main signaling pathways of glycolysis-related gene enrichment. We extracted glycolytic-related gene expression data from transcriptome data of glioma in The Cancer Genome Atlas (TCGA) database and mRNA expression data of normal brain tissue in the GTEx database for differential analysis. We established the nine-gene risk model to predict the prognosis of patients by univariate and multivariate Cox regression analysis. The reliability and accuracy of the model were verified by ROC and survival analysis. We found that this risk model can independently identify patients with poor prognosis in the high-risk group. In addition, it was confirmed that the performance of the risk score model is better than that of age, gender, and other clinical indicators in evaluating the prognosis of patients with glioma and it has a good prognostic evaluation effect.

2. Methods

2.1. Data Acquisition and Processing. First, download all data of glioma from the UCSC-Xena (https://xenabrowser.net/datapages/). The Xena-GBM dataset (TCGA http://www.tcga.org) contains 5 normal and 168 cancer samples. Download the data of normal brain tissue from the Genotype-Tissue Expression database (https://gtexportal.org/) as a control. GTEx database contains 115 normal brain tissue samples which are located in the cerebral cortex. By merging, 120 normal and 168 tumor samples were obtained. The combined gene expression profile data and clinical data in Xena were used to train the model of the prognosis of patients. The number of patients ($n = 589$), sex, age, and other clinical information of patients were included in the analysis.

2.2. Gene Set Enrichment Analysis (GSEA) of Related Pathways of Glycolysis. Pathways related to glycolysis (GLYCOLYSIS_PATHWAY, HALLMARK_GLYCOLYSIS, GLYCOLYSIS_GLUCONEOGENESIS, GLYCOLYTIC_PROCESS, and REACTOME_GLYCOLYSIS) were found from the GSEA website (http://www.broadinstitute.org/gsea/index.jsp) [19]. The GSEA was performed in the gene expression data of the training set including normal samples and glioma samples.

2.3. Differential Analysis and Modeling of Glycolytic-Related Genes. Glycolysis-related genes (GRGs) were extracted from the training set based on GSEA results, and differential analysis was performed using limma packets in R (3.61) ($P < 0.05$, logFC ≥ 1 or ≤-1). Prognosis GRGs (FDR < 0.01) were screened by the logarithmic rank test in combination with survival time. The candidate GRGs were analyzed by using Cox risk regression analysis and glmnet in R (3.61) for 10-fold cross-validation. The survival data of Cox analysis was processed by the glmnet package, and the object of survival analysis was identified to construct the Lasso regression model (the best λ value was selected by the cv.glmnet function, and the gene screening was carried out by the coef function). The optimal genes were constructed as a GRG gene pair model. We extracted the relative expression of model genes in samples and plotted the heat map. We evaluated the model's accuracy through the receiver operating characteristic curve (ROC) and distinguished the high- and low-risk groups by cutoff value of the model.

2.4. GRG Model Validation. GRG model was used to analyze the training set in terms of single-factor and multifactor Cox proportional hazard analysis and survival analysis.

2.5. Expression and Mutation of Model Genes. Use the limma package in R (3.61) to analyze the expression of 9 model genes in the training set. Use cBioPortal (http://www.cbioportal.org/) to analyze the mutations of 9 model genes in GBM samples in TCGA database.

2.6. Correction between GRG Model and Clinical Characters. Analyze the relationship between clinically relevant characteristics such as age, gender, and survival rate in the training group. Analyze the survival rate of patients with different clinical characteristics classified according to the model.

2.7. Expression Verification of Prognosis-Related Genes. Verify the expression of selected model genes related to prognosis. Use the Human Protein Atlas (HPA) (http://www.proteinatlas.org/) to validate the expression of model genes related to prognosis in glioma tissues and normal tissues and compare the consistency of previous analysis and expression differences. 50 cases of patients with glioma were screened out from our hospital from June 2018 to June 2020. Inclusion criteria were as follows: (1) pathologically confirmed; (2) new cases diagnosed by this hospital for the first time. Exclusion criteria were as follows: (1) with other malignant tumors; (2) patients who have received radiotherapy, chemotherapy, or other antineoplastic drugs before surgery [20].

Nine differentially expressed genes in tumor and normal tissues of patients with glioma were verified by qPCR. Primers were synthesized according to the PCR primer information provided by the Primer Bank database (Table 1). GAPDH was used as an internal reference, and a two-step method was used. The expression of GAPDH was detected by qPCR. Using the expression level of GAPDH as the standard value "1," the relative expression levels of each differential gene in tumor and normal tissues were calculated. The real-time PCR kit was used to detect the expression of these genes in tumor and normal tissues and to draw statistical charts. The reaction procedure was as follows: hold (predenaturation): 95°C, 30 s, 1 cycle; two-step PCR: 95°C, 5 s, 60°C, 30 s, 40 cycles; dissociation: 95°C, 15 s, 60°C, 30 s, 95°C, 15 s, 1 cycle [20].

2.8. Statistical Analysis. Measured data were expressed as mean ± standard deviation ($x \pm s$), and data were compared using a T-test. The Kaplan-Meier method was used for survival analysis [21]. The receiver operating characteristic curve (ROC curve) and ROC analysis were completed by survival ROC (1.0.3) [20]. A Cox proportional hazard regression model was used to analyze univariately and multivariately. The criterion for statistically significant difference is $P < 0.05$. And $P < 0.01$ indicates the difference has fairly significant statistical significance. A Cox proportional

Table 1: PCR primers for 9 RNAs and internal reference.

Gene	G6PC2	STC1	HDAC4	COG2	SRD5A3
Forward primer	CCCAAATCACTCAA GTCCATGC	CACGAGCTGACTTC AACAGGA	CCTGGGAATGTACG ACGCC	ACAAAGTAAGA CCGCGTATAGC	TGGCTGCACAG CTTACGAAG
Reverse primer	GGTTACCATGACAT ACCAGACAC	GGATGTGCGTT TGATGTGGG	CCCGTCTTTCCTGC GTAAC	AAGCAGTGCCGTAT TATATCGAC	TCAGCACAGTT AGGCCAACAA
Gene	MDH2	IL13RA1	TGFBI	B3GAT3	GAPDH
Forward primer	TCGGCCCAGAACAA TGCTAAA	GTCCCAGTGTA GCACCAATGA	CACTCTCAAACCTT TACGAGACC	AAGGAGTCGTCTAC TTTGCTGA	ACAACTTTGGT ATCGTGGAAGG
Reverse primer	GCGGCTTTGGTCTC GATGT	GCTCAGGTTGTGCC AAATGC	CGTTGCTAGGG GCGAAGATG	GGGCATTGGGCTTA TCTAACAG	GCCATCACGCC ACAGTTTC

regression model was used to identify the predictive model with the best explanatory and informative efficacy. A risk score staging model was established using the R package survival function coxph (). The risk score formula is described as follows [22]:

$$\begin{aligned} &\text{Risk score } 1/4 \text{ expression of gene } 1 \times \beta 1 \\ &\text{b expression of gene } 2 \times \beta 2 \\ &\text{b} \cdots \text{b expressiom of gene } n \times \beta n : \end{aligned} \tag{1}$$

The R package was used for analyzing the relationship between different clinical characteristics and survival rate in high- and low-risk groups. All statistical analyses were performed with the Statistical Package for the Social Sciences software version 16.0 (SPSS Inc., Chicago, IL, the USA) and GraphPad Prism 7 (GraphPad Software, La Jolla, CA, the USA; http://www.graphpad.com) [23].

3. Results

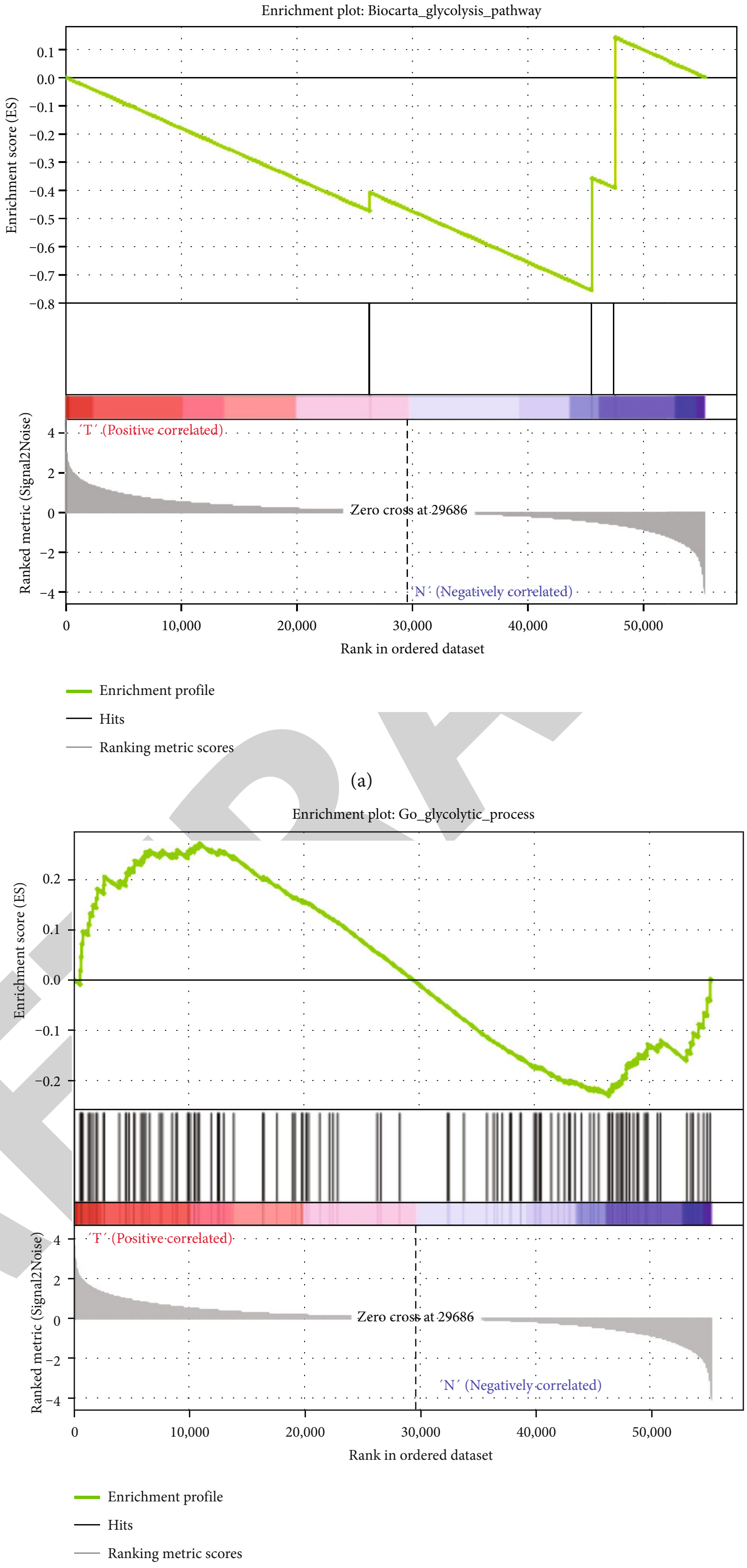
3.1. GSEA of Glycolysis-Related Pathways. The mRNA expression data and clinical information of 598 patients were obtained from TCGA. Glycolysis GSEA was performed on the GBM sample data in the training set. Results showed that the training set genes were significantly enriched in REACTOME_GLYCOLYSIS, GLYCOLYTIC_PROCESS, and HALLMARK_GLYCOLYSIS at normalized P value < 1% (Figures 1(a)–1(e)) ($P < 0.05$).

3.2. Model Construction of Glycolysis-Related Genes. Glycolysis-related genes were obtained, and these genes were extracted from the training set for differential analysis, and the results showed that there were 339 differential glycolysis-related genes. 19 genes were significantly correlated to OS ($P < 0.05$) and were entered into a stepwise multivariate Cox regression analysis (Supplement Table 1) (results of multivariate Cox regression analysis of 19 genes which significantly correlated to overall survival). Combined with the clinical survival time, through multivariable Cox regression analysis and 10-fold cross-validation, we obtained the 9 optimal glycolysis-related genes, which are G6PC2, STC1, HDAC4, COG2, SRD5A3, MDH2, IL13RA1, TGFBI, and B3GAT3 (Table 2), and we constructed a prognostic model. According to the risk score formula, patients with GBM were divided into a high-risk group ($n = 80$) and a low-risk group ($n = 80$) with the median risk score as cutoff value (Figure 2(a)). Each patient's survival (day) is shown in Figure 2(b). The patients in the high-risk score group had a higher mortality rate than those in the low-risk score group (Figure 2(c)). Figure 2(d) shows that the 5-year AUC value of the ROC curve of the model was as high as 0.763, indicating that our model had high accuracy. The heat map (Figure 2(e)) shows the expression profiles of these 9 mRNAs. With the increase of risk score in GBM patients, the expression of the risky-type mRNAs (STC1, SRD5A3, MDH2, IL13RA1, TGFBI, and B3GAT3) was all upregulated gradually, while the expression of the protective-type mRNA (HDAC4, COG2, and G6PC2) was downregulated.

3.3. Expression and Mutation of Model Genes. Use cBioPortal to analyze the mutations of model genes, and the results showed that the mutation rate of HDAC4 and MDH2 gene was the highest of 8%, while the mutation rates of G6PC2 were lowest as 2.8% (Figure 3). Analyze the expression of 9 model genes in the training set; then, the results showed that these 6 genes, STC1, SRD5A3, MDH2, IL13RA1, TGFBI, and B3GAT3, were all highly expressed, while HDAC4, COG2, and G6PC2 genes were all downregulated in patients with GBM (Figure 4).

3.4. Validation of Model. The model was applied to the training set data for verification. It was shown that risk scores were significantly related to the prognosis (Figure 5(a)) in the univariate risk regression analysis. Multivariate risk regression analysis showed that risk scores could be used as significant independent prognostic factors (Figure 5(b)). The results suggested that the risk score was effective in predicting the prognosis of patients with GBM (Table 3).

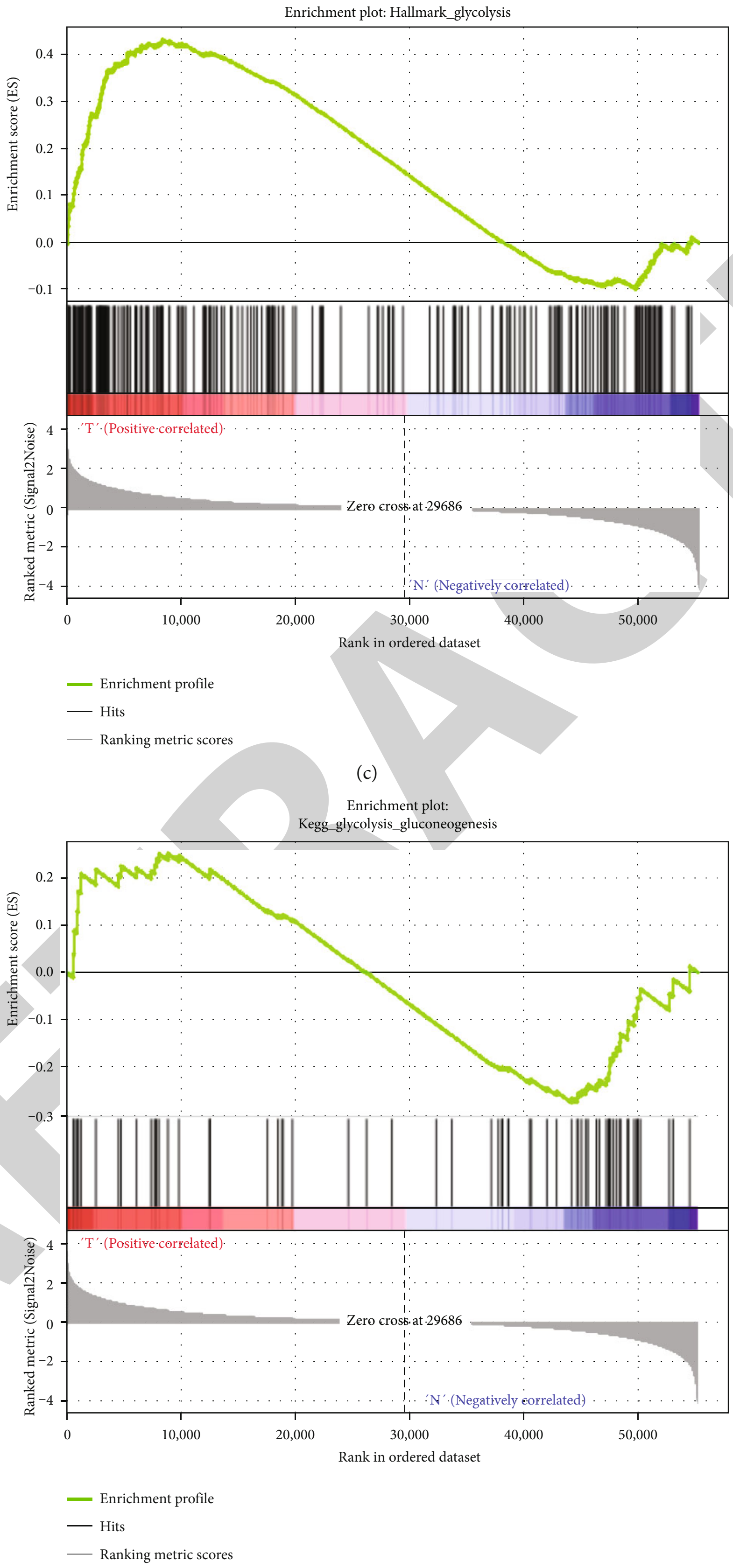
3.5. Model and Clinical Characters. After analyzing the relation between clinical traits and survival, we found that only age and risk scores were significantly related to the survival rate of patients (Figures 5(a) and 5(b)). Group the clinical traits according to the model, and analyze the survival rate of patients in different groups. We can see that the GRG model can well distinguish the older than 65-year-old group, male group, and age < 65-year-old groups of patients ($P < 0.001$), while the difference was not so obvious in female subgroups ($P = 0.016$) (Figures 5(c)–5(f)).



(a)

(b)

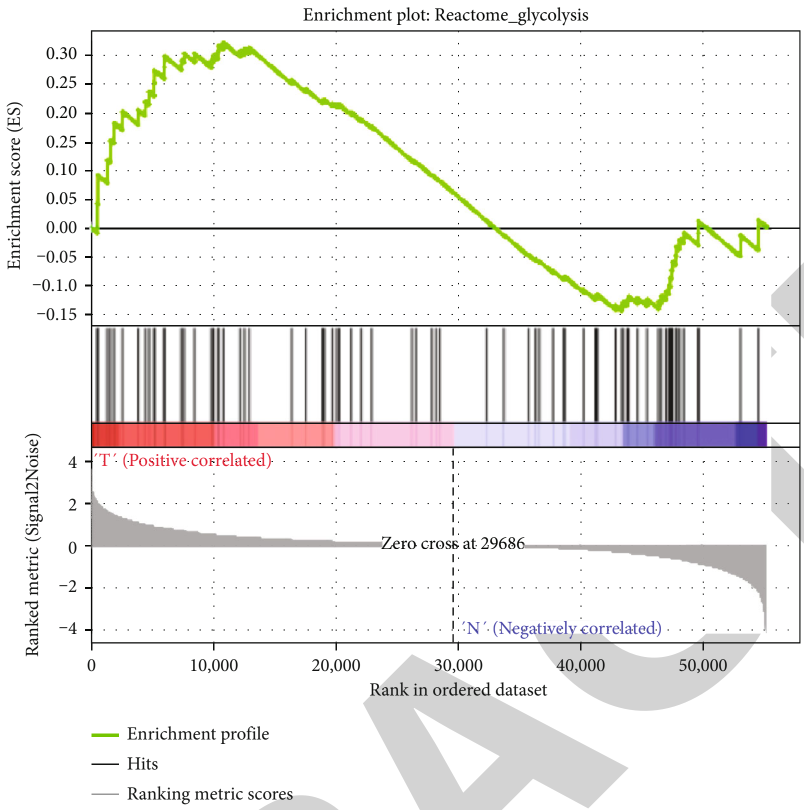
Figure 1: Continued.



(c)

(d)

Figure 1: Continued.



(e)

Figure 1: GSEA results for the enrichment plots of five gene sets (BIOCARTA_GLYCOLYSIS_PATHWAY, GLYCOLYTIC, HALLMARK_GLYCOLYSIS, GLYCOLYSIS_GLUCONEOGENESIS, and REACTOME_GLYCOLYSIS) that were significantly differentiated in normal and GBM tissues based on TCGA. GSEA: gene set enrichment analysis; GBM: glioblastoma multiforme; TCGA: The Cancer Genome Atlas.

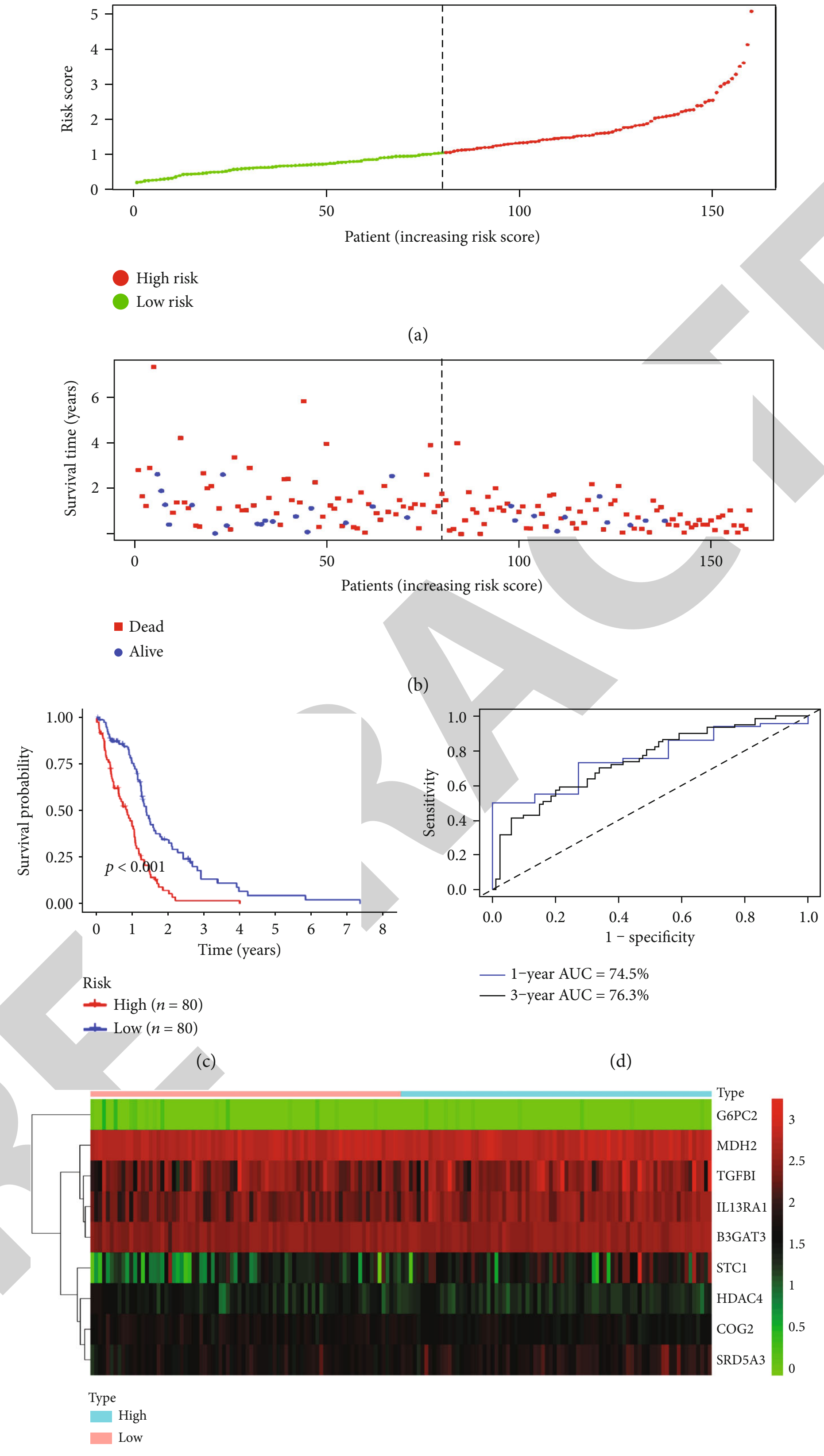
Table 2: Nine prognostic mRNAs significantly associated with overall survival in patients with GBM.

Gene ID	COEF	HR
G6PC2	-3.188458997	0.041235366
STC1	0.238492072	1.269333643
HDAC4	-0.545874088	0.579335172
COG2	-0.628281634	0.533507776
SRD5A3	0.381533075	1.4645281
MDH2	0.760142992	2.138581998
IL13RA1	0.351136696	1.420681514
TGFBI	-0.199048218	0.819510378
B3GAT3	0.57572986	1.778428057

3.6. Immunohistochemical and qPCR Verification of Prognostic Genes. The data verification results of the HPA database indicated that the expression of IL13RA1 and COG2 in cancer and adjacent tissues had not been detected in the database, and the expression of the remaining 7 genes in cancer and adjacent tissues could be verified. Among them, STC1 and TGFBI genes were not significantly expressed in tumor and normal tissues, and there was no significant difference in expression. Compared with normal tissues, the expressions of SRD5A3, MDH2, and B3GAT3 in tumor tissues were significantly upregulated, and the expression of HDAC4 and G6PC2 in tumor tissues was significantly downregulated; the verification results were basically consistent with the research analysis results (Figure 6(a)). Figure 6(b) shows the real-time quantitative PCR results of differentially expressed genes. The relative expression of each gene in the figure was calculated according to the relative expression quantity value of the internal reference gene (GAPDH). STC1, SRD5A3, MDH2, IL13RA1, TGFBI, and B3GAT3 were all upregulated in tumor tissues, while the expression of the HDAC4, COG2, and G6PC2 was downregulated. The experimental results are basically consistent with the analytical results.

4. Discussion

These nine biomarker genes (STC1, SRD5A3, MDH2, IL13RA1, TGFBI, B3GAT3, HDAC4, COG2, and G6PC2) were screened by the model; STC1 was associated with the occurrence and development of various cancers [24–26]. Chen et al. had found that STC1 was related to the prognosis of colon adenocarcinoma [24]. Kamata et al. found that fibroblast-derived STC-1 could modulate tumor-associated macrophages and lung adenocarcinoma development [25]. Zhao et al. provided an overview of (a) the possible



(a)

(b)

(c)

(d)

(e)

Figure 2: The nine-mRNA signature related to risk score predicts the overall survival of patients with glioblastoma multiforme. (a) mRNA risk score distribution. (b) Survival status. (c) Survival curves of patients in high-risk and low-risk groups. (d) The receiver operating characteristic curve (ROC) of 1 year and 3 years of the model. (e) Heat map of nine gene expression profile in The Cancer Genome Atlas.

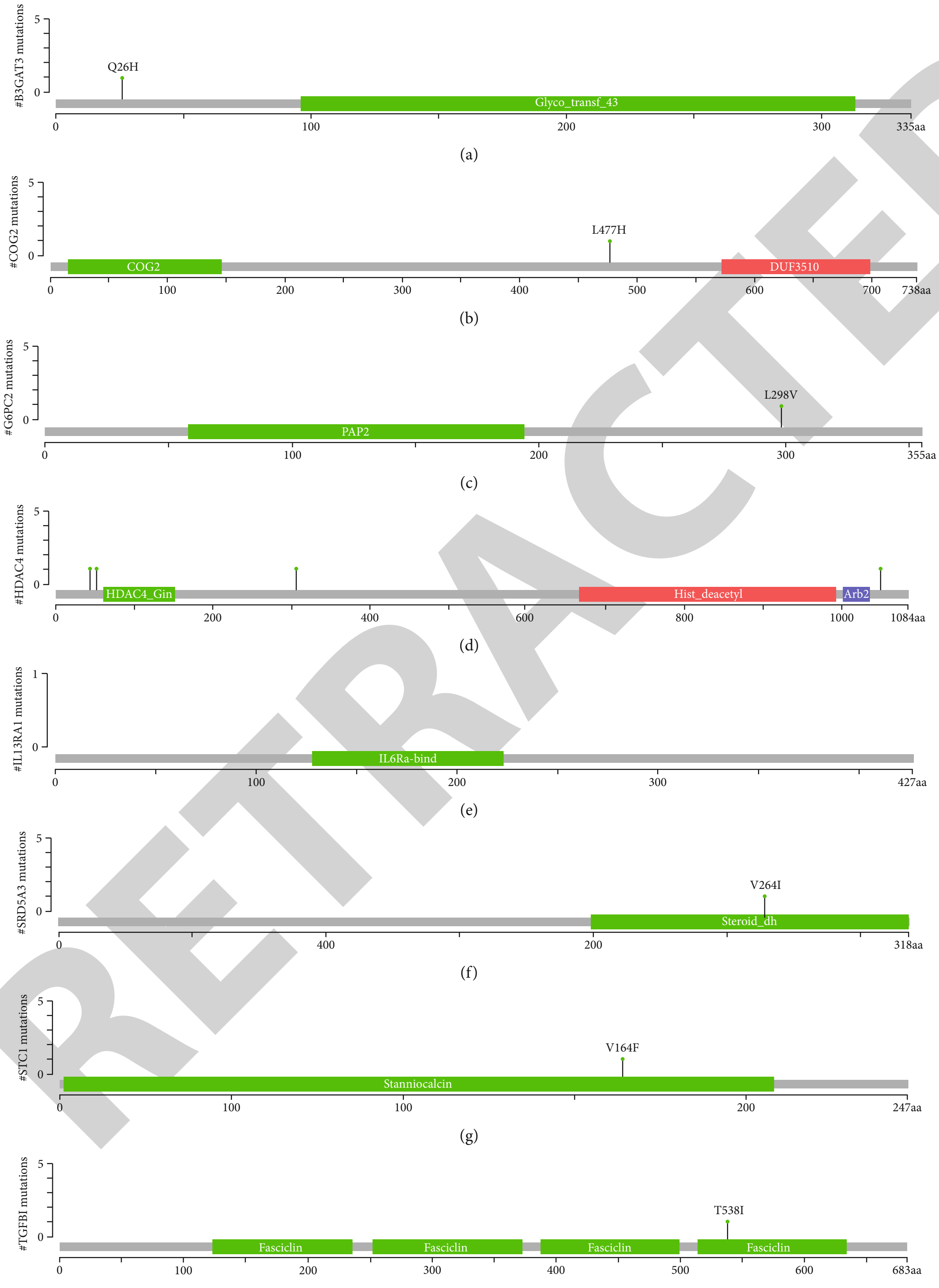
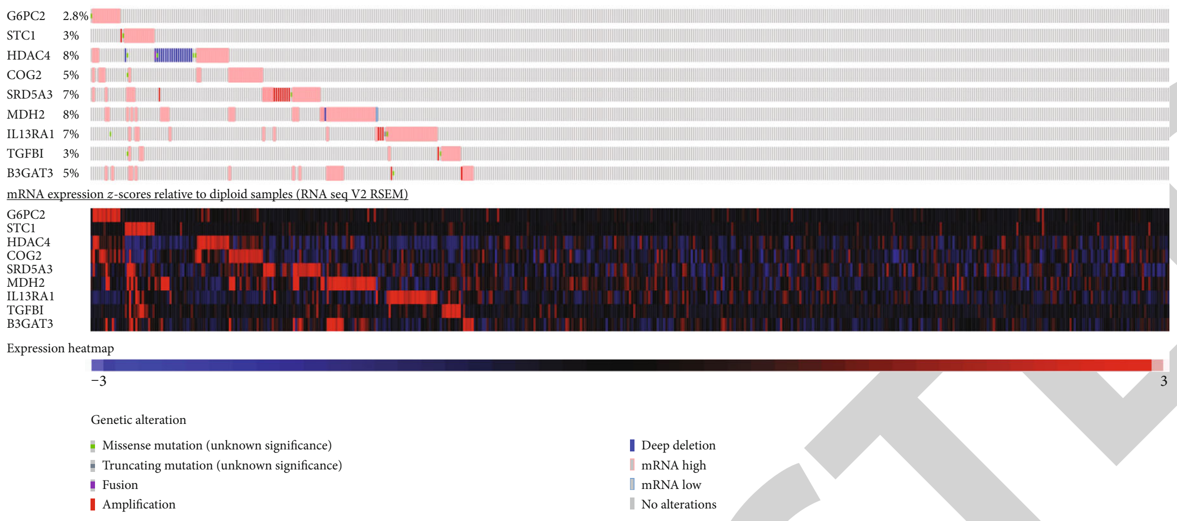


Figure 3: Continued.



(i)

Figure 3: The alteration proportion for the nine selected genes in clinical samples of glioblastoma multiforme in the cBioPortal database.

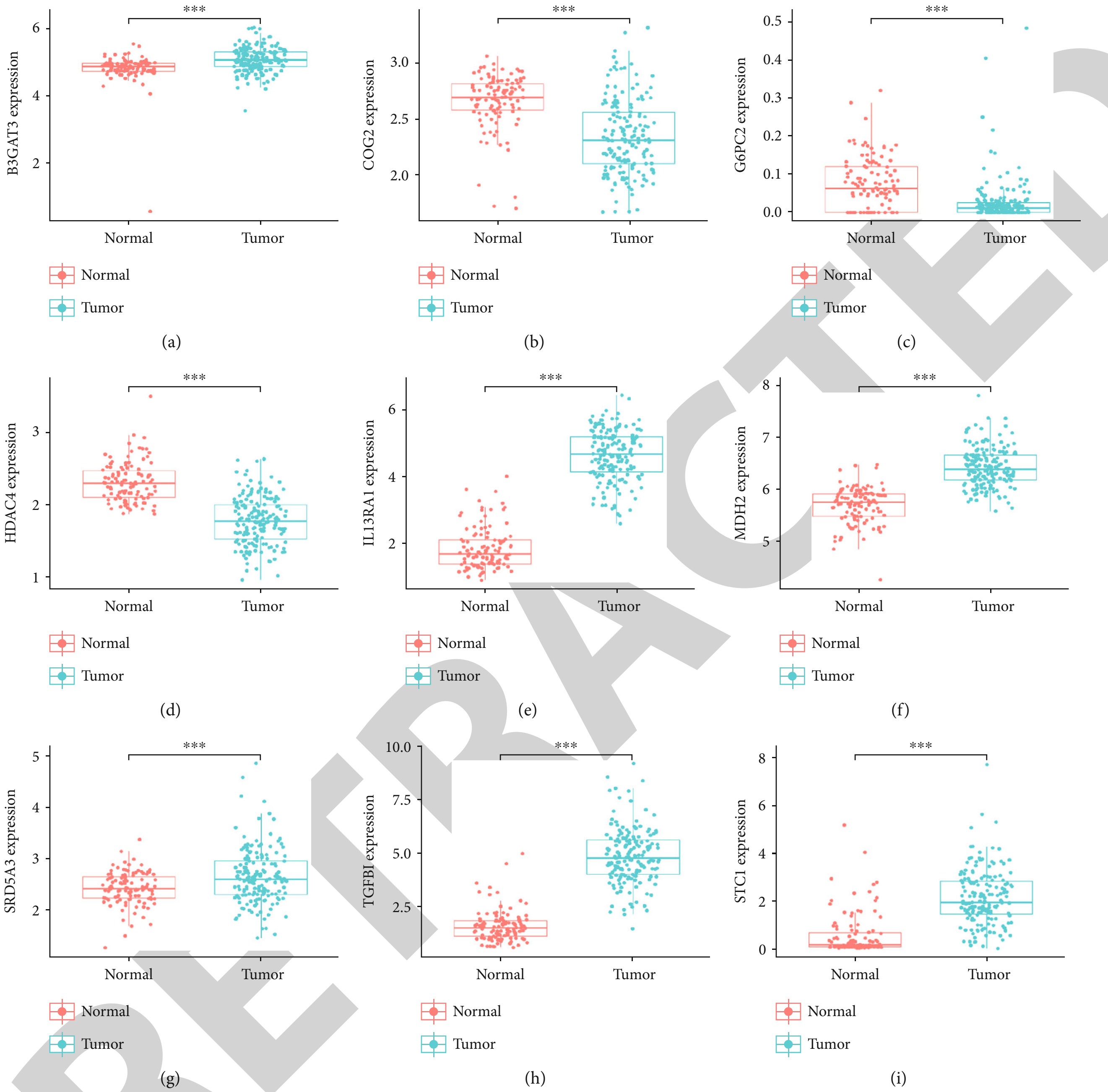
mechanisms through which STC1 affected the malignant properties of cancer in their article [26].

SRD5A3 was reported as one of the six genes associated with P4 metabolism in the liver [27]. In breast cancer, SRD5A3 was decreased significantly and primarily enriched in the hormone metabolic process [28]. Typically, the MDH2 gene was considered to play a key role in glycolysis and fatty acid metabolism [29]. The activity of the MDH2 gene was different in prostate cancer and benign cell lines at the basal level [30]. Shelar et al. found that L2HGDH suppressed both cell migration and tumor growth and these effects were mediated by the activity of malate dehydrogenase 2 (MDH2) [31]. Studies have found that the IL-13 and IL13RA1 interaction promoted cancer cell growth and metastasis, and IL13RA1 expressing in tumor cells was related to poor prognosis in patients with invasive breast cancer [32]. IL13RA1 had been previously reported to be associated with glioblastoma and was associated with multiple survival events [33, 34].

TGFBI had been found as an index of CAF abundance, which was an effective indicator of the survival of patients in various cancers [35]. Du et al. found that TGFBI related to prognosis of patients with ccRCC may become the novel prognostic biomarkers and immunotherapy targets [36]. The study had also found that tumor-associated macrophages could promote ovarian cancer cell migration by secreting transforming growth factor beta induced (TGFBI) and tenascin C [37]. High expression of B3GAT3 was related to poor prognosis of liver cancer [38]. It had been reported that HDAC4 participated in the occurrence and development of various cancers [39–41]. Low levels of AMPK could promote epithelial-mesenchymal transition in lung cancer primarily through HDAC4- and HDAC5-mediated metabolic reprogramming [39]. The knockdown of S6K1 was predicted to reduce the tumorigenicity of HCC through the regulation of hubs of genes including HDAC4 [40]. RGMB-AS1 long noncoding RNA could act as a microRNA-574 sponge thereby enhancing the aggressiveness of gastric cancer via HDAC4 upregulation [41]. Jung et al. reported that genetically elevated G6PC2 was associated with reduced risk for breast cancer in phenotype-specific analysis [42]. There was no report about the study of COG2 in any cancer so far. And COG2 may be regarded as a novel biomarker for the prognosis of patients with GBM.

GSEA is a gene set enrichment analysis that integrates data from different levels and sources. In this study, we used GSEA to analyze the mRNA expression data of 598 patients with GBM and found that five functions had significant differences. According to the NES, *N*, and *P* value, the GLYCOLYSIS with the minimum *P* value was selected for further analysis. We focused on selecting GSEA genes to predict specific functions of patient survival and explored these genes extensively. By analyzing the enrichment of the expression profile of GBM patients in glycolysis-related pathways, and using Cox regression analysis, we successfully screened glycolysis-related genes that are closely related to the survival of colon cancer patients and constructed a prognostic model. ROC analysis proved that this model had a high accuracy rate and could distinguish patients with GBM very well. By univariate and multivariate Cox regression analyses [10], 9 gene combinations rather than a single gene combination were determined to be valuable for the prognosis of patients with GBM. Compared with some known prognostic biomarkers, this selected risk marker may be targeted and more powerful prognostic in supporting clinical outcomes acting as an effective classification tool for patients with GBM.

In recent years, researchers tried to use bioinformatics methods to analyze sequencing results to detect biomarkers related to survival in glioma patients and predict their prognosis [43–46]. Jiang et al. identified genes related to low-grade glioma progression and prognosis based on integrated transcriptome analysis [44]. Liu et al. used lncRNA expression profiles to predict the prognosis of patients with



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

(i)

Figure 4: Different expressions of nine genes in the normal tissues and tumor tissues based on Genotype-Tissue Expression and The Cancer Genome Atlas database. ($***$ represents $P < 0.0001$).

glioblastoma [45]. There are also some researches focusing on the relationship between glycolysis and tumor oncogenesis, development, proliferation, and invasion [47, 48]. Most studies have focused on the relationship between glycolysis and tumor oncogenesis, development, proliferation, and invasion [49]. However, no study has investigated the relationship between glycolysis-related genes and the survival of patients with GBM. Our study first used the public TCGA database to identify and comprehensively analyze glycolysis-related mRNAs that are significantly associated with the prognosis of patients with GBM.

Although the model with nine-gene signature can be used to predict the prognosis of patients with GBM, some limitations still remain. The biological functions of the predicted genes were annotated using computational methods, and additional experiments should be performed to further reveal the mechanisms by which genes are involved in GBM tumorigenesis. The risk score model was established based on TCGA database and should be validated in other cohorts in future studies. We planned to supplement the following experiment: collect tumor samples from patients with glioma in stages I and III, and use qPCR and immunohistochemistry

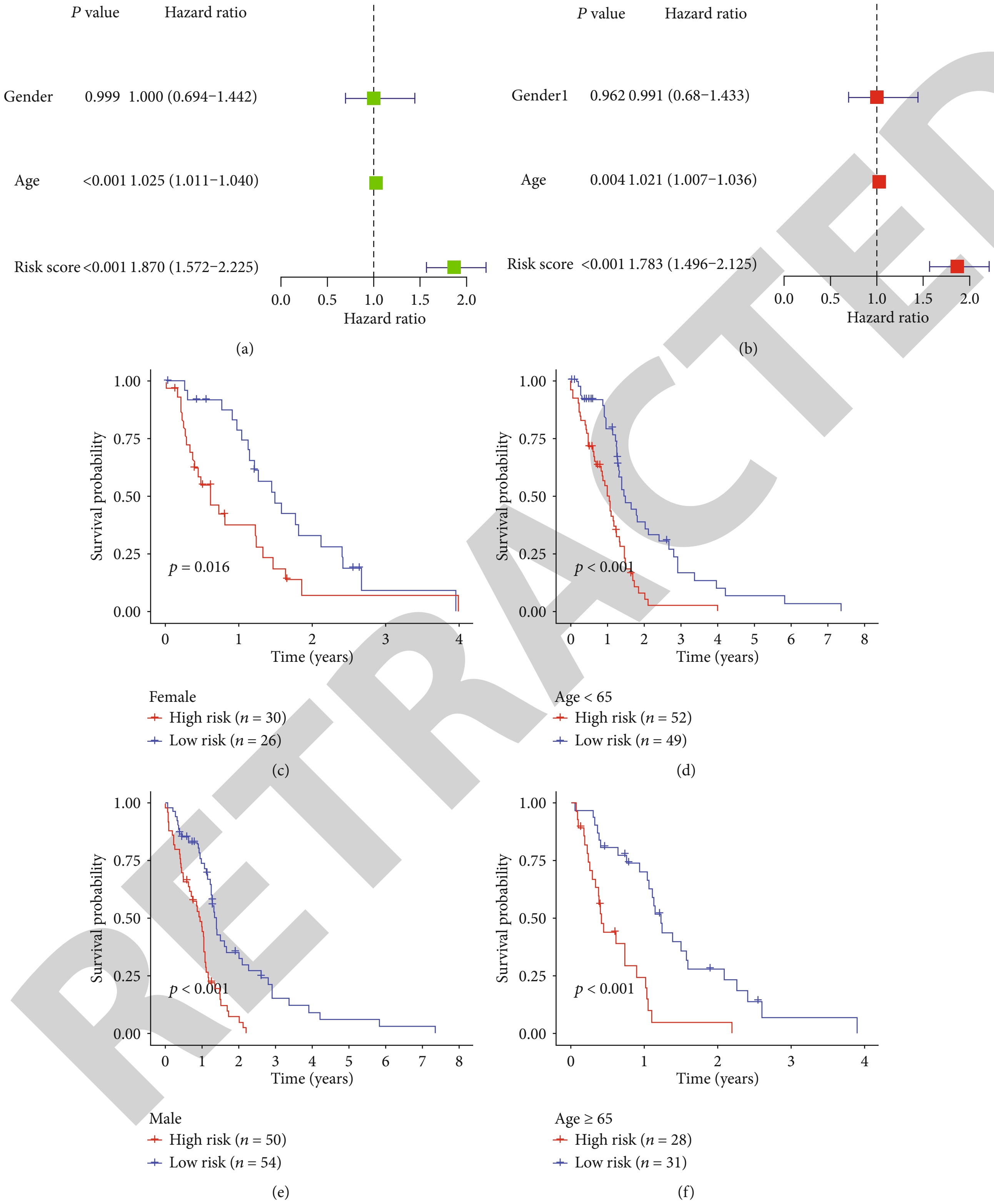


Figure 5: (a) Univariate analysis of gender, age, and risk score. (b) Multivariate analysis of gender, age, and risk score. (c) Kaplan-Meier survival curves of the female patient group. (d) Kaplan-Meier survival curves of the male patient group. (e) Kaplan-Meier survival curves of the age < 65 patient group. (f) Kaplan-Meier survival curves of the age ≧ 65 patient group.

Table 3: Univariable and multivariable analyses for each clinical feature.

Clinical feature	Number	Univariate analysis				Multivariate analysis			
		HR	HR.95L	HR.95H	P value	HR	HR.95L	HR.95H	P value
Risk score (low-risk/high-risk)	80/80	1.000184915	0.693882598	1.441699025	0.999209193	0.991023245	0.685447609	1.432825879	0.961764674
Sex (male/female)	104/56	1.025241853	1.011152205	1.03952783	0.000414349	1.02118673	1.006577098	1.036008408	0.004349668
Age (≤65/>65)	101/59	1.8701817	1.571798452	2.2252087	$1.67E-12$	1.782826112	1.495596894	2.125217669	$1.11E-10$

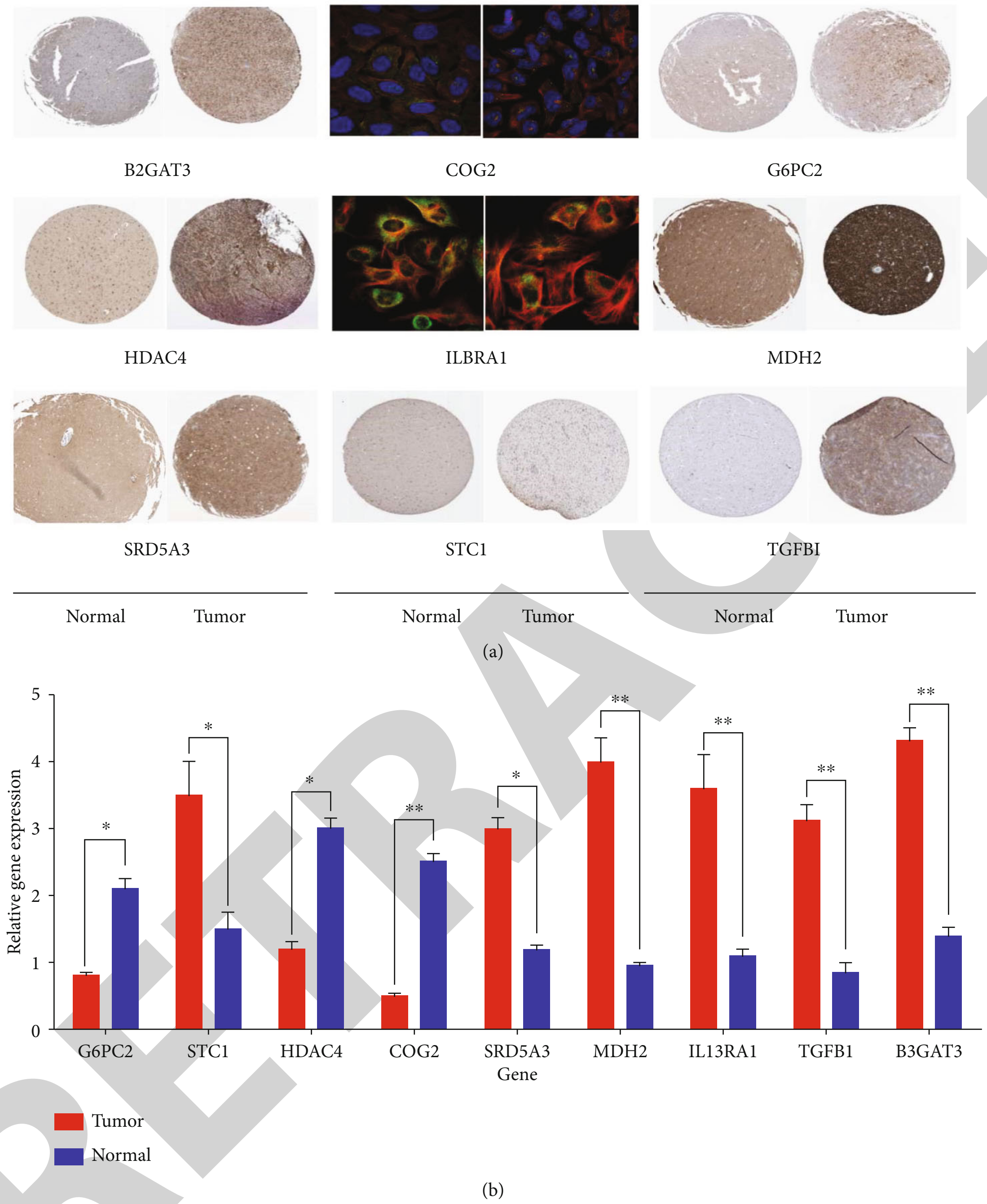


Figure 6: (a) Immunohistochemistry validated the expression of 9 prognostic-related glycolytic genes in glioma tissues and normal tissues. (b) qPCR experiments validated the expression of 9 prognostic-related glycolytic genes in glioma tissues and normal tissues.

to detect the expression differences of nine genes in tumor samples with different clinical stages (significant prognostic differences). In addition, although the gene signatures may be most effective in the early stages, their prognostic role in early GBM needs to be further evaluated.

5. Conclusion

In this study, we identified nine glycolysis-related genes associated with survival in patients with GBM using multivariate Cox regression analysis and Lasso regression analysis. The results of analysis revealed that prognostic assessment models based on nine glycolytic-related genes could accurately predict the prognosis of patients with GBM.

Data Availability

All data are available. The data in this paper are from TCGA (http://www.tcga.org) and GTEx (https://gtexportal.org/) database. Please contact us to access if it is needed.

Conflicts of Interest

There are no conflicts of interest in this study.

Authors' Contributions

Wu Pan-Xing and Feng Lu contributed to research design and drafting the manuscript. Yan Xiuyou and Ding Chao performed literature search. Xiao Bingxiang reviewed and revised the manuscript and writing guidance.

Acknowledgments

This study was funded by Taizhou Science and Technology Plan Project (XM20190667).

Supplementary Materials

Results of multivariate Cox regression analysis of 19 survival-related genes. *(Supplementary Materials)*

References

[1] L. A. Albuquerque, J. P. Almeida, L. J. de Macêdo Filho, A. F. Joaquim, and H. Duffau, "Extent of resection in diffuse low-grade gliomas and the role of tumor molecular signature-a systematic review of the literature," *Neurosurgical Review*, vol. 44, no. 3, pp. 1371–1389, 2021.

[2] R. J. Packer and T. J. MacDonald, "Integrated analysis of pediatric low-grade glioma: clinical implications and the path forward," *Neuro-Oncology*, vol. 22, no. 10, pp. 1413-1414, 2020.

[3] X. Gao, Y. Yang, J. Wang et al., "Inhibition of mitochondria NADH-ubiquinone oxidoreductase (complex I) sensitizes the radioresistant glioma U87MG cells to radiation," *Biomedicine & Pharmacotherapy*, vol. 129, p. 110460, 2020.

[4] P. M. Grazia, R. Antonio, C. Antonella, C. Antonio, F. Savina, and S. Antonio, "Diffuse intrinsic pontine glioma (DIPG): breakthrough and clinical perspective," *Current Medicinal Chemistry*, vol. 27, 2020.

[5] X. Zhang, L. Can, X. Lifei et al., "Predicting individual prognosis and grade of patients with glioma based on preoperative eosinophil and neutrophil-to-lymphocyte ratio," *Cancer Management and Research*, vol. Volume 12, pp. 5793–5802, 2020.

[6] Z. Hao, F. Fan, Y. Yu et al., "Clinical characterization, genetic profiling, and immune infiltration of TOX in diffuse gliomas," *Journal of Translational Medicine*, vol. 18, no. 1, p. 305, 2020.

[7] H. Runzhi, L. Zhenyu, X. Zhu et al., "Collagen type III alpha 1 chain regulated by GATA-binding protein 6 affects type II IFN response and propanoate metabolism in the recurrence of lower grade glioma," *Journal of Cellular and Molecular Medicine*, vol. 24, no. 18, pp. 10803–10815, 2020.

[8] H. Wei, S. Jia, C. Jiachao, B. Dong, and G. Wei, "Emerging roles and therapeutic interventions of aerobic glycolysis in glioma," *Oncotargets and Therapy*, vol. Volume 13, pp. 6937–6955, 2020.

[9] T. C. Hollon, P. Balaji, U. Esteban et al., "Rapid, label-free detection of diffuse glioma recurrence using intraoperative stimulated Raman histology and deep neural networks," *Neuro-Oncology*, vol. 23, no. 1, pp. 144–155, 2021.

[10] J. Liu, S. Li, G. Feng et al., "Nine glycolysis-related gene signature predicting the survival of patients with endometrial adenocarcinoma," *Cancer Cell International*, vol. 20, no. 1, p. 183, 2020.

[11] L. Yu, M. Lu, D. Jia et al., "Modeling the genetic regulation of cancer metabolism: interplay between glycolysis and oxidative phosphorylation," *Cancer Research*, vol. 77, no. 7, pp. 1564–1574, 2017.

[12] Z. Abbaszadeh, S. Çeşmeli, and A. Ç. Biray, "Crucial players in glycolysis: cancer progress," *Gene*, vol. 726, p. 144158, 2020.

[13] A. V. Orang, J. Petersen, R. A. McKinnon, and M. Z. Michael, "Micromanaging aerobic respiration and glycolysis in cancer cells," *Molecular metabolism*, vol. 23, pp. 98–126, 2019.

[14] Z. Wu, J. Wu, Q. Zhao, S. Fu, and J. Jin, "Emerging roles of aerobic glycolysis in breast cancer," *Clinical & Translational Oncology*, vol. 22, no. 5, pp. 631–646, 2020.

[15] M. G. Vander Heiden, "Targeting cancer metabolism: a therapeutic window opens," *Nature Reviews. Drug Discovery*, vol. 10, no. 9, pp. 671–684, 2011.

[16] P. Du, Y. Liao, H. Zhao, J. Zhang, and K. Mu, "ANXA2P2/-miR-9/LDHA axis regulates Warburg effect and affects glioblastoma proliferation and apoptosis," *Cellular Signalling*, vol. 74, p. 109718, 2020.

[17] Z. Xinyu, S. Wang, G. Lin, and D. Wang, "Down-regulation of circ-PTN suppresses cell proliferation, invasion and glycolysis in glioma by regulating miR-432-5p/RAB10 axis," *Neuroscience Letters*, vol. 735, p. 135153, 2020.

[18] J. Lu, L. Xiaobai, J. Zheng et al., "Lin28A promotes IRF6-regulated aerobic glycolysis in glioma cells by stabilizing SNHG14," *Cell Death & Disease*, vol. 11, no. 6, p. 447, 2020.

[19] W. Cheng, M. Li, J. Cai et al., "HDAC4, a prognostic and chromosomal instability marker, refines the predictive value of MGMT promoter methylation," *Journal of Neuro-Oncology*, vol. 122, no. 2, pp. 303–312, 2015.

[20] J. Xu, K. Liao, Z. Fu, and Z. Xiong, "Screening differentially expressed genes of pancreatic cancer between Mongolian and Han people using bioinformatics technology," *BMC Cancer*, vol. 20, no. 1, p. 298, 2020.

[21] X. Ren, X. Cui, S. Lin et al., "Co-deletion of chromosome 1p/19q and IDH1/2 mutation in glioma subsets of brain tumors in Chinese patients," *PLoS One*, vol. 7, no. 3, article e32764, 2012.

[22] C. Wu, X. Cai, J. Yan, A. Deng, Y. Cao, and X. Zhu, "Identification of novel glycolysis-related gene signatures associated with prognosis of patients with clear cell renal cell carcinoma based on TCGA," *Frontiers in Genetics*, vol. 11, p. 589663, 2020.

[23] L. Zhang, J. Pan, Z. Wang, C. Yang, and J. Huang, "Construction of a microRNA-based nomogram for prediction of lung metastasis in breast cancer patients," *Frontiers in Genetics*, vol. 11, p. 580138, 2021.

[24] C. Sihan, G. D. Cao, W. Wei et al., "Prediction and identification of immune genes related to the prognosis of patients with colon adenocarcinoma and its mechanisms," *World Journal of Surgical Oncology*, vol. 18, no. 1, p. 146, 2020.

[25] K. Tamihiro, T. Y. So, A. Qasim et al., "Fibroblast-derived STC-1 modulates tumor-associated macrophages and lung adenocarcinoma development," *Cell Reports*, vol. 31, no. 12, p. 107802, 2020.

[26] Z. Fangyu, G. Yang, F. Mengyu et al., "Expression, function and clinical application of stanniocalcin-1 in cancer," *Journal of Cellular and Molecular Medicine*, vol. 24, no. 14, pp. 7686–7696, 2020.

Hindawi
BioMed Research International
Volume 2021, Article ID 5546572, 12 pages
https://doi.org/10.1155/2021/5546572

Research Article

Cloud-Based Fusion of Residual Exploitation-Based Convolutional Neural Network Models for Image Tampering Detection in Bioinformatics

Amit Doegar,[1] **Srinidhi Hiriyannaiah**,[2] **G. M. Siddesh**,[3] **K. G. Srinivasa**,[1] **and Maitreyee Dutta**[1]

[1] *Department of Computer Science and Engineering, NITTTR, Chandigarh, India*
[2] *Department of Computer Science and Engineering, Ramaiah Institute of Technology, Bangalore, India*
[3] *Department of Information Science and Engineering, Ramaiah Institute of Technology, Bangalore, India*

Correspondence should be addressed to Amit Doegar; amit@nitttrchd.ac.in

Received 3 February 2021; Revised 9 March 2021; Accepted 31 March 2021; Published 12 April 2021

Academic Editor: B. D. Parameshachari

Cloud computing has evolved in various application areas such as medical imaging and bioinformatics. It raises the issues of privacy and tampering in the images especially related to the medical field and bioinformatics for various reasons. The digital images are quite vulnerable to be tampered by the interceptors. The credibility of individuals can transform through falsified information in the images. Image tampering detection is an approach to identifying and finding the tampered components in the image. For the efficient detection of image tampering, the sufficient number of features are required which can be achieved by a deep learning architecture-based models without manual feature extraction of functions. In this research work, we have presented and implemented a cloud-based residual exploitation-based deep learning architectures to detect whether or not an image is being tampered. The proposed approach is implemented on the publicly available benchmark MICC-F220 dataset with the k -fold cross-validation approach to avoid the overfitting problem and to evaluate the performance metrics.

1. Introduction

Computer vision and applications are the upcoming research areas in the field of computing and bioinformatics. The applications include object detection, video analytics, image segmentation, and image pixel analysis. Image forensics and forgery detection are the important applications that can be considered for computer vision applications. It has become easy for the online players to digitally tamper and alter the different dimensions of the images. The detection of forgery in the images is one of the key challenges in computer vision applications such as video surveillance. The different types of tampering include copy-move forgery and splicing. The recent research works on image tampering detection focus on the splicing [1] and copy-move methods [2–4]. The image tampering can be done in many ways and not restricted only copy-move and splicing methods. In some recent works, the residual methods are used for image tampering detection [5–7]. In this paper, the problem of image tampering detection with residual exploitation is dealt with the Convolutional Neural Network (CNN) models and fusion-based approach.

Recent advances in the machine learning (ML) and deep learning (DL) technologies are used for the tampering detection of the images. The underlying relationships are identified by the ML algorithms in analysing the data and making decisions. Speech and vision problems use the ML algorithms to understand and evaluate the different parameters for the speech and vision data [8]. However, these algorithms were limited to the capability of prediction of smaller amounts of data. DL techniques with CNN became more popular for solving the different challenging problems in speech and vision [9]. These techniques are used for image segmentation, classification, detection, image context, and retrieval-related tasks [10, 11]. The optimal solutions to the computer vision problems using DL have paved the way for solving different types of problems in image tampering detection as well. In this paper, the residual nature of the CNN models is exploited for the image tampering detection.

CNN has given researchers to provide an insight into the image tampering detection using the feature maps provided at each layer. It was used to detect whether the image is tampered or not initially, but not to locate the tampered regions. However, there are some attempts to locate the tampered regions using the CNN but are not accurate [12]. A nonoverlapping image patch method was used for the image segmentation in [13]. However, when the image size is small, it fails to identify the tampering. The contextual information of the image is lost and leads to incorrect prediction because they use the image patch as a part of the input to the network. Once the CNN goes deeper, there will also be gradient degradation problem and weak discrimination of features as well. The weak discrimination of the features leads to the incorrect prediction. In this paper, the traditional extraction method of the image patch for image tampering detection is replaced with a fusion-based model based on the residual nets for optimal image tampering detection.

In this paper, image forgery detection is carried out through a novel decision method by residual exploitation-based deep learning models. The proposed approach consists of three phases on the pretrained and fine-tuned spatial exploitation-based CNN models, namely, ResNet-18, ResNet-50, and ResNet-101 [5]. In the first phase, a system to extract features using residual exploitation-based CNN models in the second-phase machine learning-based classifier is deployed on the extracted features, and in the final phase, the fusion of decision outcomes based on these residual exploitation-based CNN models is done to evaluate the accuracy of the model.

The main contributions of this paper are as follows:

(i) A decision fusion-based system is proposed using the CNN-based approach for image tampering detection. The residual exploitation-based CNN models used for the fusion decision are ResNet-18, ResNet-50, and ResNet-101

(ii) The fusion decision system is implemented in two phases. First, the pretrained weights for the residual exploitation-based CNN models are used to evaluate the tampering of the images. Second, the fine-tuned weights are used to compare the results of the tampering of the images with the pretrained model

(iii) The utilization of the residual exploitation-based CNN models leads to the reduction of the number of false matches, thereby reducing the false-positive rate and ultimately increasing the accuracy of the approach

The paper is further organized as follows. In Section 2, the related work is discussed on the image tampering detection methods and the CNN methods with spatial exploitation that are used for image tampering detection. In Section 3, the fusion model using the residual exploitation-based CNN models is proposed, and it follows the regularization applied on the fusion model in Section 4. The experiments and results are discussed in Section 5 followed by the conclusion.

2. Related Work

The different areas of research identified in the image tampering detection domain are resampling detection, JPEG artefacts, detection of copy-move operations, splicing, and object removal [14]. Digital content has evolved over a period of time with the advances in the computer graphics, internet, and digital contents. The advances are utilized for many applications in computer vision and image recognition applications [15]. However, the downside of the exploitation of these applications lies in the fact of analysis of creating fake images and videos. The current research focuses on identifying the forgeries in the images using different DL models and techniques. In this section, we discuss some of the related work based on some of image tampering detection methods and the residual exploitation of CNN models used for the image tampering detection methods.

In the method of copy-move forgeries, the image is divided into overlapping blocks, and correlation is determined for the cloned blocks. A patch detection-based algorithm [16] was used to approximate the neighbours for the forgery detection. Geometrical-based transformations with invariant features of the image were used for the copy-move forgery detection. Local binary pattern (LBP) and steerable pyramid transform (SPT) were used for image forgery detection [17, 18]. These methods are used for the traditional extraction of the tampered regions for the forgery detection. However, these methods fail for the images that are small in size and provide inaccurate tampered regions of the image.

DL methods are widely used for image forgery detection in recent works [19, 20]. The image manipulation tasks that are generally used are generic manipulations, resampling [21], and splicing [20]. One of the works in [22] used Gaussian-based CNN for Steganalysis. In [23], a stacked autoencoder was used for image tampering detection. Further, CNN combined with LSTM was used for image tampering detection using the various layers of CNN. Residual-based networks such as ResNet 50 were used in [15] for image tampering detection using the input of computer-generated images.

In 2015, a variant of CNN called U-Net was proposed in [24]. U-Net gained a huge success in neuronal structure segmentation, and because of its features which are propagated among layers, its framework is path breaking in the above field. The context information in U-Net is captured by successive layers; later, the output feature is up sampled and finally combined with the high-resolution features propagated by a symmetric expanding path. This enables precise location and also reduces the loss of detail information. This helped to propose some image segmentation methods [25, 26] based on U-Net. Most of the time in image splicing forgery detection, we need to segment out the tampered region in an image which is impossible to do with human eyes. Hence, image splicing forgery detection can be understood as a complicated image segmentation task which is independent of the human visual system. Extraction of discriminative features plays a vital role in locating tampered regions of an image by providing the differences of image attributes. Even though U-Net can extract relatively shallow discriminative features, only two sides of the U-Net structure are interactive; this is not enough to locate the tampered regions. Besides, the gradient degradation problem [26] is observed when the network architecture becomes much more deeper.

VLAD [27] is a representation used in image recognition which is encoded by the residual vectors with respect to a dictionary. The formulated probabilistic version of VLAD [27] is used to form Fischer vector [28]. Both the representations are powerful for image retrieval and classification. Encoding residual vectors [28] is preferred over encoding original vectors for vector quantization. Multigrid method [29] is widely used in computer graphics and low-level vision to solve Partial Differential Equations (PDEs). This method develops subproblems at multiple scales, where each subproblem gives the residual solution between the finer and the coaster scale. Hierarchical basis preconditioning [30] is an alternative to Multigrid which relies on the variables that represent the residual vectors between the coarser and finer scales. As the standard solvers are unaware of the residual nature of the solutions, they converge slower compared to the Multigrid or Hierarchical basis preconditioning solvers [30]. These methods suggest that preconditioning or good reformulation can make the optimization easy.

Image classification using deep learning techniques involves the same three steps that are followed in machine learning algorithms for image classification. Those three steps are preprocessing, feature extraction, and classification. First, the input dataset is divided into two sets for training and testing. Both training and testing images are then preprocessed to resize the images according to the pretrained network size [31]. Further, these preprocessed images are sent through various layers of the network until the fully connected layer (FC-1000) extract features from the images sent. The classifier model is trained by passing the features extracted by FC-1000. The prediction of test images is done using the trained classifier model. Naïve Bayes, K-nearest neighbour, and multiclass model using SVM learner are the three classifiers used in this model.

In 2015, the ResNet architecture was proposed [32] which won the championship in the classification task of ImageNet match. For a few stacked layers in ResNet, residual mapping is defined as $y = F(x) + x(1)$ Equation (1), where x represents the input, the operation $F(x) + x$ is performed by a shortcut connection and element-wise addition, and y represents the output. The gradient degradation problem is a serious problem in image splicing forgery detection. This is generally seen in deeper networks; hence, the residual mapping technique is proposed to overcome this problem. The differences of image essence attributes are hard to discover through the multilayer structure as the discrimination of image essence attribute features will be weaker. To solve the above issue and to simultaneously strengthen the learning way of CNN, the residual mapping should be utilized more efficiently.

To make full use of features to detect tampering and to fuse features, adaptive attention mechanism and residual refinement network [33] are used which are robust to various postprocessing, such as blur, noise, and JPEG recompression. Residual-based [34] descriptors have proven extremely effective for a number of image forensic applications. Experimental results based on residual-based fully convolutional network [35] for image tampering detection for various datasets performed better than some existing methods in generalization ability, localization ability, and robustness against additional operations.

As CNN contains numerous parameters, weights, layers (spatial filters), biases, and so on, nowadays, they are widely used for detecting image forgery. The convolution operation in CNN considers the neighbourhood of the pixels in an image which results in different sizes of layers (spatial features). Various sizes of filters encapsulate the images with unique levels of granularity. The coarse-grained features of the image are extracted using the large-sized filters, while the fine-grained portions of the images are extracted using the small filters. Various researches on adjusting the size of the filters are conducted to optimize the performance of CNN to extract both coarse-grained and fine-grained features of an image.

The evaluation metrics play a vital role in estimating the tampering in images. There are two types of metrics used for the evaluation, pixel-based and image-based [36]. In the pixel-based method, the classification of the pixels is done as copy-move and authentic, whereas in the image-based method, the classification of image is done as either tampered or authentic. The measures used at image level are TP (true positive): tampered images are detected as tampered images, TN (true negative): nontampered images are detected as nontampered images, FP (false positive): nontampered images are detected as tampered images, and FN (false negatives): tampered images are detected as nontampered images. In this paper, the proposed method uses image-based methods to evaluate the accuracy. Among the existing methods discussed in this section, the CNN model is used to extract the spatial features of the image which includes the geometry, texture, wavelet, and transformations. The weights of the majority of the above-discussed models need to be altered each time for a new dataset of images as they use pretrained weights. In the proposed system, a fusion of decision-making is involved for image tampering detection based on the CNN models. The proposed fusion model is discussed in further sections.

3. Proposed System

The architecture of the proposed fusion system of residual exploitation-based CNN models is as shown in Figure 1. The residual exploitation-based CNN models chosen are ResNet-18, Resnet-50, and ResNet-101. It consists of three stages, namely, data preprocessing, fusion model, and the classification. In the data preprocessing stage, the input image is preprocessed based on the dimensions required by the fusion models. A support vector machine (SVM) is used for the classification of the image as tampered/forged or not.

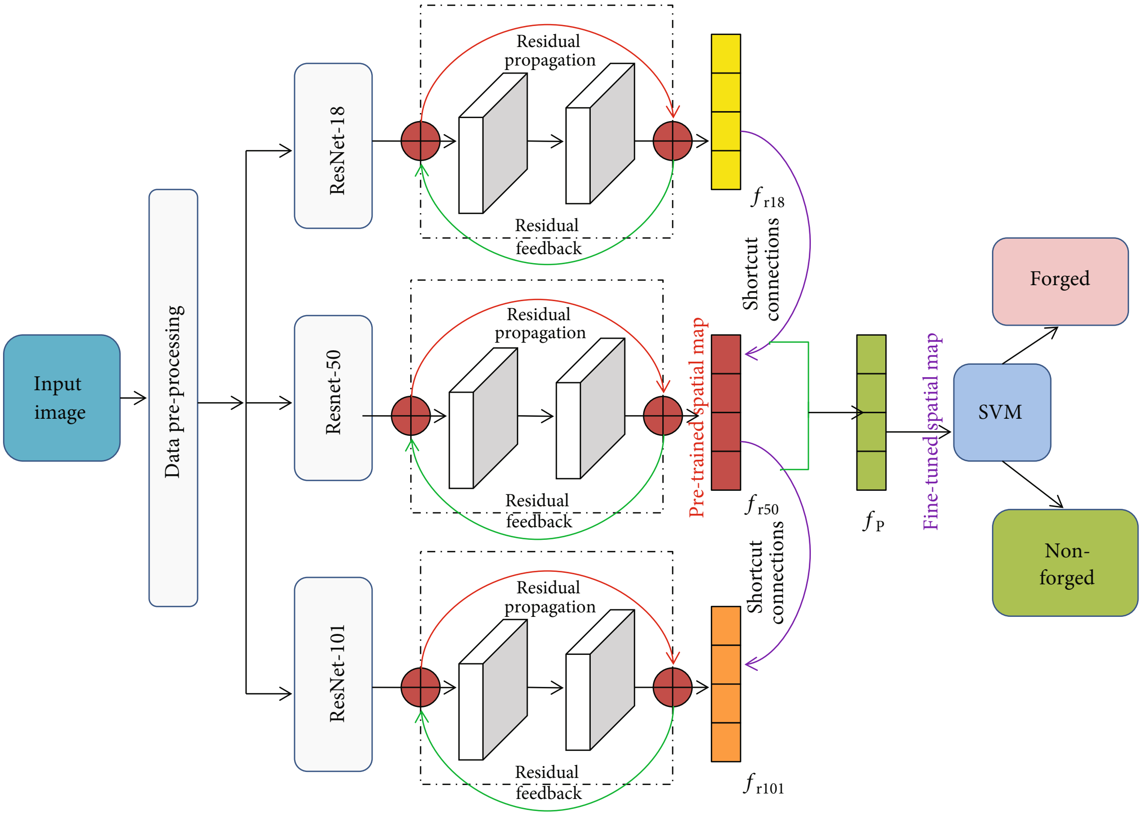


Figure 1: Architecture of fusion of residual exploitation-based CNN models.

The proposed system is implemented in two parts, i.e., pretrained and fine-tuned. In the pretrained implementation, regularization is not applied, and the pretrained weights are used for the classification. Regularization is a set of techniques that can prevent overfitting in neural networks and thus improve the accuracy of convolutional neural network-based models. Thus, to minimize the effect of overfitting regularization is applied in fine-tuned model implementation. In the fine-tuned implementation, regularization is applied for the classification. Initially, we discuss the residual exploitation-based CNN models, and then, the strategy used for the regularization is discussed in the further sections.

3.1. Data Preprocessing. In this stage, the input image that needs to be identified whether it is tampered or not is subjected to preprocessing. The dimensions of the input image required for ResNet-18 is 224×224. The dimensions of the input image required for ResNet-50 is 224×224. The dimensions of the input image required for ResNet-101 is 224×224 . The input image is preprocessed first based on the dimensions required for each of the residual exploitation-based CNN models. Each CNN model then takes the input image to produce the feature vector in the further stages.

3.2. Residual Exploitation-Based CNN Models for Image Classification. The different residual exploitation-based CNN models that are considered for fusion are ResNet-18, ResNet-50, and ResNet-101. These models are used for the image classification problems numerously. In this section, these models are discussed briefly. The residual deep learning models considered are summarized as shown in Table 1.

3.2.1. ResNet-18. It is a CNN trained on the ImageNet dataset with 18 layers deep and can classify the images upto 1000 categories. The network has learnt rich representations of the images with 11.7 million parameters. The network has an image input size of 224-by-224.

3.2.2. ResNet-50. It is a CNN trained on the ImageNet dataset with 50 layers deep and can classify the images upto 1000 categories. The network has learnt rich representations of the images with 25.6 million parameters. The network has an image input size of 224-by-224.

3.2.3. ResNet-101. It is a CNN trained on the ImageNet dataset with 101 layers deep and can classify the images upto 1000 categories. The network has learnt rich representations of the images with 44.6 million parameters. The network has an image input size of 224-by-224.

3.2.4. SVM. SVM is used as a classifier, as it is more suitable, popular, efficient, and widely used for binary classification problems as compared to other classifiers. Performance of the proposed approach is evaluated at image level by calculating the performance metrics as precision, false-positive rate (FPR), and recall, also known as true positive rate (TPR), F -score, and accuracy.

Table 1: Residual exploitation-based CNN models for image classification [36].

Residual exploitation-based CNN models	Depth	Number of parameters (million)	Input size
ResNet-18	18	11.7	224-by-224
ResNet-50	50	25.6	224-by-224
ResNet-101	101	44.6	224-by-224

3.3. Fusion Model and Regularization. The proposed system is first implemented as a CNN with pretrained weights for the image classification. Afterwards, the proposed system is implemented as a fusion of the residual exploitation-based CNN models as discussed in the previous section. Initially, the input image is passed to the residual exploitation-based CNN models to obtain their feature maps, respectively. The feature map from the ResNet-18 is denoted as f_{r18}, the feature map from the ResNet-50 is denoted as f_{r50}, and the feature map from the ResNet-101 is denoted as f_{r101}. For the fusion model, the pretrained CNN output feature mapping fp is used. This feature map f_{p} is a combination of the feature maps obtained from the residual exploitation-based CNN models as shown in Equation (1).

$$f_{\mathrm{p}} = f_{\mathrm{r18}} + f_{\mathrm{r50}} + f_{\mathrm{r101}}. \tag{1}$$

The fusion model uses feature map f_{p} as a local descriptor for input patch to extract the features of the image. The image for the fusion model is represented as a function $Y_{\mathrm{fusion}} = f(x)$ where x is the patch in the input image. For a test image size $m \times n$, a sliding window of size $p \times p$ is used to compute the local descriptor Y_i is computed as shown in Equation (2). It is obtained as a concatenation of all the input patches X_i, and the new image representation is given by Equation (3) where s is the size of the stride used for transforming the input patch; this new image representation fusion is used as the feature map for the classification by the SVM as tampered or nontampered. In Equation (2), W_s represents the weights of the shortcut connections from the residual features of ResNet-18, ResNet-50, and ResNet-101 models.

$$Y_{\mathrm{fusion}} = [Y_1 + Y_2 + \cdots + Y_T] + W_s, \tag{2}$$

$$f_{\mathrm{fusion}} = \frac{m-w}{s} + 1 * \frac{n-w}{s} + 1. \tag{3}$$

For fine tuning of the parameters of the fusion model, the initialization of the weight kernels is used as shown in Equation (4). In this equation, W_{f} represents the weights of the fusion model, W_{r18} represents the weights of the ResNet-18 model, W_{r50} represents the weights of the ResNet-50 model, and W_{r101} represents the weights of the ResNet-101 model. The weight of the fusion model W_{f} is initialized as shown in Equation (5). The initialization of the weights acts as a regularization term and facilitates the fusion model to learn robust features of detecting the forgery rather than the complex image representations.

$$W_{\mathrm{f}} = \left[W_{\mathrm{r18}j} + W_{\mathrm{r50}j} + W_{\mathrm{r101}j}\right], \quad \text{where } j = 1, 2, 3, \tag{4}$$

$$W_{\mathrm{f}} = \left[W_{\mathrm{r18}}^{4k-2} + W_{\mathrm{r50}}^{4k-2} + W_{\mathrm{r101}}^{4k}\right] + W_{\mathrm{s}}, \quad \text{where } k = ((j+1) \bmod 11) + 1. \tag{5}$$

4. Experiments and Results

In this section, the experiments and results of the proposed fusion model are discussed. The experiment is carried out in two stages. In the first stage, the residual exploitation-based CNN models are used with the pretrained weights, in the second stage, the fusion model with the strategy of weight initialization as discussed in the previous section. The configuration of the system used for the experiments is shown in Table 2.

4.1. Dataset. The dataset used for the experiment is benchmark publicly available MICC-F220 [37] of 110 nonforged images and 110 forged images with 3 channels, i.e., color images of size 722 × 480 to 800 × 600 pixels with 10 different combinations of geometrical and transformations attacks as shown in Figures 2 and 3. To avoid the problem of overfitting and to generalize the approach k-fold cross-validation with the value of k as 5 is used for training and testing sampling of images.

4.2. Baseline Models and Metrics. The baseline models that are used for the comparison of the fusion model are summarized as follows.

(i) SIFT: It uses the forensic method of the image tampering detection using a scale invariant features transform (SIFT) approach [37].

(ii) SURF: It uses a speeded up robust features (SURF) and hierarchical agglomerative clustering (HAC) for the image tampering detection [38].

(iii) DCT: It uses discrete cosine transform (DCT) features for each block and through lexicographical sorting of block-wise DCT coefficients for the image tampering detection [39].

(iv) PCA: It uses PCA on the image blocks to reduce the dimension space and perform lexicographical sorting for the image tampering detection [40].

(v) CSLBP: It uses center-symmetric local binary pattern (CSLBP) based on the combined features of Hessian points for the image tampering detection [41].

(vi) SYMMETRY: It uses the local symmetry value of an image to compute the key points for image tampering detection [42].

Table 2: Configuration of system.

Hardware	Intel(R) Xeon(R) Silver 4110 CPU with 2.10 GHZ, 128 GB RAM
GPU	Tesla P4
Software	Ubuntu 18.04 with Matlab release R2020a

Figure 2: Original image.

(vii) CLUSTERING strategy: It uses SIFT features with a clustering strategy to detect image tampering [43].

The basic metrics that are used for the evaluation of the fusion model are false-positive rate (FPR), recall (R), precision (P), f-score, and accuracy as shown in the equations (Equations (7) to (10)). The confusion matrix is used as the basis for the evaluation of the tampered and nontampered images as shown in Table 3, and the notations used are as follows:

(i) TP: Tampered image detected as tampered.

(ii) FN: Tampered image detected as nontampered.

(iii) FP: Nontampered image detected as tampered.

(iv) TN: Nontampered image detected as nontampered.

$$\text{FPR} = \frac{\text{FP}}{(\text{FP} + \text{TN})}, \tag{6}$$

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})}, \tag{7}$$

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})}, \tag{8}$$

$$F1\ \text{Score} = \frac{2 * (\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})}, \tag{9}$$

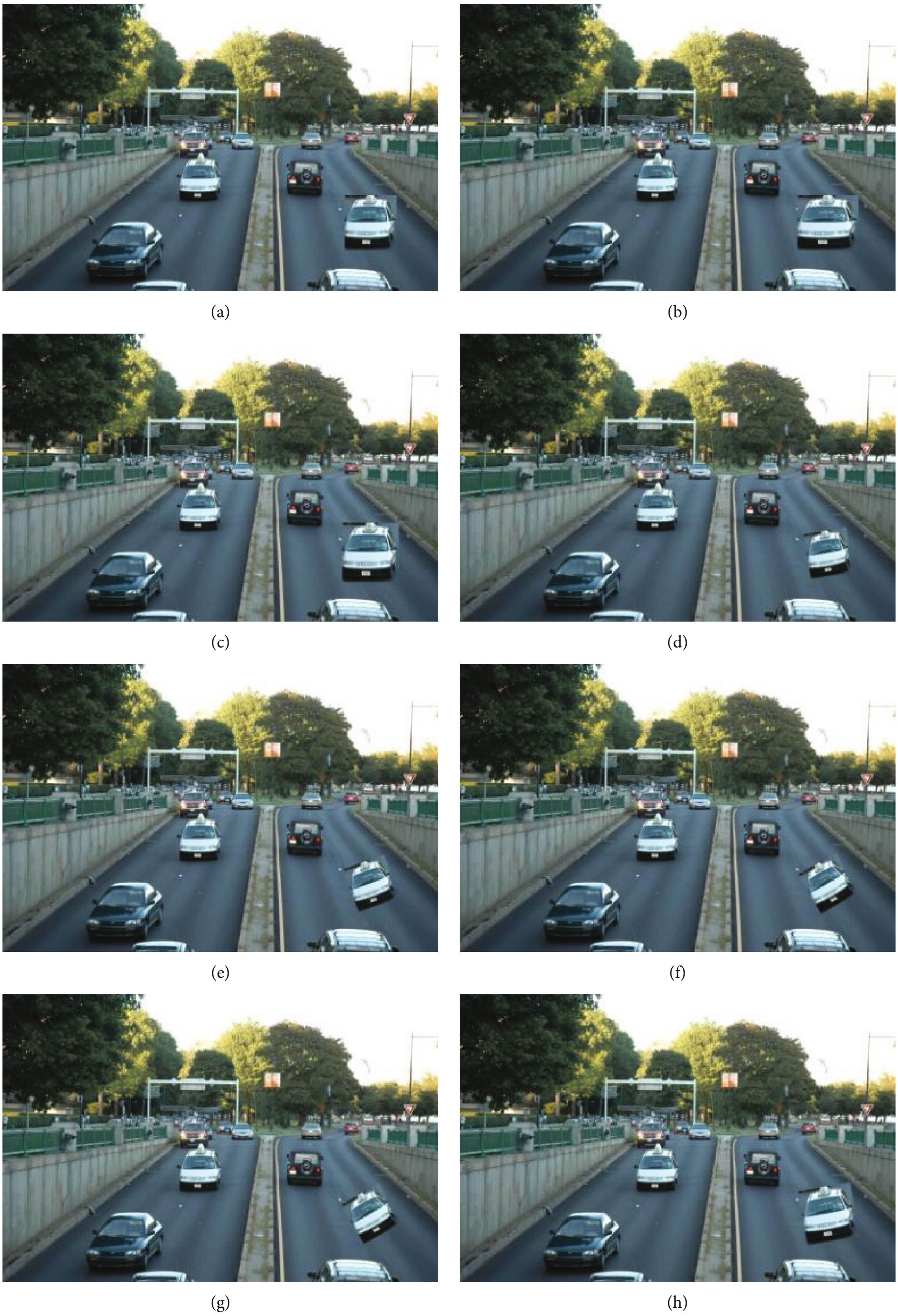
$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN}))}. \tag{10}$$

4.3. Pretrained Residual Exploitation-Based CNN Models. In this section, the results of the pretrained residual-based CNN models are discussed. The three models, namely, ResNet-18, ResNet-50, and ResNet-101 are used with the pretrained weights for the image tampering detection. Table 4 shows the confusion matrix for the ResNet-18 model. It can be observed that the accuracy of the ResNet-18 model is 92.27%, and the percentage of the prediction of the correct tampered images is 50% and correct nontampered images is 42.27%. However, the wrong tampered image prediction is 7.23%. Table 5 shows the confusion matrix for the ResNet-50 model. It can be observed that the accuracy of the ResNet-50 model is 92.27%, and the percentage of the prediction of the correct tampered images is 50% and correct nontampered images is 42.27%. However, the wrong tampered image prediction is 7.23%. Table 6 shows the confusion matrix for the ResNet-101 model. It can be observed that the accuracy of the ResNet-101 model is 91.81%, and the percentage of the prediction of the correct tampered images is 50% and correct nontampered images is 41.82%. However, the wrong nontampered prediction is 8.18%.

The ROC curve is used to estimate the AUC values for the pretrained residual exploitation-based convolutional neural networks as shown in Figure 4. Figure 4(a) represents the ROC curve for the pretrained ResNet-18 model with AUC of 97.57%. Figure 4(b) represents the ROC curve for the pretrained ResNet-50 model with AUC of 97.57%. Figure 4(c) represents the ROC curve for the pretrained ResNet-101 model with AUC of 96.52%.

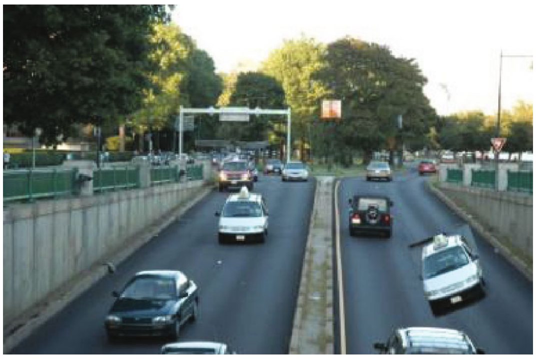
4.4. Fine-Tuned Residual Exploitation-Based CNN Models. In this section, the results of the fine-tuned residual exploitation-based models are discussed. The three models, namely, ResNet-18, ResNet-50, and ResNet-101 are used with the fine-tuned weights for image tampering detection. Table 7 shows the confusion matrix for the fine-tuned ResNet-18 model. It can be observed that the accuracy of the fine-tuned ResNet-18 model is 95.0%, and the percentage of the prediction of the correct tampered images is 50% and correct nontampered images is 45.0%. However, the wrong tampered image prediction is 5.0%. Table 8 shows the confusion matrix for the fine-tuned ResNet-50 model. It can be observed that the accuracy of the fine-tuned ResNet-50 model is 90.90%, and the percentage of the prediction of the correct tampered images is 50% and correct nontampered images is 40.91%. However, the wrong tampered image prediction is 9.09%. Table 9 shows the confusion matrix for the fine-tuned ResNet-101 model. It can be observed that the accuracy of the fine-tuned ResNet-101 model is 87.27%, and the percentage of the prediction of the correct tampered images is 45.45% and correct nontampered images is 41.82%. However, the prediction of the wrong tampered images is 8.18%, and wrong nontampered image prediction is 4.55%.

The ROC curve is used to estimate the AUC values for the fine-tuned residual exploitation-based models as shown in



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 3: Continued.



(i)

Figure 3: (a–i) Tampered images with different combinations of geometrical and transformations attacks.

Table 3: Confusion matrix.

Actual images	Predicted tampered	Predicted nontampered
Tampered	(True positive) TP	(False negative) FN
Nontampered	(False positive) FP	(True negative) TN

Table 4: Confusion matrix for the pretrained ResNet-18 model.

Image dataset	Tampered image prediction	Nontampered image prediction
Tampered images	50%	0%
Nontampered images	7.23%	42.27%

Table 5: Confusion matrix for the pretrained ResNet-50 model.

Image dataset	Tampered image prediction	Nontampered image prediction
Tampered images	50%	0%
Nontampered images	7.23%	42.27%

Table 6: Confusion matrix for the pretrained ResNet-101 model.

Image dataset	Tampered image prediction	Nontampered image prediction
Tampered images	50%	0%
Nontampered images	8.18%	41.82%

Figure 5. Figure 5(a) represents the ROC curve for the fine-tuned ResNet-18 model with AUC of 97.0. Figure 5(b) represents the ROC curve for the fine-tuned ResNet-50 model with AUC of 93.91%. Figure 5(c) represents the ROC curve for the ResNet-101 model with AUC of 92.0%.

4.5. Fusion Model. In this section, the results of the fusion models are discussed. Table 10 shows the confusion matrix for the pretrained fusion models. It can be observed that the accuracy of the pretrained fusion model is 90.91%, and the percentage of the prediction of the correct tampered images is 49.55% and correct nontampered images is 41.36%. However, the wrong tampered image prediction is 8.64%, and wrong nontampered image prediction is 0.45%.

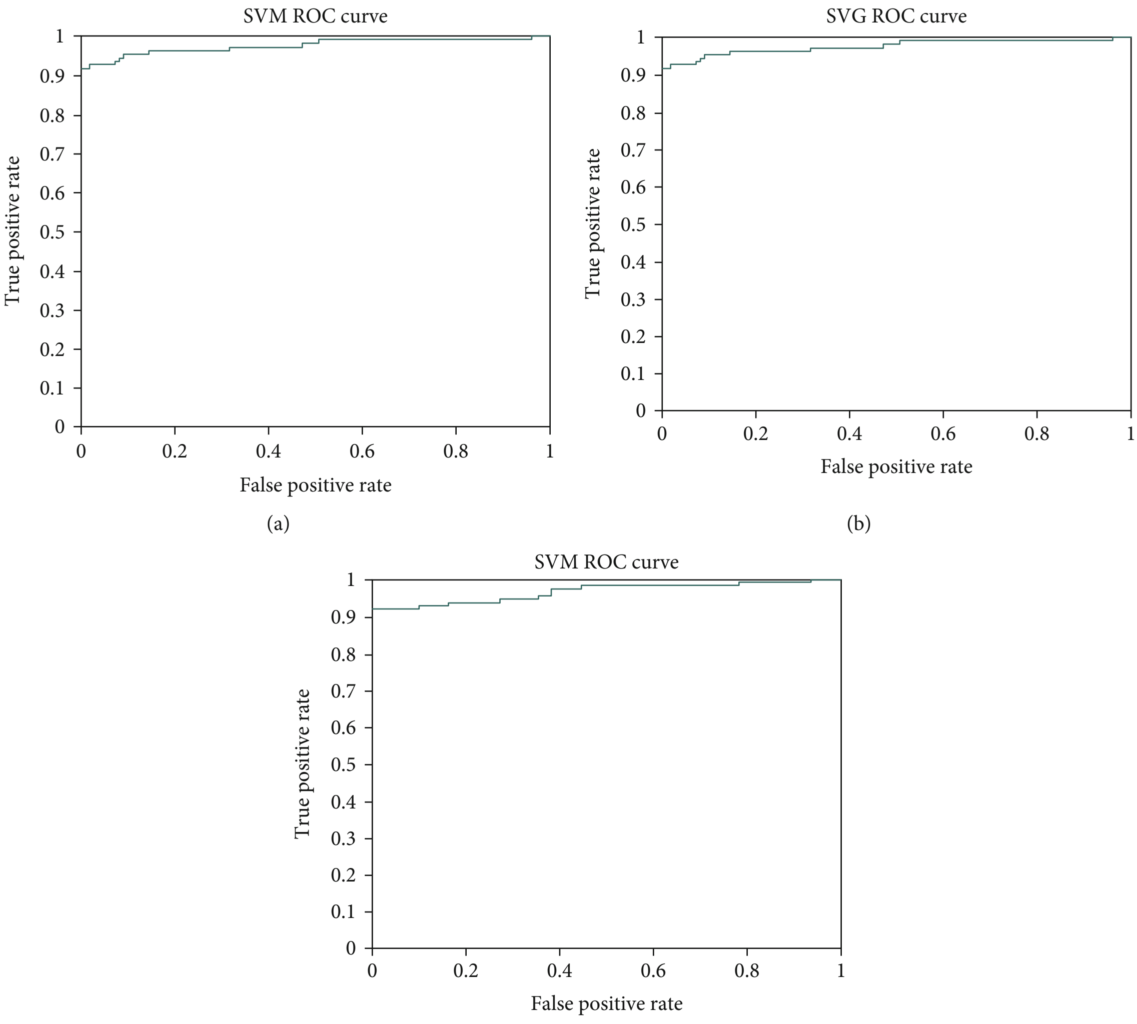
Table 11 shows the confusion matrix for the fine-tuned fusion model. It can be observed that the accuracy of the fine-tuned fusion network is 93.18%, and the percentage of the prediction of the correct tampered images is 50% and the correct nontampered images is 43.18%. However, the wrong tampered image prediction is 6.82%. It can be clearly observed that the percentage of wrong tampered image prediction is less as compared to the pretrained residual exploitation-based models. The accuracy of the fine-tuned fusion model is higher than the pretrained fusion model.

4.6. Performance Comparison

4.6.1. Performance Comparison with Pretrained Residual Exploitation-Based Models. In this section, the performance comparison of the fusion model is carried out with pretrained residual exploitation-based CNN models. The metrics used for the comparison are precision, recall, f-score, and accuracy. The results of the performance comparisons are as shown in Table 12. The precision and recall metrics are important to determine the effectiveness of the CNN models. According to Equations (7) to (10), the values in Table 12 were obtained.

4.6.2. Performance Comparison with Fine-Tuned Residual Exploitation-Based Models. In this section, the performance comparison of the fusion model is carried out with fine-tuned residual exploitation-based CNN models. The metrics used for the comparison are precision, recall, f-score, and accuracy. The results of the performance comparison are shown in Table 13.

It can be observed from the values in Table 13 that the proposed fusion model achieves comparatively more precision, recall, and f-score than the fine-tuned residual exploitation-based CNN models. The results of the performance comparison of the fusion model with the baseline



(a)

(b)

(c)

Figure 4: ROC curves for the pretrained residual-based models.

Table 7: Confusion matrix for the fine-tuned ResNet-18 model.

Image dataset	Tampered image prediction	Nontampered image prediction
Tampered images	50%	0%
Nontampered images	5.0%	45.0%

Table 8: Confusion matrix for the fine-tuned ResNet-50 model.

Image dataset	Tampered image prediction	Nontampered image prediction
Tampered images	50%	0%
Nontampered images	9.09%	40.91%

Table 9: Confusion matrix for the fine-tuned ResNet-101 model.

Image dataset	Tampered image prediction	Nontampered image prediction
Tampered images	45.45%	4.55%
Nontampered images	8.18%	41.82%

models are as shown in Table 14. The metrics used for the comparison are the FPR and TPR as they give the correctness of the model for the image tampering detection. The FPR for baseline 1 [37] is 8%, baseline 2 [38] is 3.64%, baseline 3 [39] is 84%, baseline 4 [40] is 86%, baseline 5 [41] is 2.89%, baseline 6 [42] is 5.45%, baseline 7 [43] is 7.63%, proposed pretrained fusion model is 17.27%, and proposed fine-tuned fusion model is 13.63%. The TPR for baseline 1 [37] is 100%, baseline 2 [38] is 73.64%, baseline 3 [39] is 89%, baseline 4 [40] is 87%, baseline 5 [41] is 96%, baseline 6 [42] is 83.64%, baseline 7 [43] is 97.87%, proposed pretrained fusion

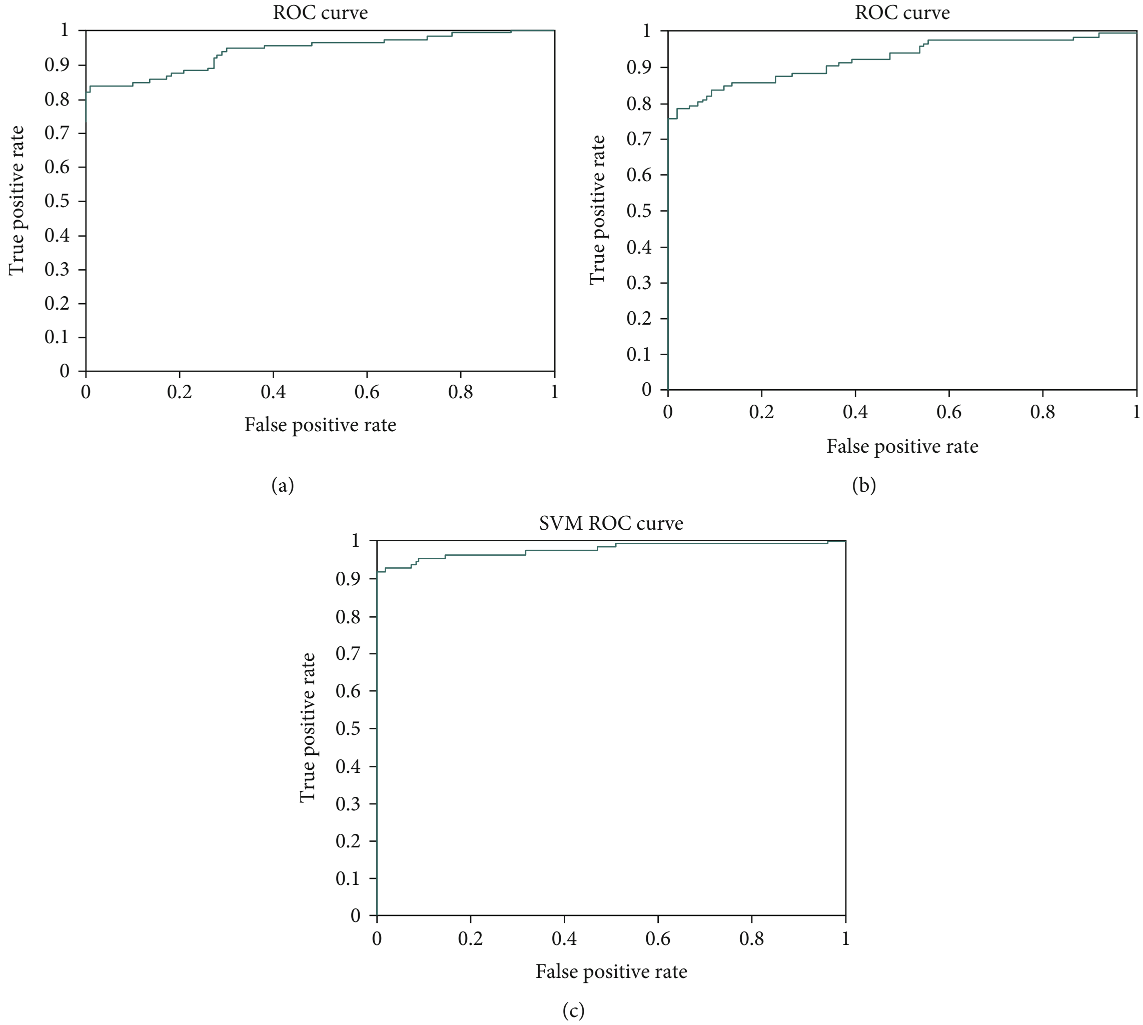


Figure 5: ROC curves for the fine-tuned residual-based models.

Table 10: Confusion matrix for the decision fusion for pretrained models.

Image dataset	Tampered image prediction	Nontampered image prediction
Tampered images	49.55%	0.45%
Nontampered images	8.64%	41.36%

Table 11: Confusion matrix for the decision fusion for fine-tuned models.

Image dataset	Tampered image prediction	Nontampered image prediction
Tampered images	50.0%	0%
Nontampered images	6.82%	43.18%

Table 12: Comparison of metrics for the pretrained residual exploitation-based models.

Model	Precision (%)	Recall (%)	*F*-score (%)	Accuracy (%)
ResNet-18	86.61	100	92.82	92.27
ResNet-50	86.61	100	92.82	92.27
ResNet-101	85.93	100	92.43	91.81
Fusion model	85.15	99.09	91.59	90.91

Table 13: Comparison of metrics for the fine-tuned residual exploitation-based models.

Model	Precision (%)	Recall (%)	*F*-score (%)	Accuracy (%)
ResNet-18	90.90	100	95.23	95.0
ResNet-50	84.61	100	91.66	90.90
ResNet-101	84.74	90.90	87.71	87.27
Fusion model	88.0	100	93.61	93.18

Table 14: Comparison of the proposed fusion-based models with baseline models.

Approach	FPR (%)	TPR (%)
SIFT [37]	8	100
SURF [38]	3.64	73.64
DCT [39]	84	89
PCA [40]	86	87
CSLBP [41]	2.89	96
SYMMETRY [42]	5.45	83.64
CLUSTERING strategy [43]	7.63	97.87
Pretrained fusion model (proposed)	17.27	99.09
Fine-tuned fusion model (proposed)	13.63	100

model is 99.09%, and proposed fine-tuned fusion model is 100%. Therefore, it can be observed that the fusion model has higher TPR as compared to the baseline models due to the weight initialization strategy used for the fusion model.

5. Conclusion

Image tampering detection helps to differentiate between the original and the manipulated or fake images. In this paper, a decision fusion of residual exploitation-based CNN models is implemented for image tampering detection. The idea was to use the residual exploitation-based CNN models, namely, ResNet-18, ResNet-50, and ResNet-101, and then combine all these models to obtain the decision to detect the tampering of the image. Regularization of the weights of the pretrained models is implemented to arrive at a decision of the image tampering. The experiments carried out indicate that the fusion-based approach gives more accuracy than the state-of-the-art approaches. In the future, the fusion decision can be improved with other weight initialization strategies for image tampering detection.

Data Availability

The data used for the research is already taken from public repository, and the link is provided in the paper.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] V. Manu and B. Mehtre, “Visual artifacts based image splicing detection in uncompressed images,” in *2015 IEEE International Conference on Computer Graphics, Vision and Information Security (CGVIS)*, pp. 145–150, Bhubaneswar, India, 2015.

[2] Y. Cao, T. Gao, L. Fan, and Q. Yang, “A robust detection algorithm for copy-move forgery in digital images,” *Forensic Science International*, vol. 214, no. 1–3, pp. 33–43, 2012.

[3] M. F. Hashmi, V. Anand, and A. G. Keskar, “Copy-move image forgery detection using an efficient and robust method combining un-decimated wavelet transform and scale invariant feature transform,” *AASRI Procedia*, vol. 9, pp. 84–91, 2014.

[4] J. Li, X. Li, B. Yang, and X. Sun, “Segmentation-based image copy-move forgery detection scheme,” *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 3, pp. 507–518, 2015.

[5] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, NV, USA, 2016.

[6] X. Bi, Y. Wei, B. Xiao, and W. Li, “RRU-Net: the ringed residual U-Net for image splicing forgery detection,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1–10, Long Beach, CA, USA, 2019.

[7] A. K. Jaiswal and R. Srivastava, “Image splicing detection using deep residual network,” in *Proceedings of 2nd International Conference on Advanced Computing and Software Engineering (ICACSE)*, UP, India, 2019.

[8] M. Heikkilä, M. Pietikäinen, and C. Schmid, “Description of interest regions with local binary patterns,” *Pattern Recognition*, vol. 42, no. 3, pp. 425–436, 2009.

[9] Y. LeCun, B. Boser, J. S. Denker et al., “Backpropagation applied to handwritten zip code recognition,” *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989.

[10] X. Liu, Z. Deng, and Y. Yang, “Recent progress in semantic image segmentation,” *Artificial Intelligence Review*, vol. 52, no. 2, pp. 1089–1106, 2019.

[11] D. Ciresan, A. Giusti, L. Gambardella, and J. Schmidhuber, “Deep neural networks segment neuronal membranes in electron microscopy images,” *Advances in Neural Information Processing Systems*, vol. 25, pp. 2843–2851, 2012.

[12] S. Ren, K. He, R. Girshick, X. Zhang, and J. Sun, “Object detection networks on convolutional feature maps,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 7, pp. 1476–1481, 2016.

[13] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, Boston, Massachusetts, 2015.

[14] J. H. Bappy, A. K. Roy-Chowdhury, J. Bunk, L. Nataraj, and B. Manjunath, “Exploiting spatial structure for localizing manipulated image regions,” in *Proceedings of the IEEE international conference on computer vision*, pp. 4970–4979, California, USA, 2017.

[15] E. R. De Rezende, G. C. Ruppert, A. Theophilo, E. K. Tokuda, and T. Carvalho, “Exposing computer generated images by using deep convolutional neural networks,” *Signal Processing: Image Communication*, vol. 66, pp. 113–126, 2018.

[16] D. Cozzolino, G. Poggi, and L. Verdoliva, “Efficient dense-field copy–move forgery detection,” *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 11, pp. 2284–2297, 2015.

[17] G. Muhammad, M. H. Al-Hammadi, M. Hussain, and G. Bebis, “Image forgery detection using steerable pyramid transform and local binary pattern,” *Machine Vision and Applications*, vol. 25, no. 4, pp. 985–995, 2014.

[18] C. Guillemot and O. Le Meur, “Image inpainting: overview and recent advances,” *IEEE Signal Processing Magazine*, vol. 31, no. 1, pp. 127–144, 2013.

[19] B. Bayar and M. C. Stamm, “Design principles of convolutional neural networks for multimedia forensics,” *Electronic Imaging*, vol. 2017, no. 7, pp. 77–86, 2017.

[20] Y. Rao and J. Ni, “A deep learning approach to detection of splicing and copy-move forgeries in images,” in *2016 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–6, Abu Dhabi, United Arab, 2016.

[21] B. Bayar and M. C. Stamm, “On the robustness of constrained convolutional neural networks to jpeg post-compression for image resampling detection,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2152–2156, New Orleans, LA, USA, 2017.

[22] Y. Qian, J. Dong, W. Wang, and T. Tan, “Deep learning for steganalysis via convolutional neural networks,” in *Media Watermarking, Security, and Forensics, Vol. 9409*, International Society for Optics and Photonics, 2015.

[23] Y. Zhang, J. Goh, L. L. Win, and V. L. Thing, “Image region forgery detection: a deep learning approach,” in *Proceedings of the Singapore Cyber-Security Conference (SG-CRC) 2016 - Cyber-Security by Design*, pp. 1–11, Singapore, 2016.

[24] O. Ronneberger, P. Fischer, and T. Brox, “U-net: convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, Cham, 2015.

[25] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, “3D U-net: learning dense volumetric segmentation from sparse annotation,” in *International conference on medical image computing and computer-assisted intervention*, pp. 424–432, Springer, Cham, 2016.

[26] V. Iglovikov and A. Shvets, “Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation,” 2018, http://arxiv.org/abs/11801.05746.

[27] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid, “Aggregating local image descriptors into compact codes,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 9, pp. 1704–1716, 2011.

[28] F. Perronnin and C. Dance, “Fisher kernels on visual vocabularies for image categorization,” in *2007 IEEE conference on computer vision and pattern recognition*, pp. 1–8, Minneapolis, MN, USA, 2007.

[29] M. Brezina, A. J. Cleary, R. D. Falgout et al., “Algebraic multigrid based on element interpolation (AMGe),” *SIAM Journal on Scientific Computing*, vol. 22, no. 5, pp. 1570–1592, 2001.

[30] T. Vatanen, T. Raiko, H. Valpola, and Y. LeCun, “Pushing stochastic gradient towards second-order methods–backpropagation learning with transformations in nonlinearities,” in *International Conference on Neural Information Processing*, pp. 442–449, Springer, Berlin, Heidelberg, 2013.

[31] P. Korus and J. Huang, “Multi-scale fusion for improved localization of malicious tampering in digital images,” *IEEE Transactions on Image Processing*, vol. 25, no. 3, pp. 1312–1326, 2016.

[32] R. Girshick, “Fast r-cnn,” in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.

[33] Y. Zhu, C. Chen, G. Yan, Y. Guo, and Y. Dong, “A R-net: adaptive attention and residual refinement network for copy-move forgery detection,” *IEEE Transactions on Industrial Informatics*, vol. 16, no. 10, pp. 6714–6723, 2020.

[34] D. Cozzolino, G. Poggi, and L. Verdoliva, “Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection,” in *Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security*, pp. 159–164, New York, NY, USA, 2017.

[35] B. Chen, X. Qi, Y. Zhou, G. Yang, Y. Zheng, and B. Xiao, “Image splicing localization using residual image and residual-based fully convolutional network,” *Journal of Visual Communication and Image Representation*, vol. 73, p. 102967, 2020.

[36] O. M. Al-Qershi and B. E. Khoo, “Evaluation of copy-move forgery detection: datasets and evaluation metrics,” *Multimedia Tools and Applications*, vol. 77, no. 24, pp. 31807–31833, 2018.

[37] I. Amerini, L. Ballan, R. Caldelli, A. Del Bimbo, and G. Serra, “A sift-based forensic method for copy–move attack detection and transformation recovery,” *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 3, pp. 1099–1110, 2011.

[38] A. Kuznetsov, “Digital image forgery detection using deep learning approach,” *Journal of Physics: Conference Series*, vol. 1368, no. 3, article 032028, 2019.

[39] A. J. Fridrich, B. D. Soukal, and A. J. Lukáš, “Detection of Copy-Move Forgery in Digital Images,” *in Proceedings of Digital Forensic Research Workshop*, CiteSeerX, 2003.

[40] A. C. Popescu and H. Farid, *Exposing digital forgeries by detecting duplicated image regions*, Tech. Rep. TR2004-515, Dartmouth, 2004.

[41] D. M. Uliyan, H. A. Jalab, A. Abdul Wahab, S. Sadeghi, and S. Sadeghi, “Image region duplication forgery detection based on angular radial partitioning and Harris key-points,” *Symmetry*, vol. 8, no. 7, p. 62, 2016.

[42] D. Vaishnavi and T. Subashini, “Application of local invariant symmetry features to detect and localize image copy move forgeries,” *Journal of Information Security and Applications*, vol. 44, pp. 23–31, 2019.

[43] M. Abdel-Basset, G. Manogaran, A. E. Fakhry, and I. El-Henawy, “2-Levels of clustering strategy to detect and locate copy-move forgery in digital images,” *Multimedia Tools and Applications*, vol. 79, pp. 5419–5437, 2020.

Hindawi
BioMed Research International
Volume 2024, Article ID 9898717, 1 page
https://doi.org/10.1155/2024/9898717

Retraction

Retracted: Multiomics Analysis of Transcriptome, Epigenome, and Genome Uncovers Putative Mechanisms for Dilated Cardiomyopathy

BioMed Research International

Received 12 March 2024; Accepted 12 March 2024; Published 20 March 2024

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope

(2) Discrepancies in the description of the research reported

(3) Discrepancies between the availability of data and the research described

(4) Inappropriate citations

(5) Incoherent, meaningless and/or irrelevant content included in the article

(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] L. Liu, J. Huang, Y. Liu et al., "Multiomics Analysis of Transcriptome, Epigenome, and Genome Uncovers Putative Mechanisms for Dilated Cardiomyopathy," *BioMed Research International*, vol. 2021, Article ID 6653802, 29 pages, 2021.

Hindawi
BioMed Research International
Volume 2021, Article ID 6653802, 29 pages
https://doi.org/10.1155/2021/6653802

Research Article

Multiomics Analysis of Transcriptome, Epigenome, and Genome Uncovers Putative Mechanisms for Dilated Cardiomyopathy

Li Liu,[1] Jianjun Huang,[2] Yan Liu,[3] Xingshou Pan,[3] Zhile Li,[3] Liufang Zhou,[3] Tengfang Lai,[3] Chengcai Chen,[4] Baomin Wei,[3] Jianjiao Mo,[3] Qinjiang Wei,[3] Wei Yan,[3] Xiannan Huang,[3] Zhen Zhang,[3] Zhuohua Zhang,[3] Meidan Huang,[5] Fengzhen He,[5] and Zhaohe Huang[3]

[1] *Department of Cardiology, Youjiang Medical University for Nationalities, Affiliated Hospital of Youjiang Medical University for Nationalities, Baise, 533000 Guangxi, China*
[2] *Department of Neurology, Youjiang Medical University for Nationalities, Affiliated Hospital of Youjiang Medical University for Nationalities, Baise, 533000 Guangxi, China*
[3] *Department of Cardiology, Affiliated Hospital of Youjiang Medical University for Nationalities, Baise, 533000 Guangxi, China*
[4] *Department of Ultrasound, Affiliated Hospital of Youjiang Medical University for Nationalities, Baise, 533000 Guangxi, China*
[5] *Graduate School of Youjiang Medical University for Nationalities, Baise, 533000 Guangxi, China*

Correspondence should be addressed to Li Liu; liuli011258@sina.com and Zhaohe Huang; bshuangzhaohe@163.com

Received 16 December 2020; Revised 8 February 2021; Accepted 24 February 2021; Published 30 March 2021

Academic Editor: Parameshachari B. D.

Objective. Multiple genes have been identified to cause dilated cardiomyopathy (DCM). Nevertheless, there is still a lack of comprehensive elucidation of the molecular characteristics for DCM. Herein, we aimed to uncover putative molecular features for DCM by multiomics analysis. *Methods*. Differentially expressed genes (DEGs) were obtained from different RNA sequencing (RNA-seq) datasets of left ventricle samples from healthy donors and DCM patients. Furthermore, protein-protein interaction (PPI) analysis was then presented. Differentially methylated genes (DMGs) were identified between DCM and control samples. Following integration of DEGs and DMGs, differentially expressed and methylated genes were acquired and their biological functions were analyzed by the clusterProfiler package. Whole exome sequencing of blood samples from 69 DCM patients was constructed in our cohort, which was analyzed the maftools package. The expression of key mutated genes was verified by three independent datasets. *Results*. 1407 common DEGs were identified for DCM after integration of the two RNA-seq datasets. A PPI network was constructed, composed of 171 up- and 136 downregulated genes. Four hub genes were identified for DCM, including C3 (degree = 24), GNB3 (degree = 23), QSOX1 (degree = 21), and APOB (degree = 17). Moreover, 285 hyper- and 321 hypomethylated genes were screened for DCM. After integration, 20 differentially expressed and methylated genes were identified, which were associated with cell differentiation and protein digestion and absorption. Among single-nucleotide variant (SNV), C>T was the most frequent mutation classification for DCM. MUC4 was the most frequent mutation gene which occupied 71% across 69 samples, followed by PHLDA1, AHNAK2, and MAML3. These mutated genes were confirmed to be differentially expressed between DCM and control samples. *Conclusion*. Our findings comprehensively analyzed molecular characteristics from the transcriptome, epigenome, and genome perspectives for DCM, which could provide practical implications for DCM.

1. Introduction

DCM is the most common inherited cardiomyopathies, characterized by left ventricular dilation and consecutive systolic dysfunction [1]. This disease is the third most common cause of heart failure [2]. About 70% of cases are considered idiopathic [2]. Many factors can induce the occurrence of DCM such as myocarditis, alcohol consumption, drugs, and other toxins [3]. Despite some progress in therapy and diagnosis, DCM patients' prognosis remains unsatisfactory. Given the high prevalence of DCM, understanding the potential molecular characteristics is of importance to reduce DCM-related morbidity and mortality.

Research on the genetics of DCM may provide an in-depth understanding of the pathogenesis of DCM, which assists make better clinical decisions, thereby speeding up the implementation of precision medicine [4]. In recent years, DNA methylation has been widely involved in the regulation of gene expression. Abnormal methylation is closely involved in the pathogenesis of DCM [5]. For example, nuclear DNA methylation in cardiomyocytes has a distinct relationship with left ventricular remodeling and heart failure for DCM patients [6]. Thus, genetic testing has become a promising and effective tool for screening main genetic or epigenetic changes in DCM. Genetic mutations include single nucleotide variants (SNVs), small insertion–deletion, copy number alterations, and translocations. Although it is heritable, DCM exhibits extensive genetic heterogeneity [7]. WES has become a robust diagnostic tool for DCM patients [8]. According to WES studies, a mutation (c.333+2T>C) of TNNI3K has been detected in a Chinese family with DCM [9].

The development of bioinformatics provides high-throughput data at the transcriptome, genome, and epigenome levels for DCM [10]. It is of significance to comprehensively analyze the multiomics to reveal synergistic interactions. Hence, in this study, we aimed to elucidate the molecular characteristics as therapeutic targets for DCM as well as their biological functions by multiomics analysis.

2. Materials and Methods

2.1. DCM Datasets. RNA sequencing (RNA-seq) data of left ventricle samples from 166 healthy donors and 166 DCM patients were obtained from the Gene Expression Omnibus (GEO) repository (https://www.ncbi.nlm.nih.gov/gds/; accession: GSE141910). The GSE141910 dataset was based on the GPL16791 platform. Furthermore, we also downloaded RNA-seq data of left ventricle tissues from 18 healthy donors and 15 DCM patients from the GSE126569 dataset on the GPL16791 and GPL20301 platforms. Raw data were normalized by quantile normalization using the normalizeBetweenArrays function in the limma package [11]. Cardiac DNA methylation profiles from 8 control and 9 DCM specimens were recruited from the GSE42510 dataset on the GPL8490 platform [12]. The correlations between different samples were calculated based on the gene expression and methylation profiles.

2.2. Differential Expression and Methylation Analysis. Differentially expressed (DEGs) or methylated (DEMs) genes between DCM and control left ventricle tissues were identified in line with the criteria of $|\text{fold change (FC)}| \geq 1.5$ and adjusted p value < 0.05. All DEGs or DEMs were visualized into scatter plots, volcano plots, and heat maps.

Table 1: Clinical features for 69 DCM patients.

Clinical features	Number of patients
Gender	
Female	15
Male	54
Age	52.68 ± 12.46
Age of onset	48.37 ± 13.33
BMI (kg/m^2)	23.02 ± 3.12
Number of onsets	3.41 ± 3.36
Hypertension	
Yes	18
No	51
Drinking	
Yes	35
No	34
Smoking	
Yes	36
No	33
Left atrial diameter (LAD)	45.45 ± 7.97
Left ventricular internal diameter (LVIDd)	67.72 ± 9.40
Left ventricular ejection fraction (LVEF (%))	30 ± 11.45
RAS (mm)	43.84 ± 9.54
Right ventricular diameter (RVD (mm))	24.13 ± 5.41
Right ventricular outflow tract (RVOT (mm))	31.28 ± 5.96
Interventricular septal diameter (IVSd (mm))	8.75 ± 1.45
Left ventricular posterior wall diameter (LVPWd (mm))	8.80 ± 1.18

2.3. Functional Enrichment Analysis. Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis of specified genes was carried out using the clusterProfiler package in R [13]. GO terms were composed of biological process, cellular component, and molecular function. Terms with adjusted p value < 0.05 were considered significantly enriched. The top ten terms were presented for each category.

2.4. A Protein-Protein Interaction (PPI) Network. Common DEGs were imported into the STRING database (version 11.0; https://string-db.org/) [14]. With the criteria of confidence > 0.7, the PPI network was visualized using the Cytoscape software (version 3.7.2; https://cytoscape.org/). GO and KEGG enrichment analysis was presented for genes in the PPI network.

2.5. Joint Analysis of RNA-seq and DNA Methylation. Upregulated and hypomethylated genes and downregulated and hypermethylated genes were obtained by joint analysis of DEGs and DEMs. The biological functions of differentially

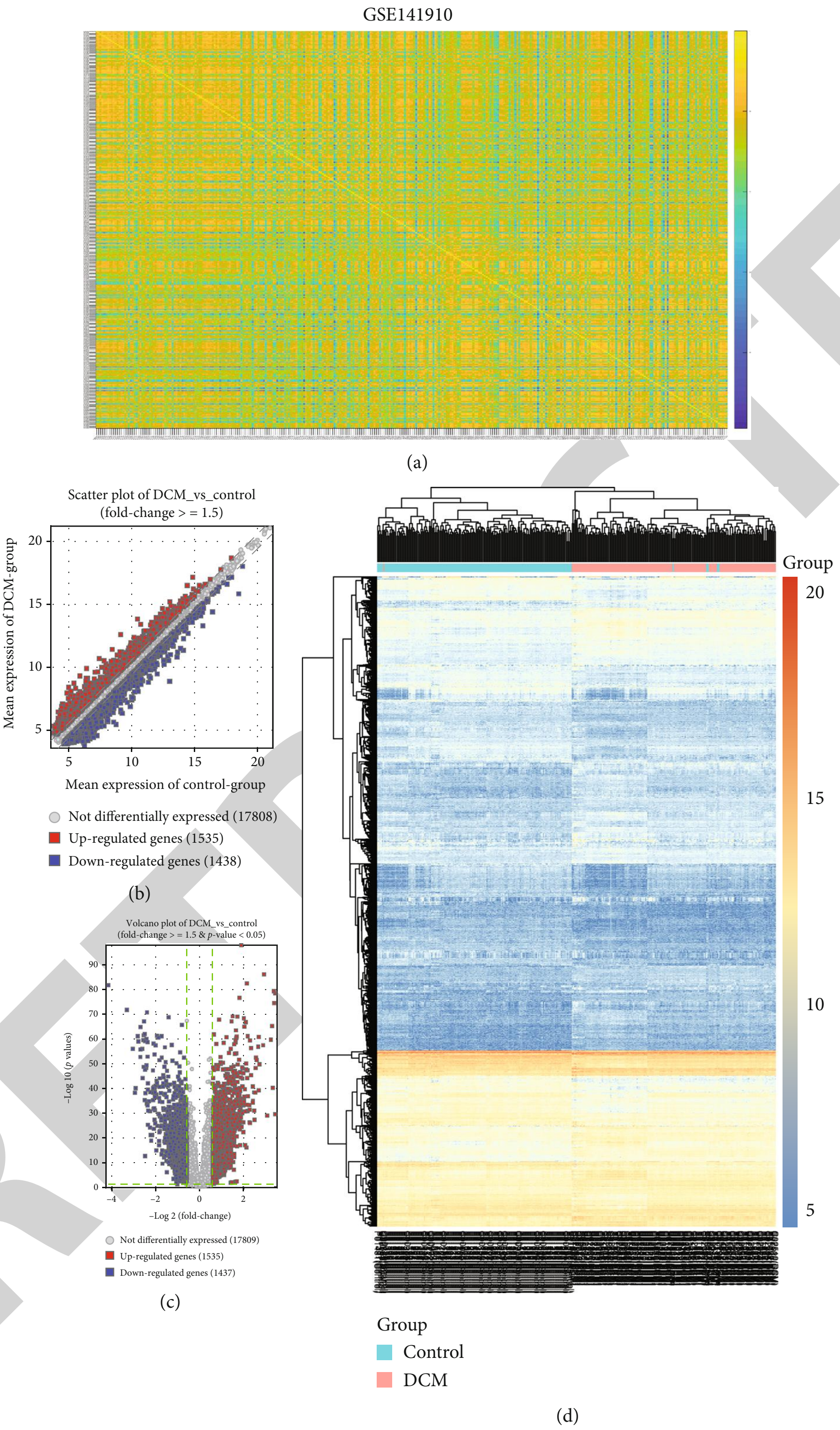
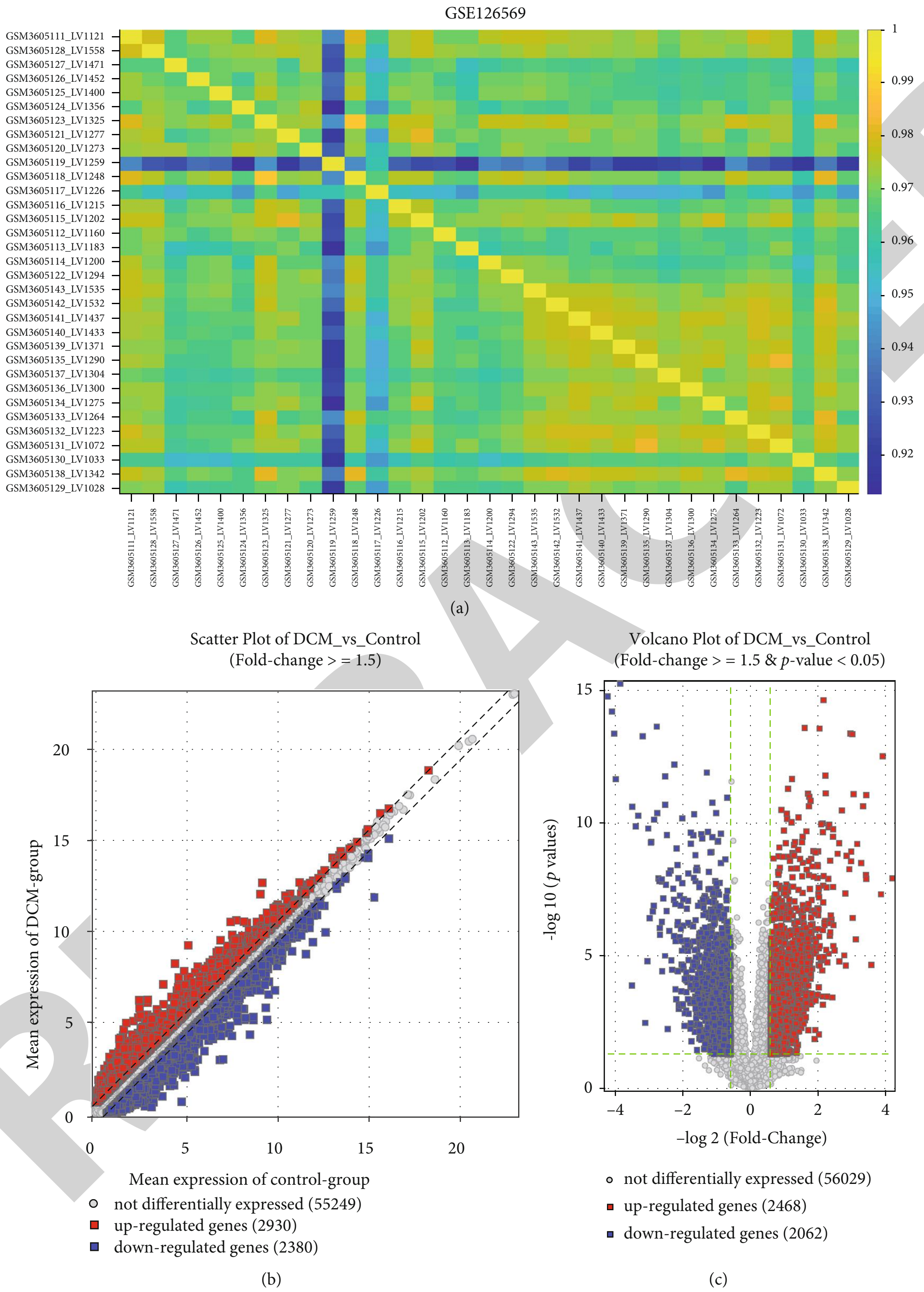


Figure 1: Differentially expressed genes between DCM and control left ventricle specimens in the GSE141910 dataset. (a) Heat map depicting the correlation between DCM and control left ventricle samples. The closer to yellow, the higher the correlation coefficient. (b) Scatter plots for up- and downregulated genes with |FC | ≥1.5 between DCM and control left ventricle specimens. (c) Volcano plots for DEGs with |FC | ≥1.5 and adjusted p value < 0.05 between DCM and control left ventricle specimens. (d) Hierarchical clustering heat map for DEGs between DCM and control groups. Red: upregulated genes; blue: downregulated genes; grey: no differentially expressed genes.

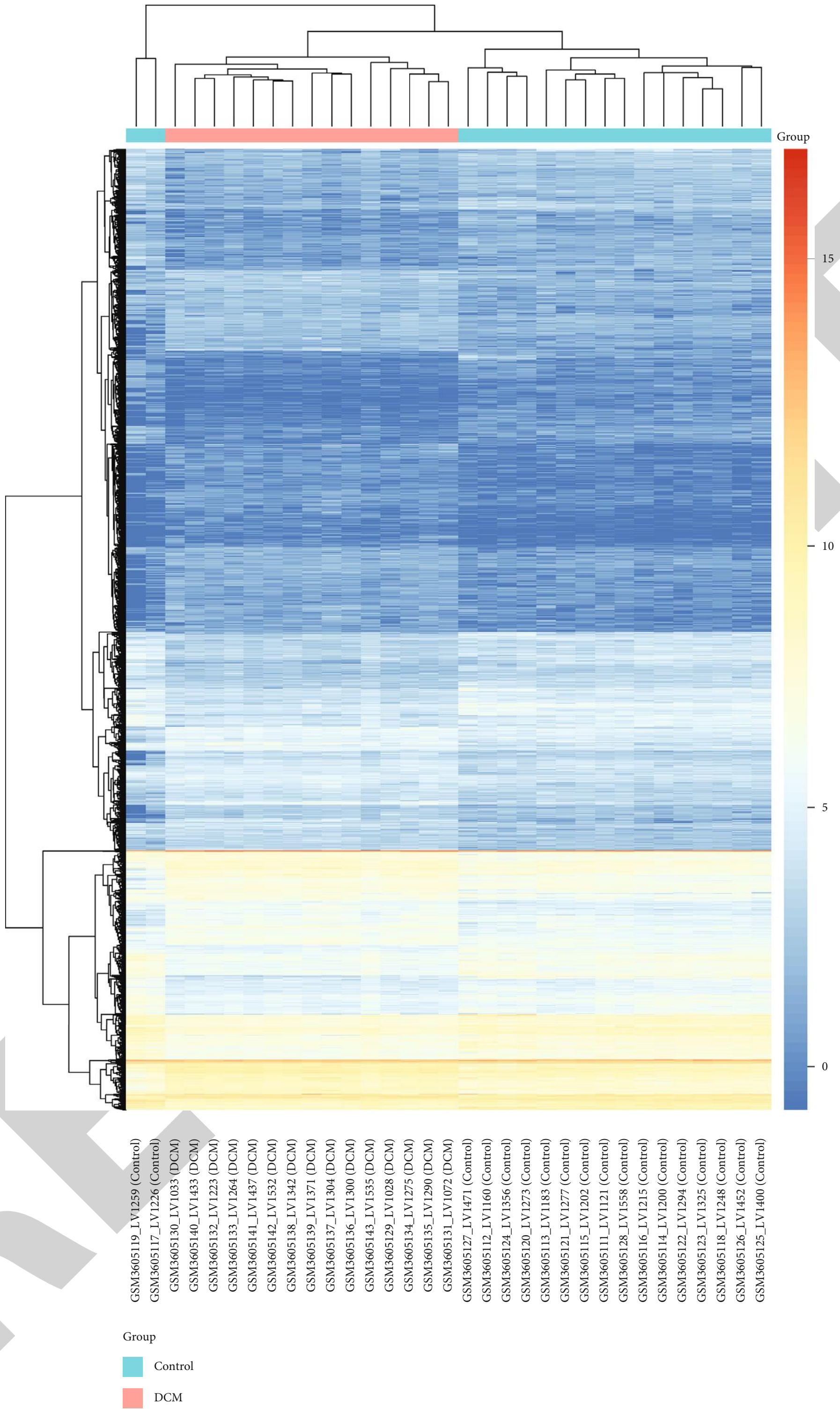


(a)

(b)

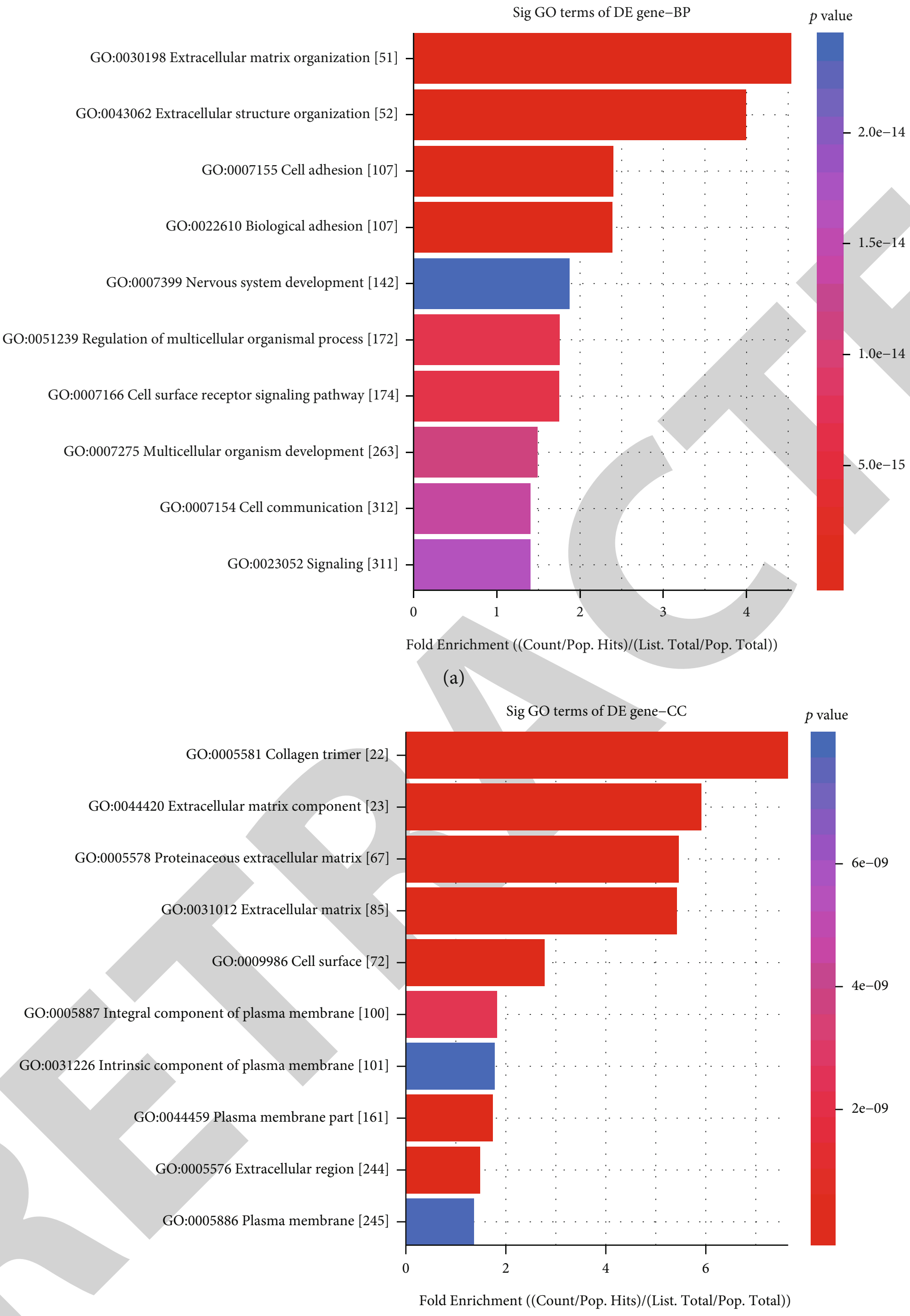
(c)

Figure 2: Continued.



(d)

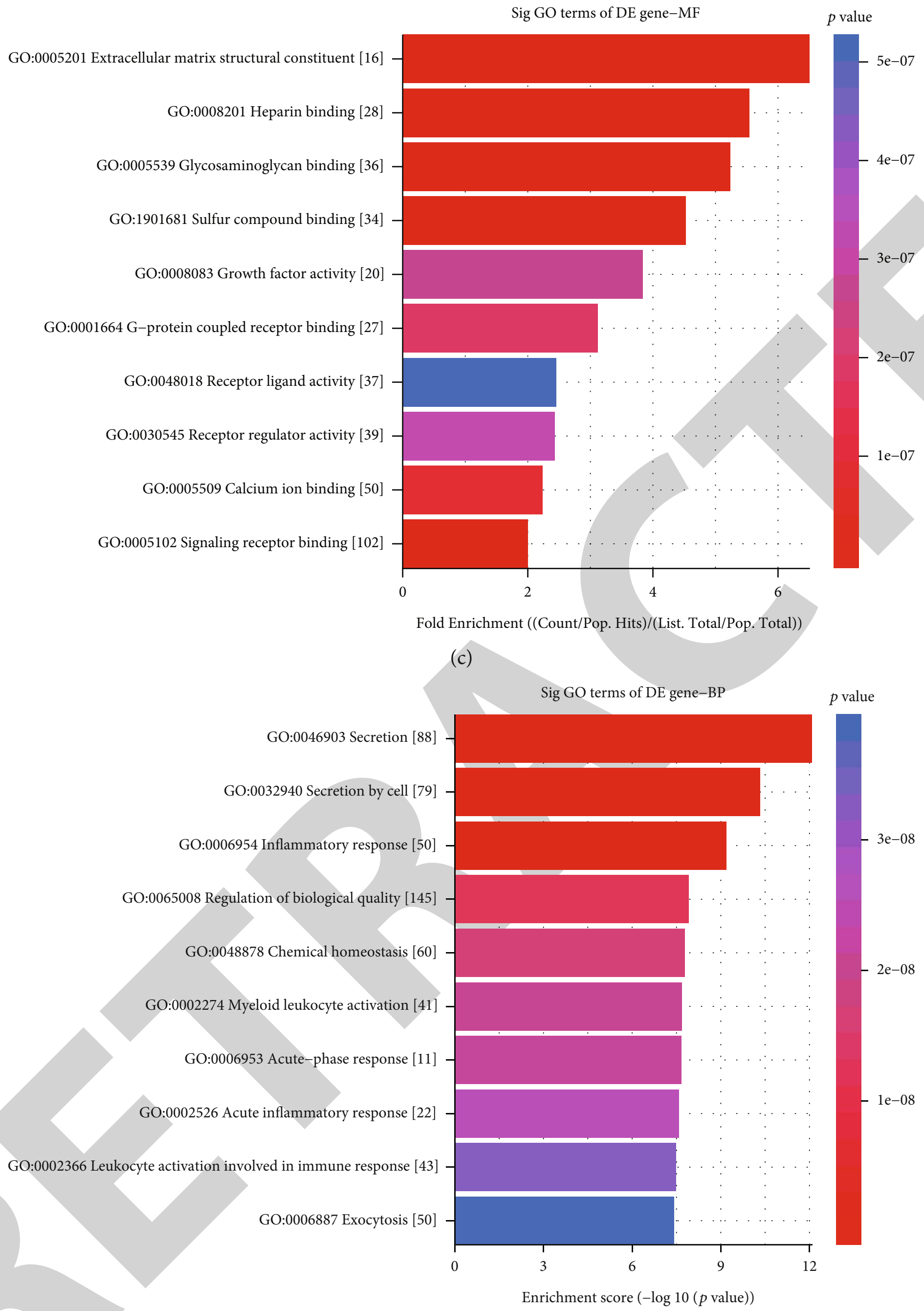
Figure 2: Differentially expressed genes between DCM and control left ventricle specimens in the GSE126569 dataset. (a) Heat map for the correlation between DCM and control left ventricle samples. The closer to yellow, the higher the correlation coefficient. (b) Scatter plots showing up- and downregulated genes with |FC | ≥1.5 between DCM and control groups. (c) Volcano plots for DEGs with |FC | ≥1.5 and adjusted p value < 0.05 between DCM and control groups. (d) Heat map for DEGs between DCM and control groups. Red: upregulated genes; blue: downregulated genes; grey: no differentially expressed genes.



(a)

(b)

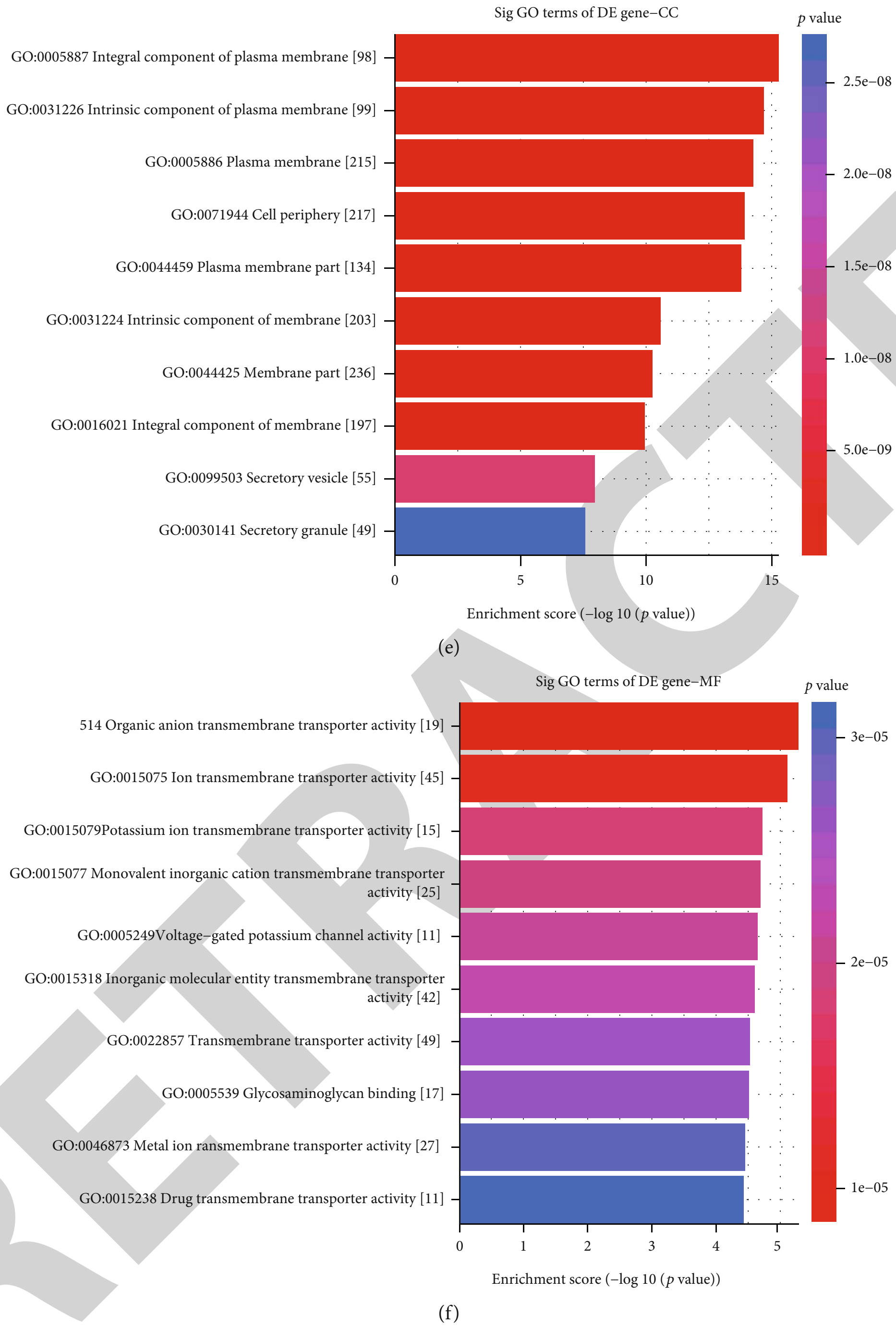
Figure 3: Continued.



(c)

(d)

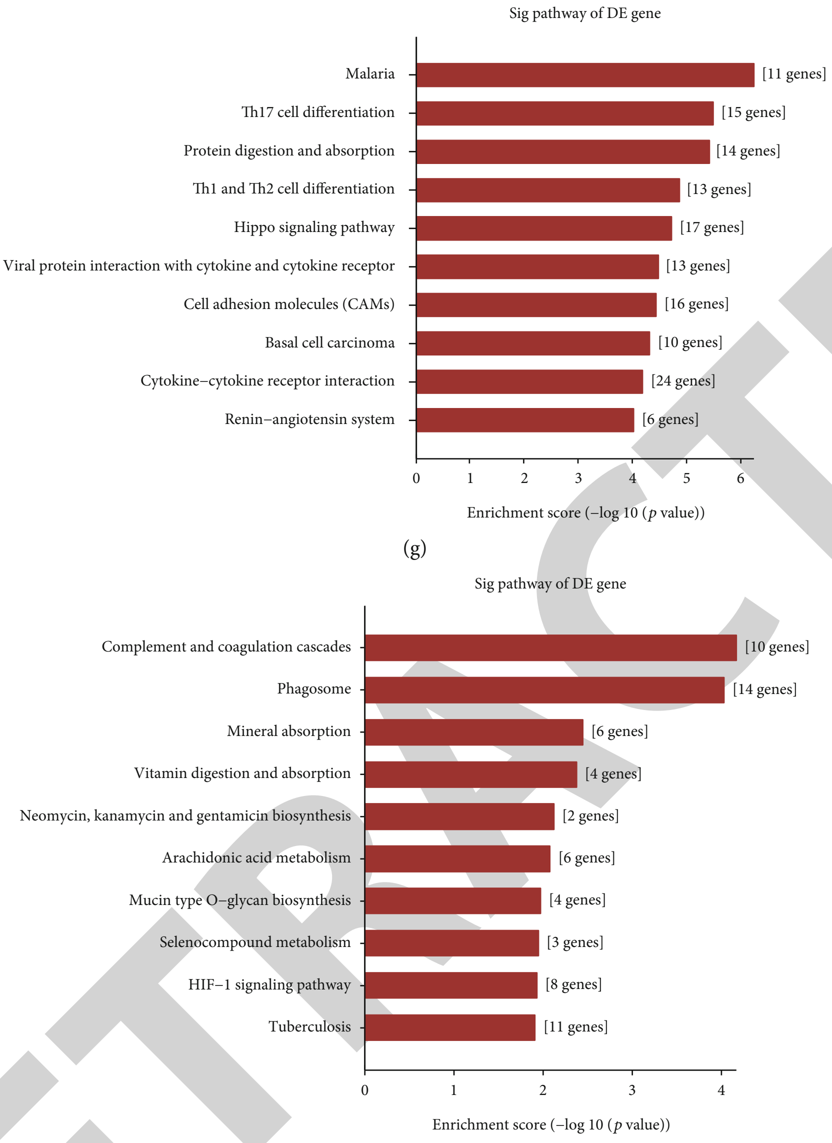
Figure 3: Continued.



(e)

(f)

Figure 3: Continued.



(g)

(h)

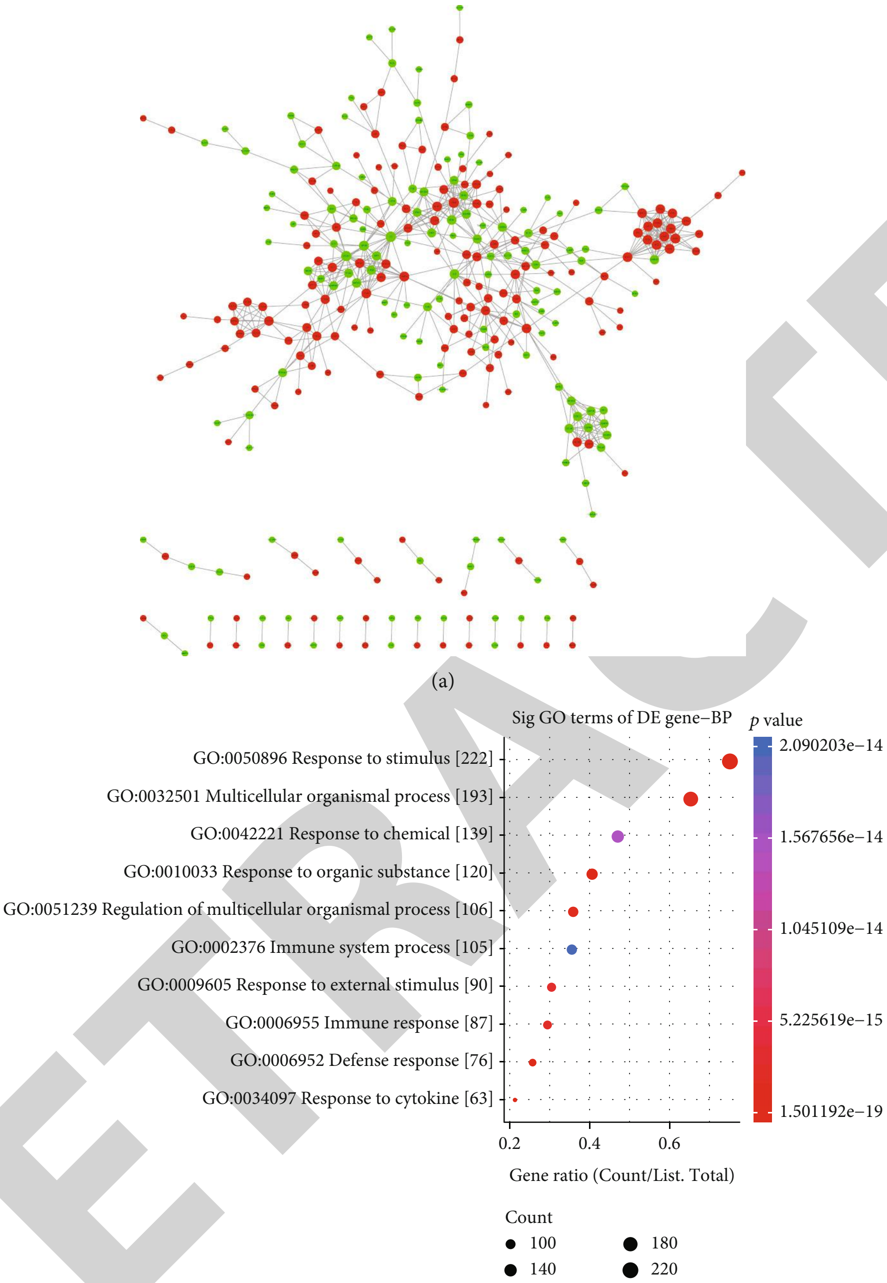
Figure 3: GO and KEGG enrichment analysis for common DEGs in the GSE141910 and GSE126569 datasets. The top ten GO enrichment analysis results for upregulated genes including biological process (a), cellular component (b), and molecular function (c). The top ten GO-biological process (d), cellular component (e), and molecular function (f) results for downregulated genes. (g, h) The top ten KEGG pathway enrichment analysis results for up- and downregulated genes, respectively.

expressed and methylated genes were analyzed by GO and KEGG enrichment analysis.

2.6. Whole Exome Sequencing (WES). Blood samples from 69 DCM patients were harvested from the Affiliated Hospital of Youjiang Medical University for Nationalities. The clinical features of these patients are listed in Table 1. WES was achieved by Wuhan Huada Medical Laboratory Co., Ltd. Our research was in line with the guidelines of the Declaration of Helsinki and was approved by the Ethics Committee of Affiliated Hospital of Youjiang Medical University for Nationalities (YYFY-LL-2016-01). All patients provided written informed consent. The following mutation data were filtered as follows: (1) 1000G_EAS mutation < 0.1; (2) homozygous mutation (Otherinfo = ‘ hom’); and (3) the mutation type that exhibited the greatest influence on the same gene for the same specimen. The selected mutation data were saved into the Mutation Annotation Format (maf) format, which were visualized using the maftools package in R [15].

2.7. Validation of Mutant Genes in Independent Datasets. RNA-seq of blood samples from 8 DCM patients and 8 healthy participants was obtained from the GSE101585 dataset on the GPL20301platform. The expression of the top 5 genes with the highest mutation frequency according to the WES results was validated in the GSE101585, GSE141910, and GSE126569 datasets.

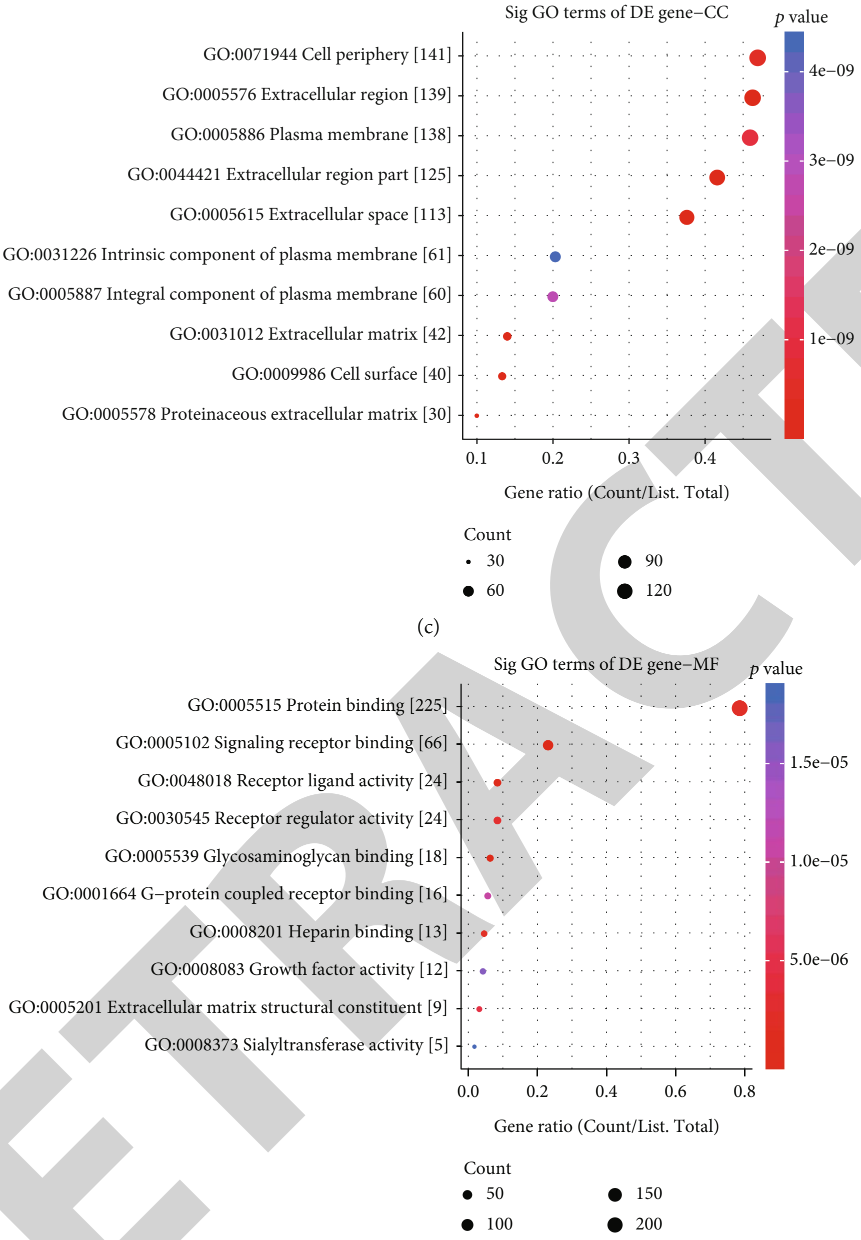
2.8. Statistical Analyses. Statistical analyses were carried out using R language packages (version 4.0.2; https://www.r-



(a)

(b)

Figure 4: Continued.



(c)

(d)

Figure 4: Continued.

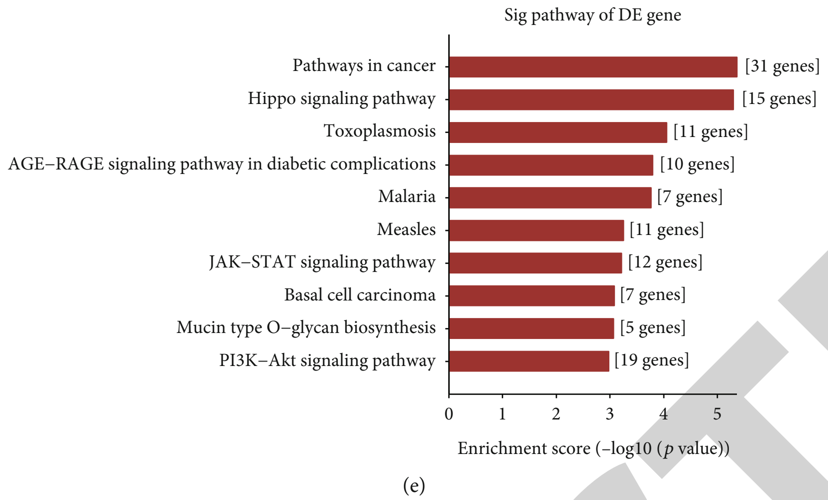


Figure 4: Construction of a PPI network for common DEGs and function annotation analysis. (a) A PPI network was established based on common DEGs. Red represents upregulation and green represents down-regulation. Nodes are proportional to the size of the circle. The top ten GO enrichment analysis results for genes in the network including biological process (b), cellular component (c), and molecular function (d). (e) The top ten KEGG pathway enrichment analysis results for genes in the network.

project.org/). Differences with p value < 0.05 were statistically significant.

3. Results

3.1. Screening DEGs for DCM. Two datasets including GSE141910 and GSE126569 were included for screening DEGs between DCM and control left ventricle specimens. In the GSE141910 dataset, there were high correlations between 166 healthy and 166 DCM specimens (Figure 1(a)). With the threshold of $|FC| \geq 1.5$, 1535 genes were upregulated and 1438 were downregulated in DCM compared to control samples (Figure 1(b)). DEGs including 1535 up- and 1437 downregulated genes were screened for DCM under the criteria of $|FC| \geq 1.5$ and adjusted p value < 0.05 (Figure 1(c)). Heat map depicted that these DEGs distinctly distinguished DCM samples from normal samples (Figure 1(d)). In the GSE126569 dataset, there were 18 healthy and 15 DCM left ventricle samples. The high correlation was found between these samples (Figure 2(a)). 2930 genes were upregulated, and 2380 genes were downregulated in DCM than in control specimens (Figure 2(b)). As shown in Figure 2(c), we identified DEGs between the two groups, composed of 2468 up- and 2062 downregulated genes. The differences in expression patterns of these DEGs are depicted in Figure 2(d).

3.2. Potential Biological Functions of Common DEGs. After overlapping the DEGs in the GSE141910 and GSE126569 datasets, 1407 common DEGs were identified for DCM. Potential biological functions of up- and downregulated genes were separately analyzed. Our GO enrichment analysis results showed that upregulated genes were significantly enriched in extracellular matrix organization (Figure 3(a)), collagen trimer (Figure 3(b)), and extracellular matrix structural constituent (Figure 3(c)). Meanwhile, downregulated genes were distinctly enriched in secretion (Figure 3(d)), integral and intrinsic component of plasma membrane (Figure 3(e)), and organic anion transmembrane transporter activity (Figure 3(f)). KEGG enrichment analysis results demonstrated that upregulated genes were significantly associated with Th17 cell differentiation, protein digestion and absorption, Th1 and Th2 cell differentiation, Hippo signaling pathway, cytokine and cytokine receptor, and cell adhesion molecules (Figure 3(g)). In Figure 3(h), downregulated genes were significantly enriched in complement and coagulation cascades as well as phagosome.

3.3. A PPI Network Based on Common DEGs. We further analyzed the interactions between common DEGs by the STRING database. The interactions were visualized into a PPI network via the Cytoscape software. As a result, there were 307 nodes in the PPI network, composed of 171 up- and 136 downregulated genes (Figure 4(a)). The top four genes with the highest degree were selected as hub genes, including C3 (degree = 24), GNB3 (degree = 23), QSOX1 (degree = 21), and APOB (degree = 17). The expression of C3, QSOX1, and APOB was significantly upregulated, and GNB3 was distinctly downregulated in DCM compared to controls (Figure 4(a)). GO enrichment analysis results indicated that the genes in the PPI network were distinctly enriched in response to stimulus (Figure 4(b)), cell periphery (Figure 4(c)), extracellular region (Figure 4(c)), and protein binding (Figure 4(d)). KEGG pathway enrichment analysis revealed that these genes had significant associations with pathways in cancer, Hippo, JAK-STAT, and PI3K-Akt signaling pathways (Figure 4(e)).

3.4. Identification of DMGs for DCM. We further analyzed the DNA methylation for DCM using the GSE42510 dataset. There were distinct correlations between 8 control and 9 DCM specimens (Figure 5(a)). With the threshold of $|FC| \geq 1.5$, 1122 hypermethylated and 1314 hypomethylated genes were screened between DCM and control samples (Figure 5(b)). Differentially methylated genes with fold-change ≥ 1.5 and adjusted p value < 0.05 were identified for DCM compared to controls, including 285 hypermethylated

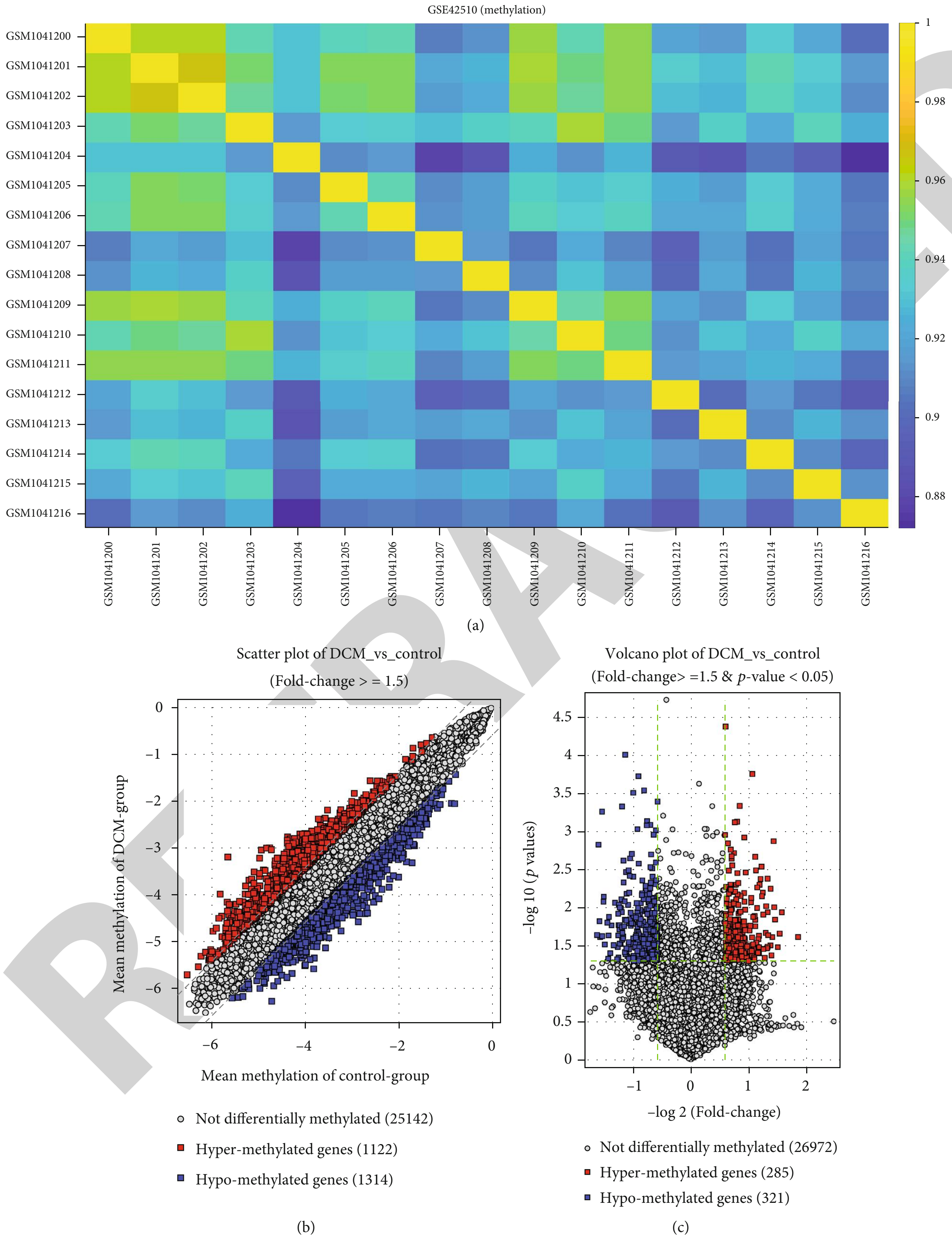
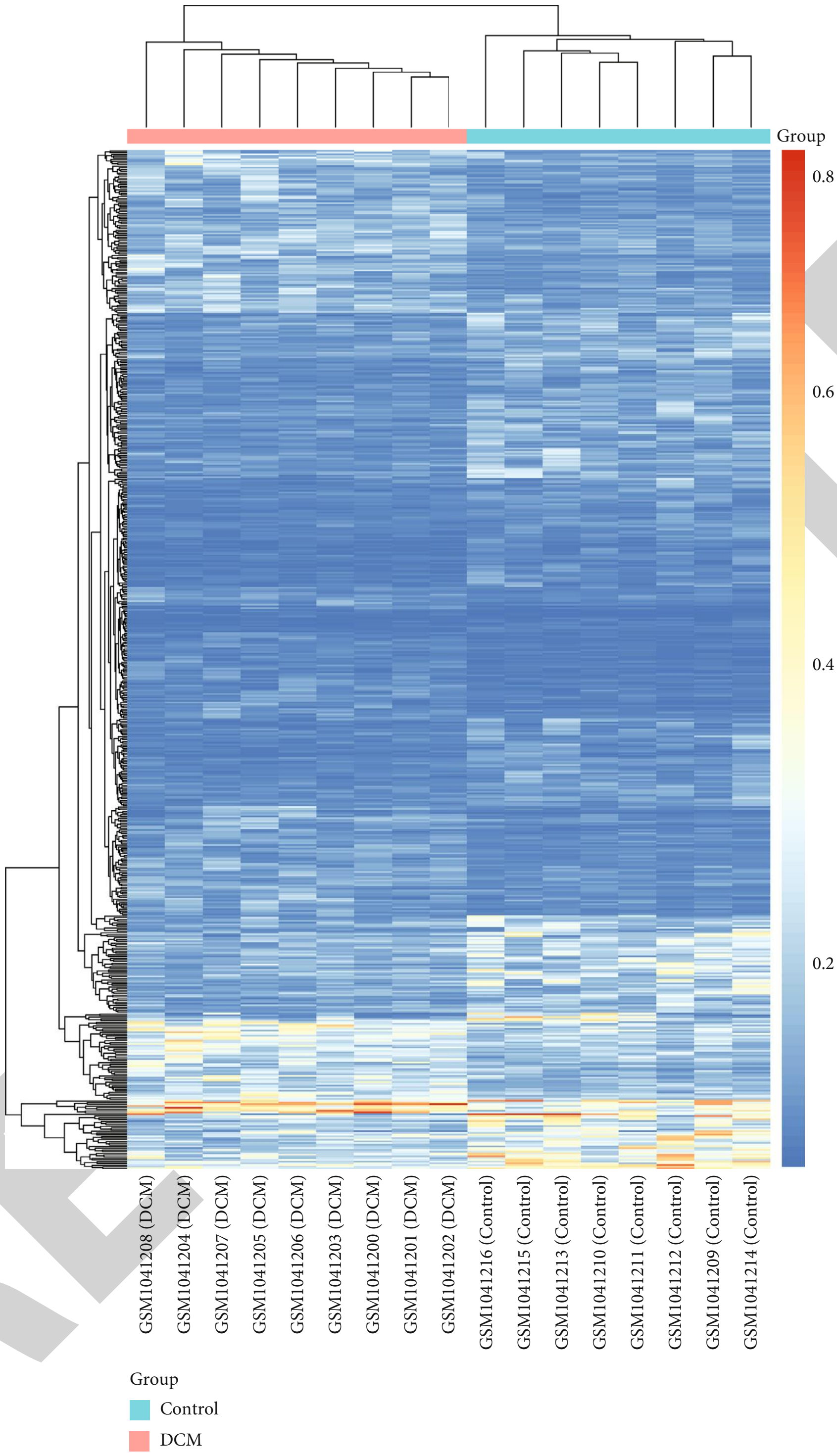


Figure 5: Continued.

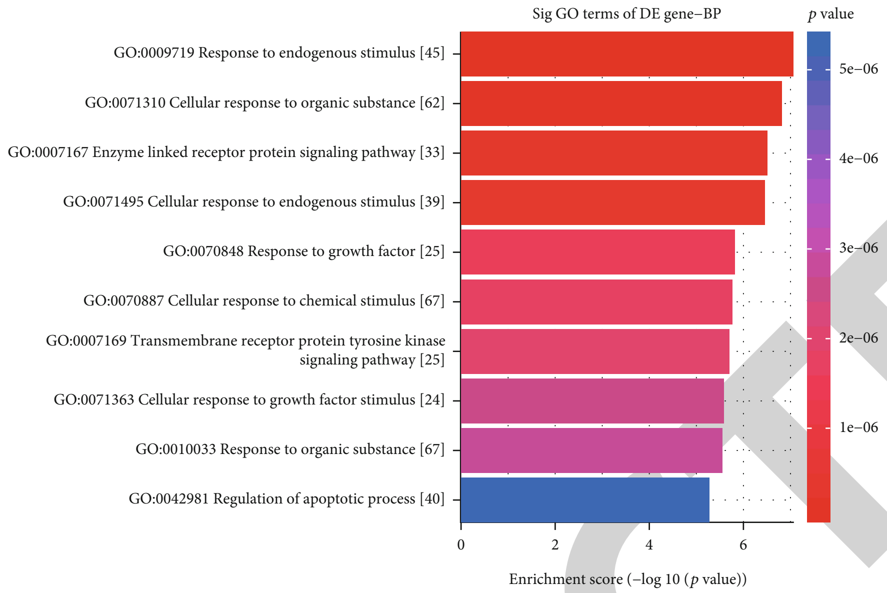


(d)

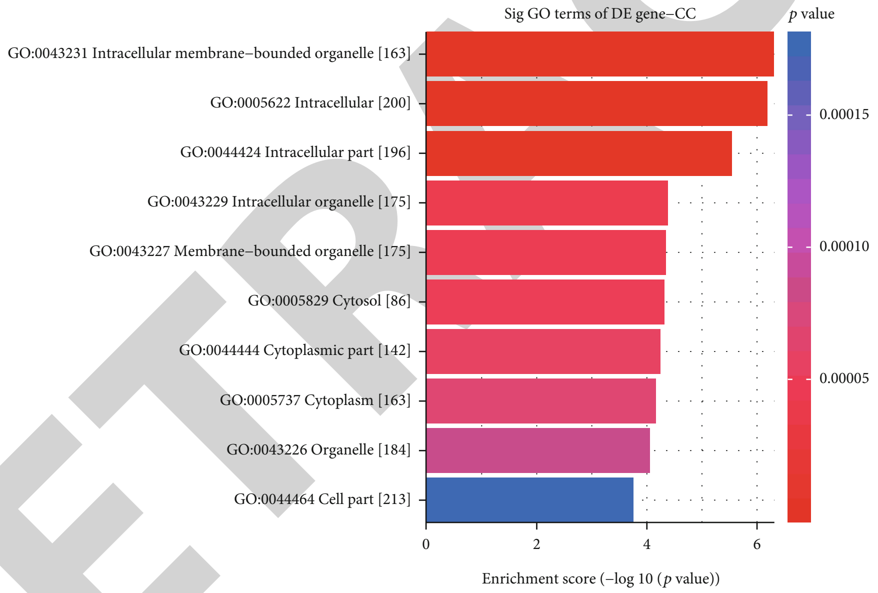
Figure 5: Differentially methylated genes between DCM and control left ventricle tissues in the GSE42510 dataset. (a) Heat map showing the correlation between DCM and control samples. The closer to yellow, the higher the correlation coefficient. (b) Scatter plots depicting hyper- and hypomethylated genes with |FC | ≥1.5 between DCM and control groups. (c) Volcano plots for differentially methylated genes with | FC | ≥1.5 and adjusted p value < 0.05 between DCM and control groups. (d) Heat map for differentially methylated genes between the two groups. Red: hypermethylated genes; blue: hypomethylated genes; grey: no differentially methylated genes.

and 321 hypomethylated genes (Figure 5(c)). The distinct differences in their methylation levels were found between DCM and control groups (Figure 5(d)).

3.5. Exploring Underlying Biological Functions of DMGs. We analyzed which biological processes the DEMs were mainly involved in. GO enrichment analysis results showed that

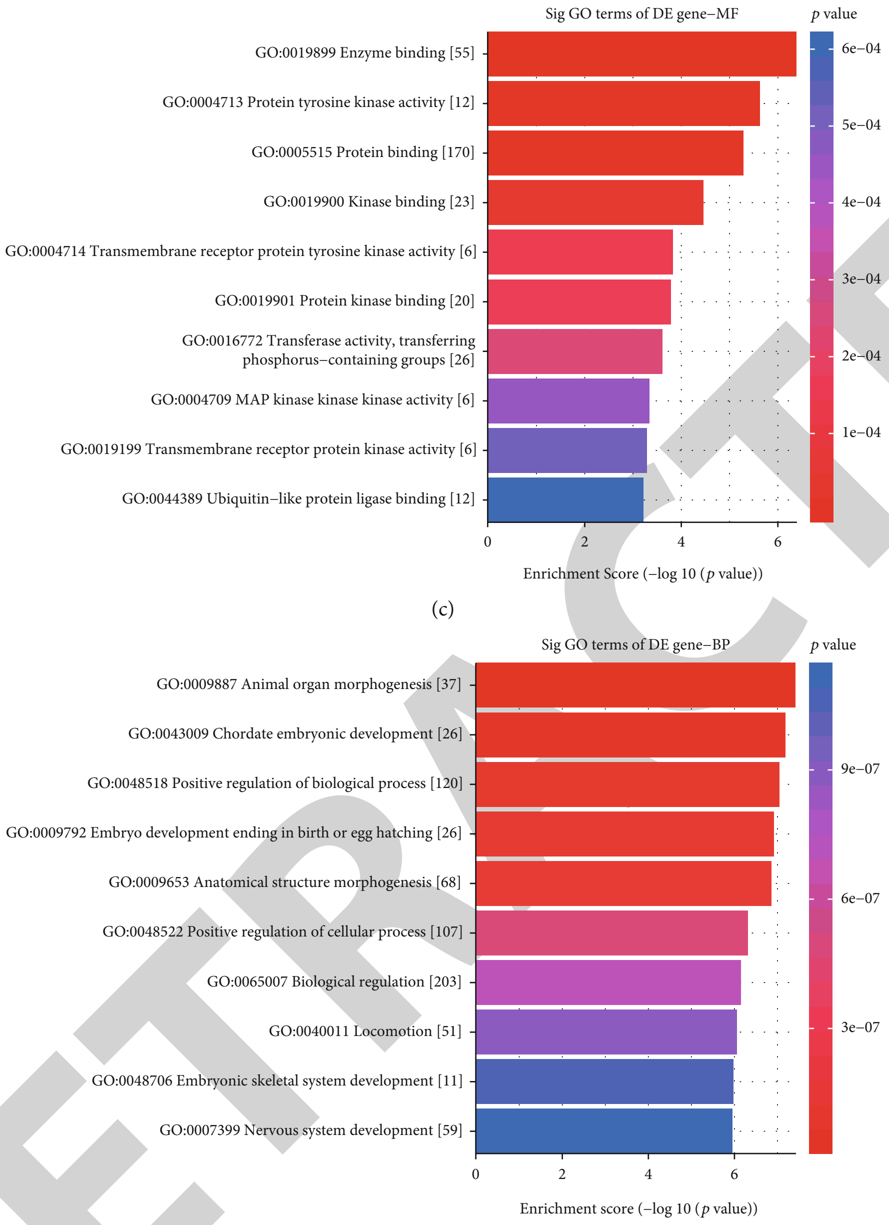


(a)



(b)

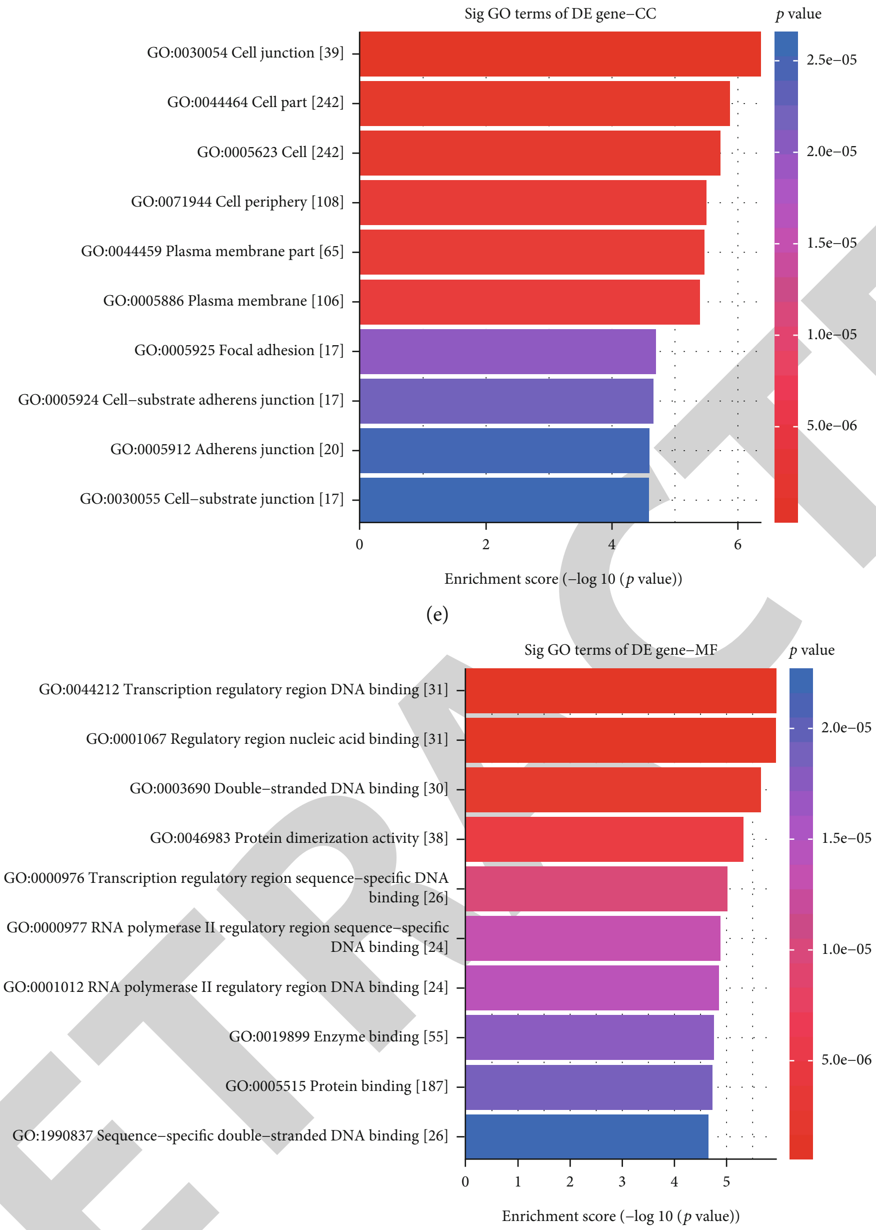
Figure 6: Continued.



(c)

(d)

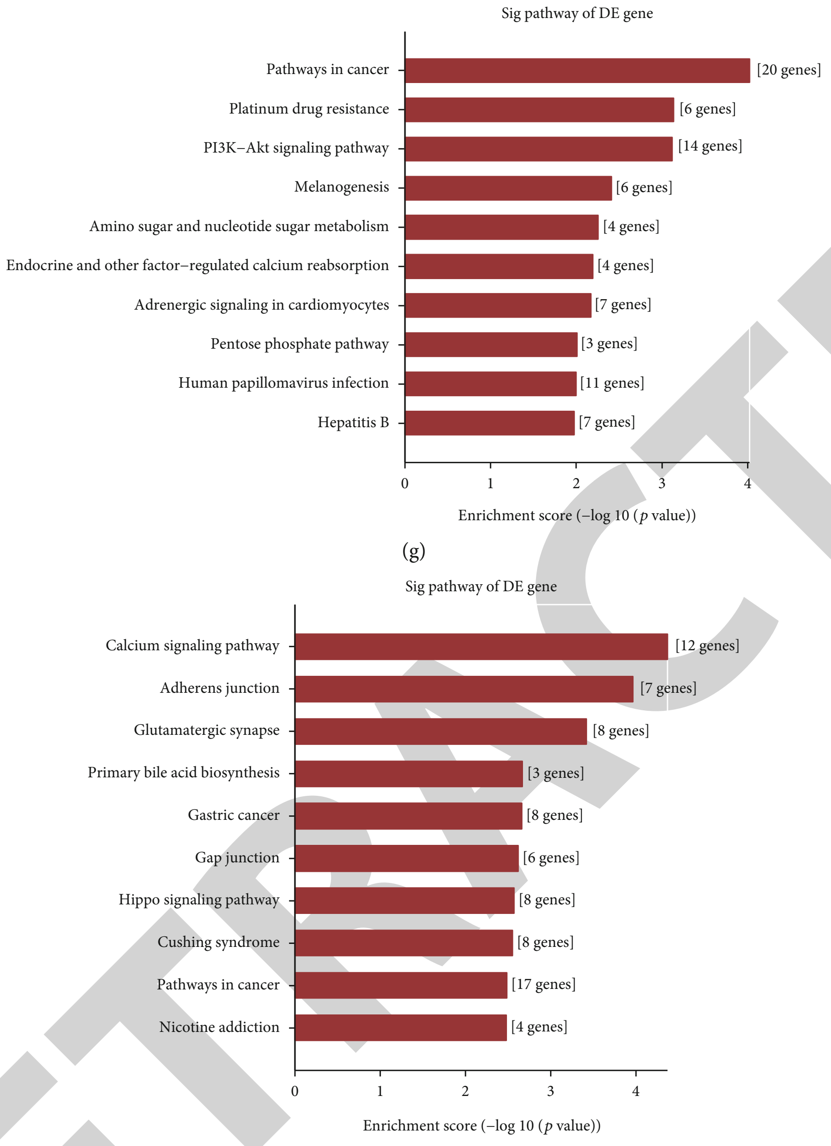
Figure 6: Continued.



(e)

(f)

Figure 6: Continued.



(g)

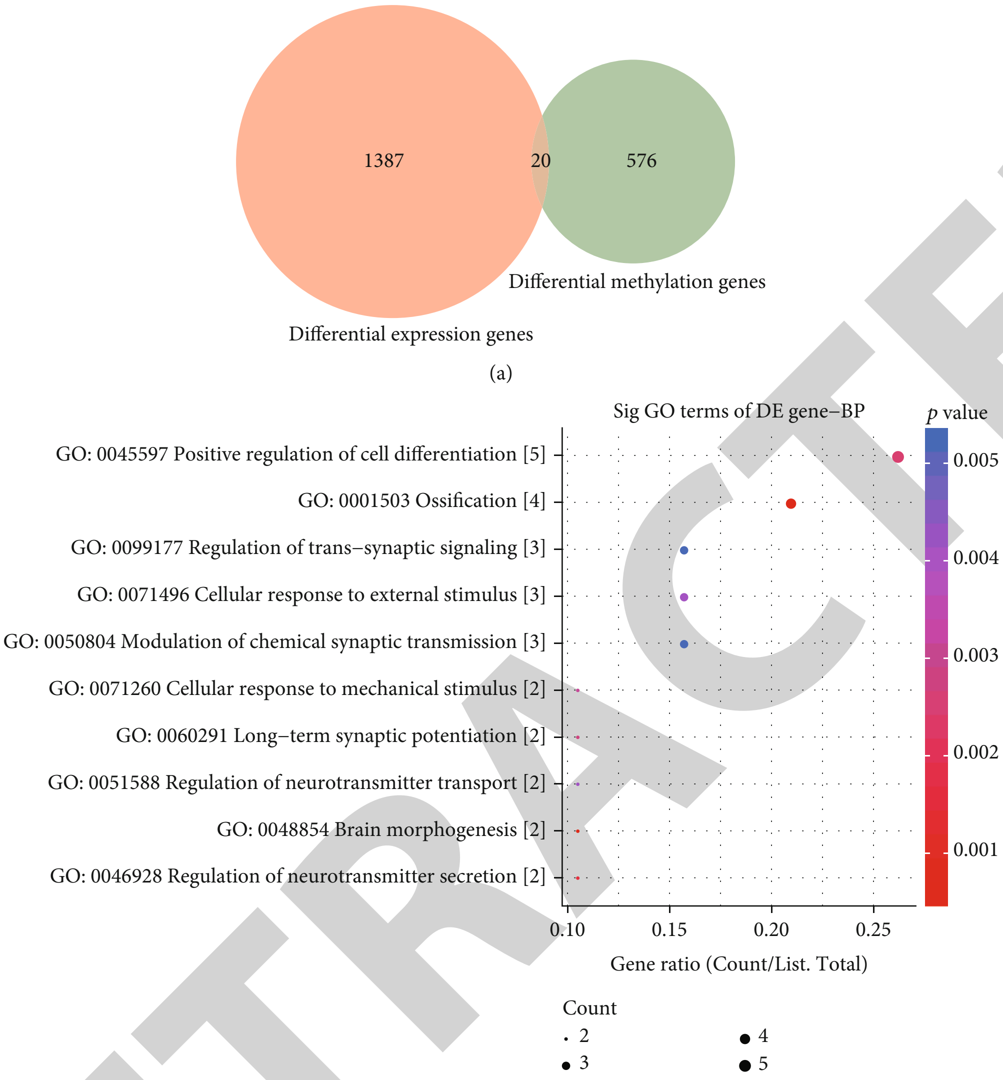
(h)

Figure 6: GO and KEGG enrichment analysis of differentially methylated genes. The top ten GO enrichment analysis results for hypermethylated genes including biological process (a), cellular component (b), and molecular function (c). The top ten GO-biological process (d), cellular component (e), and molecular function (f) results for hypomethylated genes. (g, h) The top ten KEGG pathway enrichment analysis results for hyper- and hypomethylated genes.

hypermethylated genes were significantly enriched in response to endogenous stimulus (Figure 6(a)), intracellular membrane-bounded organelle (Figure 6(b)), and enzyme binding (Figure 6(c)). Hypomethylated genes were distinctly associated with animal organ morphogenesis (Figure 6(d)), chordate embryonic development (Figure 6(d)), cell junction (Figure 6(e)), transcription regulatory region DNA binding (Figure 6(f)), and regulatory region nucleic acid binding (Figure 6(f)). As shown in KEGG pathway enrichment analysis, hypermethylated genes were mainly enriched in pathways in cancer, platinum drug resistance, and PI3K-Akt signaling pathway (Figure 6(g)). In Figure 6(h), hypomethylated genes were primarily enriched in calcium signaling pathway, adherens junction, and glutamatergic synapse.

3.6. Differentially Expressed and Methylated Genes for DCM. Conjoint analysis of DEGs and DEMs was further presented for DCM. As shown in Figure 7(a), 20 differentially expressed and methylated genes were identified between DCM and control left ventricle samples. GO enrichment analysis results showed that these differentially expressed and methylated genes were distinctly correlated to positive regulation of cell differentiation (Figure 7(b)), synapse (Figure 7(c)), and growth factor binding (Figure 7(d)). KEGG pathway enrichment analysis results indicated that these genes were significantly enriched in protein digestion and absorption (Figures 7(e) and 7(f)) and amoebiasis (Figures 7(e) and 7(g)).

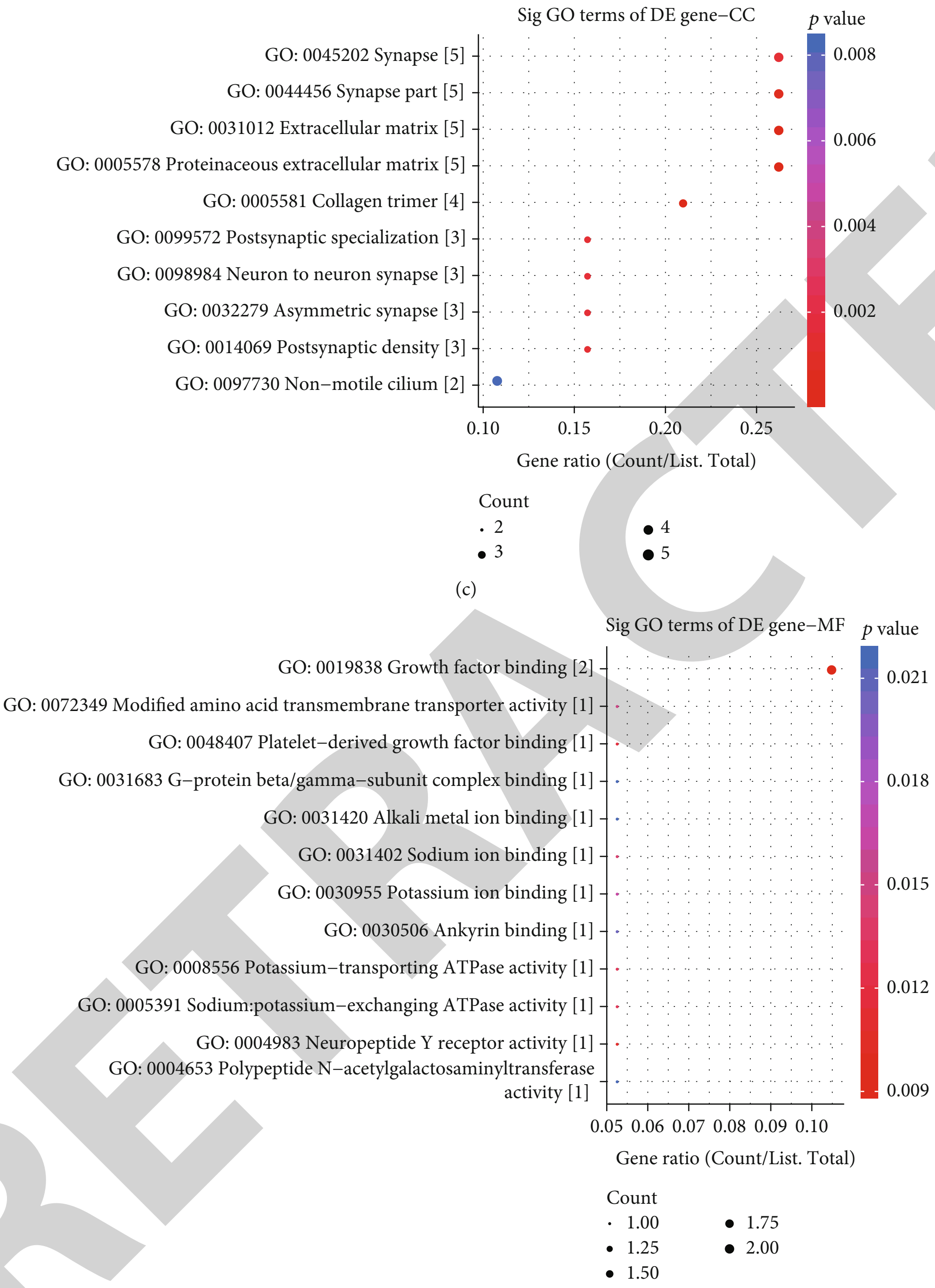
3.7. Single-Nucleotide Polymorphisms (SNP) in DCM. 69 DCM blood samples were analyzed using WES. Missense



(a)

(b)

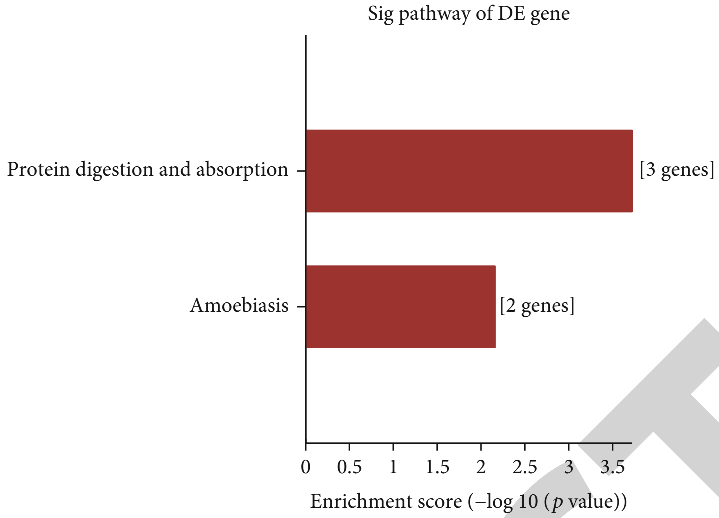
Figure 7: Continued.



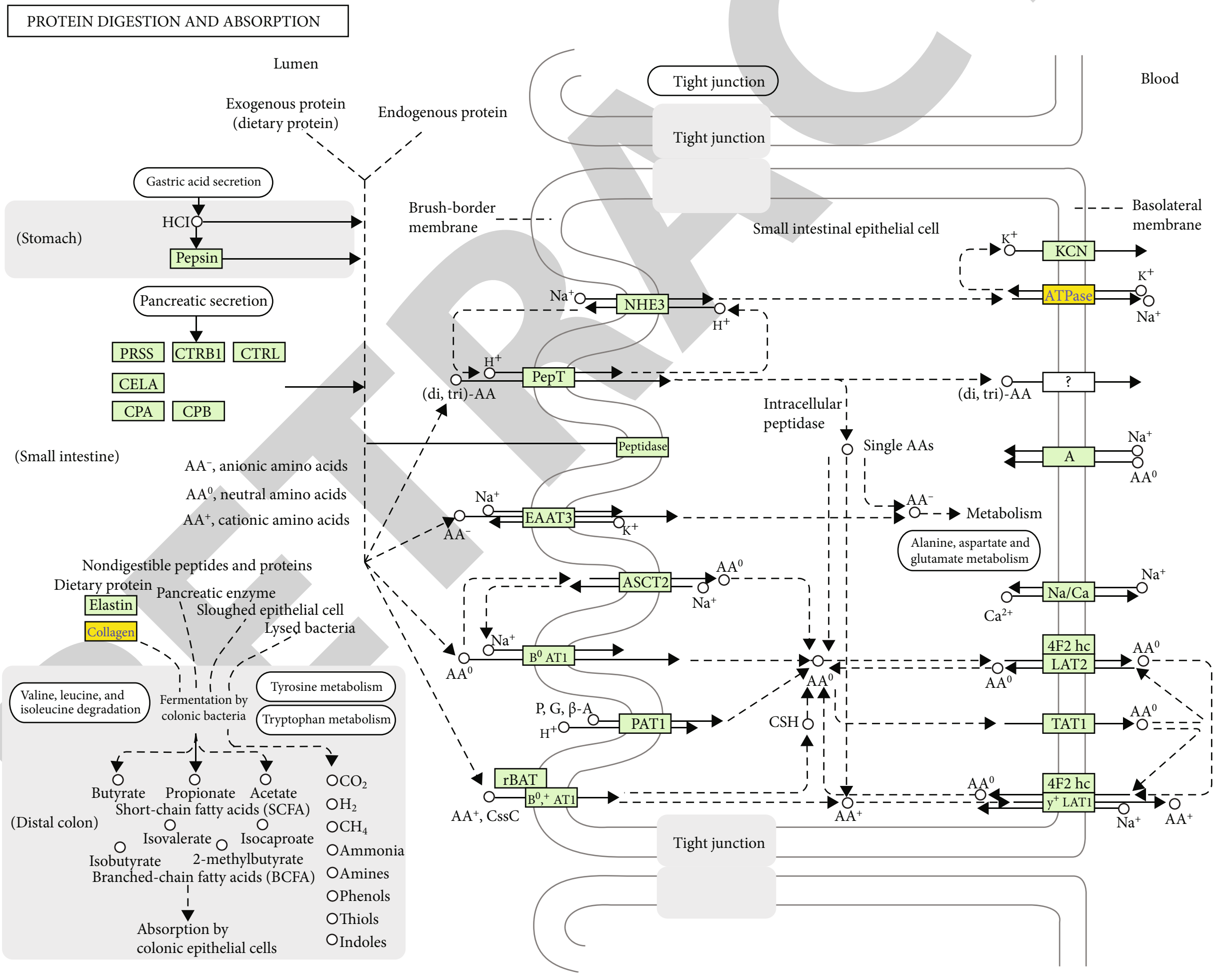
(c)

(d)

Figure 7: Continued.

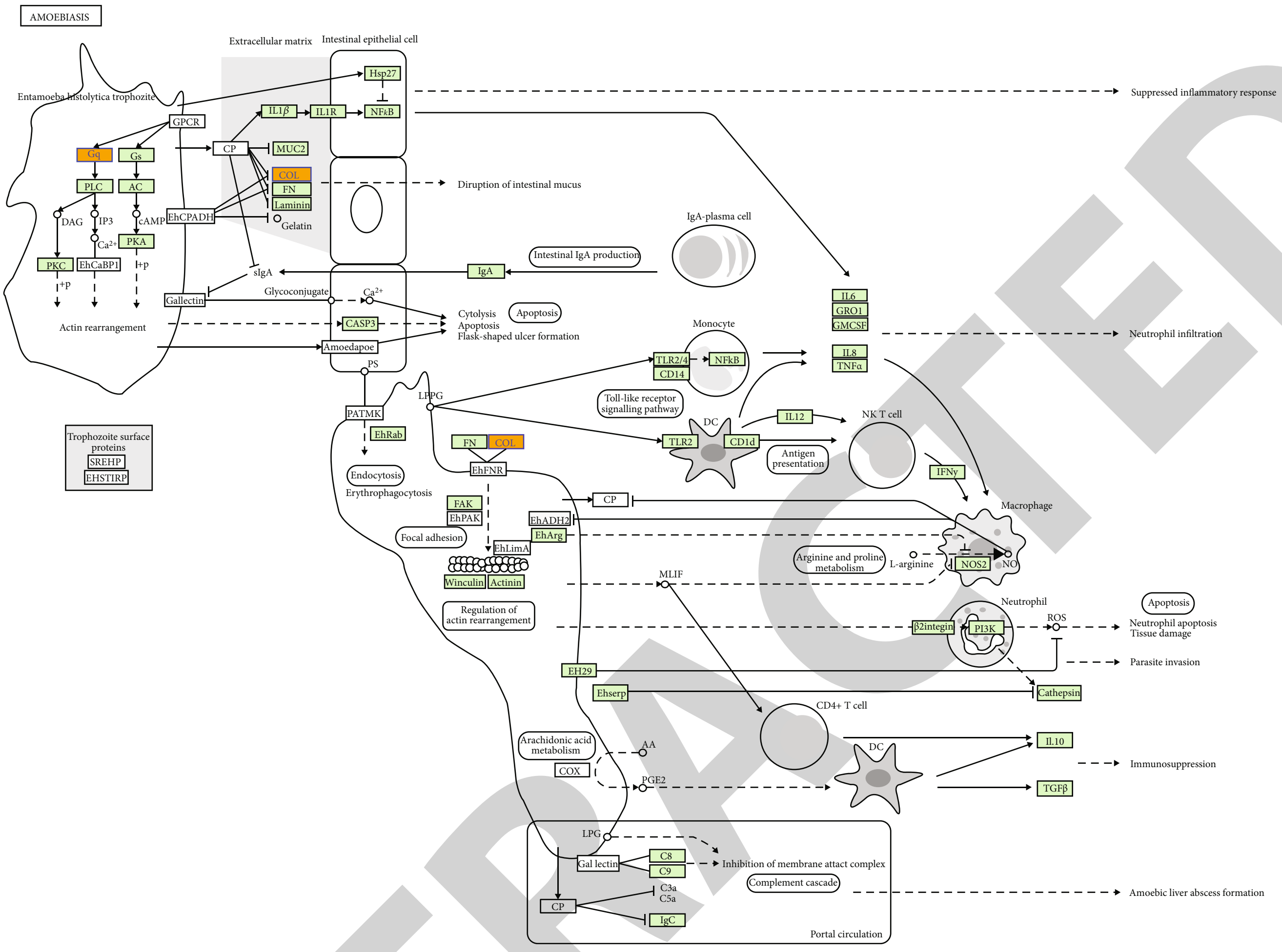


(e)



(f)

Figure 7: Continued.



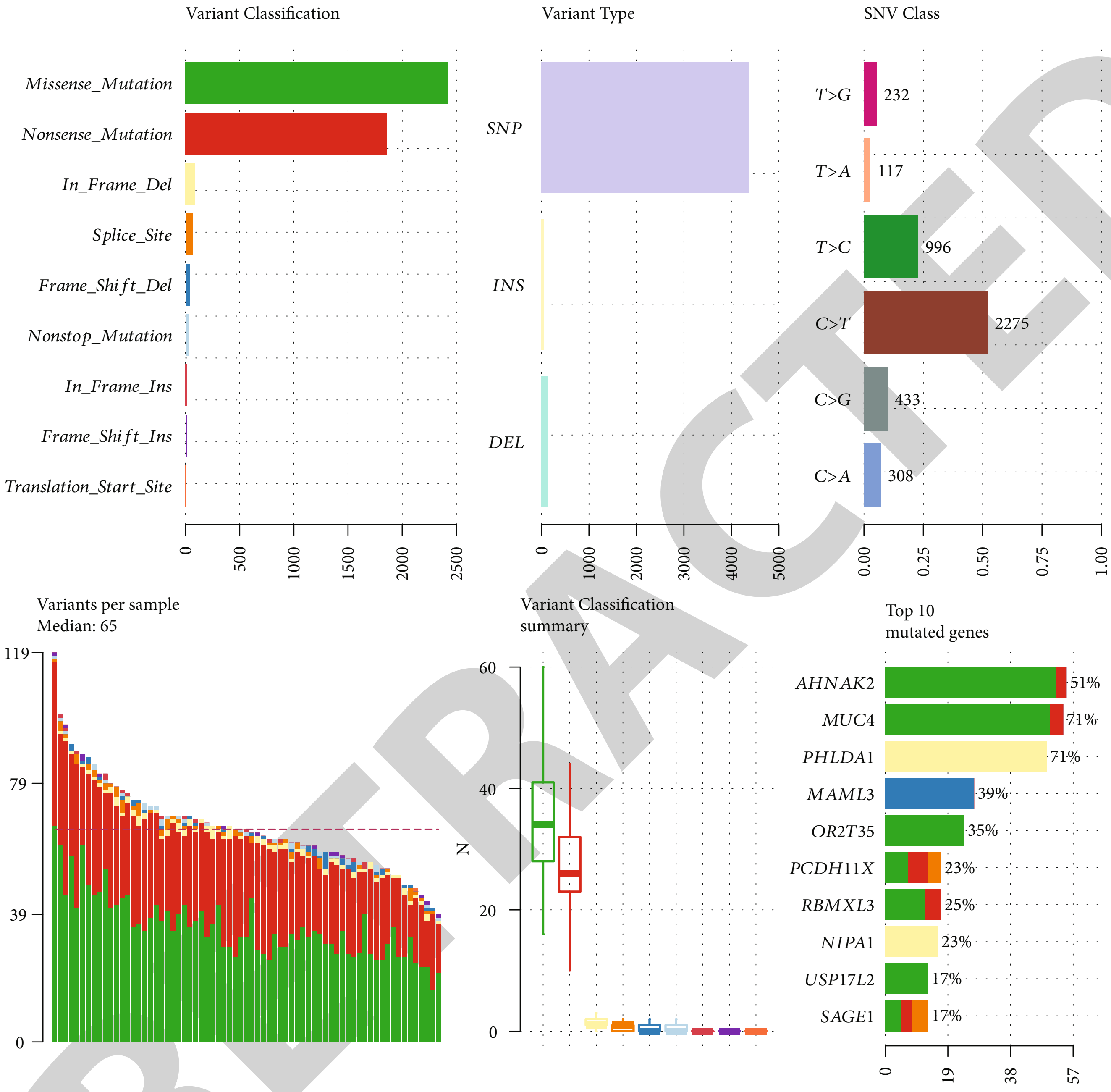
(g)

Figure 7: Conjoint analysis of RNA-seq and DNA methylation for DCM. (a) Venn diagram depicting the 20 differentially expressed and methylated genes between DCM and control left ventricle samples. The top ten GO-biological process (b), cellular component (c), and molecular function (d) results for differentially expressed and methylated genes. (e) KEGG enrichment analysis including (f) protein digestion and absorption and (g) amoebiasis.

mutation and nonsense mutation were the top two variant classification (Figure 8(a)). Table 2 lists the mutation frequency about different mutation types among these DCM samples. SNP was the most common variant type. C>T was the most common single-nucleotide variant (SNV) classification, followed by T>C. The median variants per sample were 65. The five mutated genes were as follows: AHNAK2, MUC4, PHLDA1, MAML3, and OR2T35. Figure 8(b) visualizes the top 100 genes with the highest mutation frequency, such as PHLDA1 and MUC4. In Figure 8(c), MUC4 (mainly missense mutation) and PHLDA1 (mainly in frame deletion) occupied the highest frequency mutation among 69 DCM samples (71%). We further explored the correlations in cooccurrence between different mutated genes. As shown in Figure 8(d), among 69 DCM blood samples, IRS4 mutation was positively correlated with COL4A5 mutation. MUC4 mutation had a significant correlation with PCDH11X mutation. PCDH11X mutation exhibited a distinct association with CYLC1 mutation.

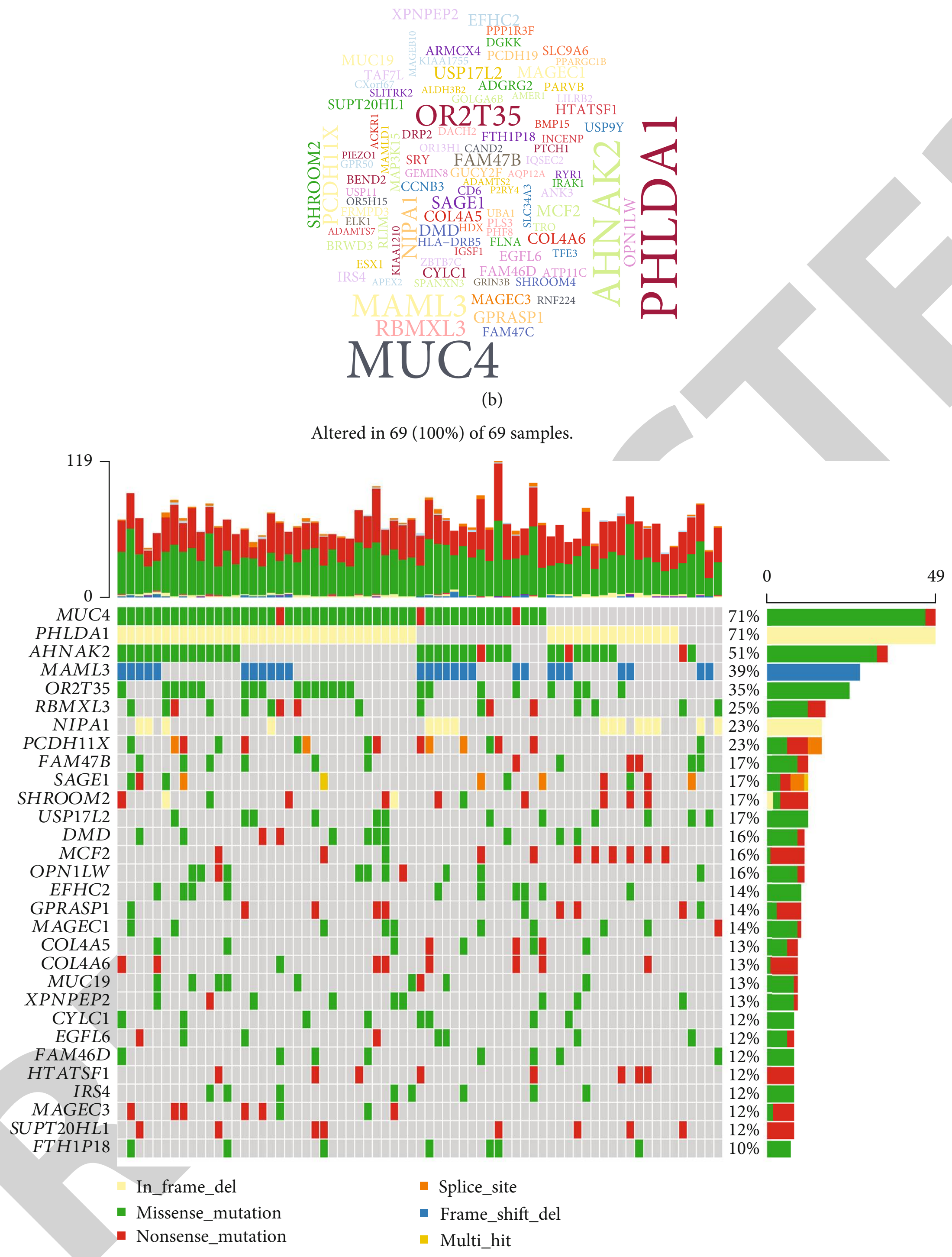
3.8. *Validation of the Expression of Mutant Genes in the Blood and Left Ventricle of DCM.*

The expression of mutant genes was detected and validated in blood and left ventricle samples of DCM patients and controls. In the GSE101585 dataset, there were distinct differences between DCM and control blood samples based on the gene expression profiles after preprocessing (Figure 9(a)). Heat map visualized the correlations between different samples at the mRNA expression levels (Figure 9(b)). In Figure 9(c), we found that there were 2517 up- and 3987 downregulated genes with |FC|>2 in DCM compared to control groups. With the threshold of |FC|>2 and adjusted p value < 0.05, 146 up- and 675 downregulated genes were identified for DCM (Figure 9(d)). There were distinct differences in the expression of these DEGs between DCM and control groups (Figure 9(e)). The mutant genes including AHNAK2, MAML3, MUC4, OR2T35, and PHLDA1 were differentially expressed in DCM compared to control blood samples (Figure 9(f)). In the GSE141910 dataset, AHNAK2 (p value = $3.7e-15$) was lowly expressed,



(a)

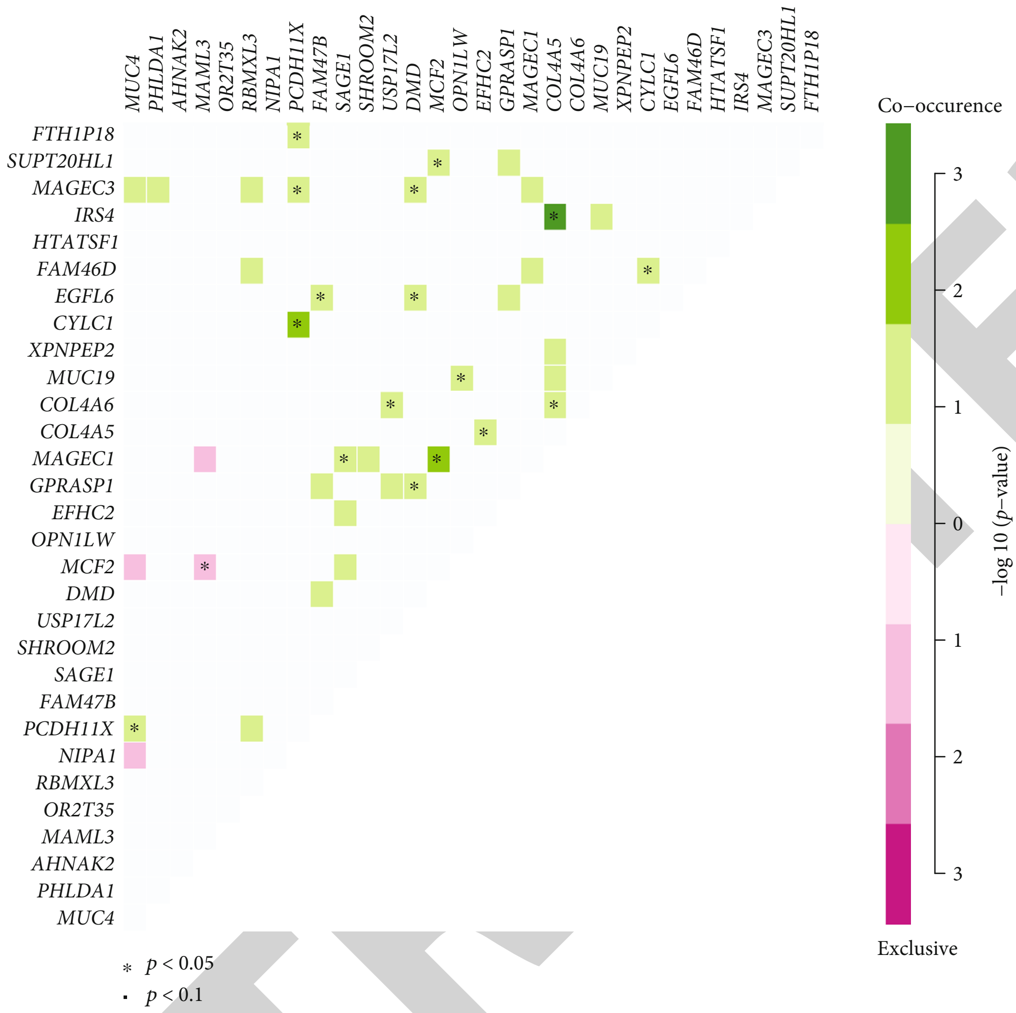
Figure 8: Continued.



(b)

(c)

Figure 8: Continued.



(d)

Figure 8: Landscape of mutation for DCM. (a) Mutation classification, type, SNV class, variants per sample, variant classification summary, and the top ten mutated genes for DCM. (b) Map of the top 100 genes with mutation frequencies. The larger the font, the higher the mutation frequency. (c) Waterfall diagram depicting the mutations of the top 30 genes in each sample. (d) Correlation analysis among the top 30 genes in mutation frequency.

Table 2: WES results for 69 DCM blood samples.

ID	Summary	Mean	Median
Frame shift deletion	42	0.609	0
Frame shift insertion	14	0.203	0
In frame deletion	90	1.304	1
In frame insertion	16	0.232	0
Missense mutation	2426	35.159	34
Nonsense mutation	1861	26.971	26
Nonstop mutation	33	0.478	0
Splice site	68	0.986	1
Translation start site	1	0.014	0
Total	4551	65.957	65

and MAML3 (p value < $2.22e-16$) and PHLDA1 (p value = 0.0068) were highly expressed in DCM than in control left ventricle tissues (Figure 9(g)). Consistently, in the GSE126569 dataset, PHLDA1 (p value = $2.5e-07$) and MAML3 (p value = 0.0045) exhibited higher expression levels and AHNAK2 (p value = 0.0001) showed lower expression levels in DCM compared to control left ventricle samples (Figure 9(h)). No significant difference in MUC4 was found between the two groups.

4. Discussion

Without the effective treatment strategies, DCM is the major cause of heart failure [16]. Diverse genetic and environment factors to the myocardium contribute to the occurrence of DCM [17]. From the transcriptome, genome, and epigenome perspectives, our study comprehensively analyzed molecular characteristics for DCM, which could deepen the understanding of the pathogenesis for DCM and assist clinicians to specify more reasonable clinical decisions.

Due to the wide heterogeneity of the population, we integrated the two datasets to obtain 1407 common DEGs between DCM and control left ventricle samples. Functional

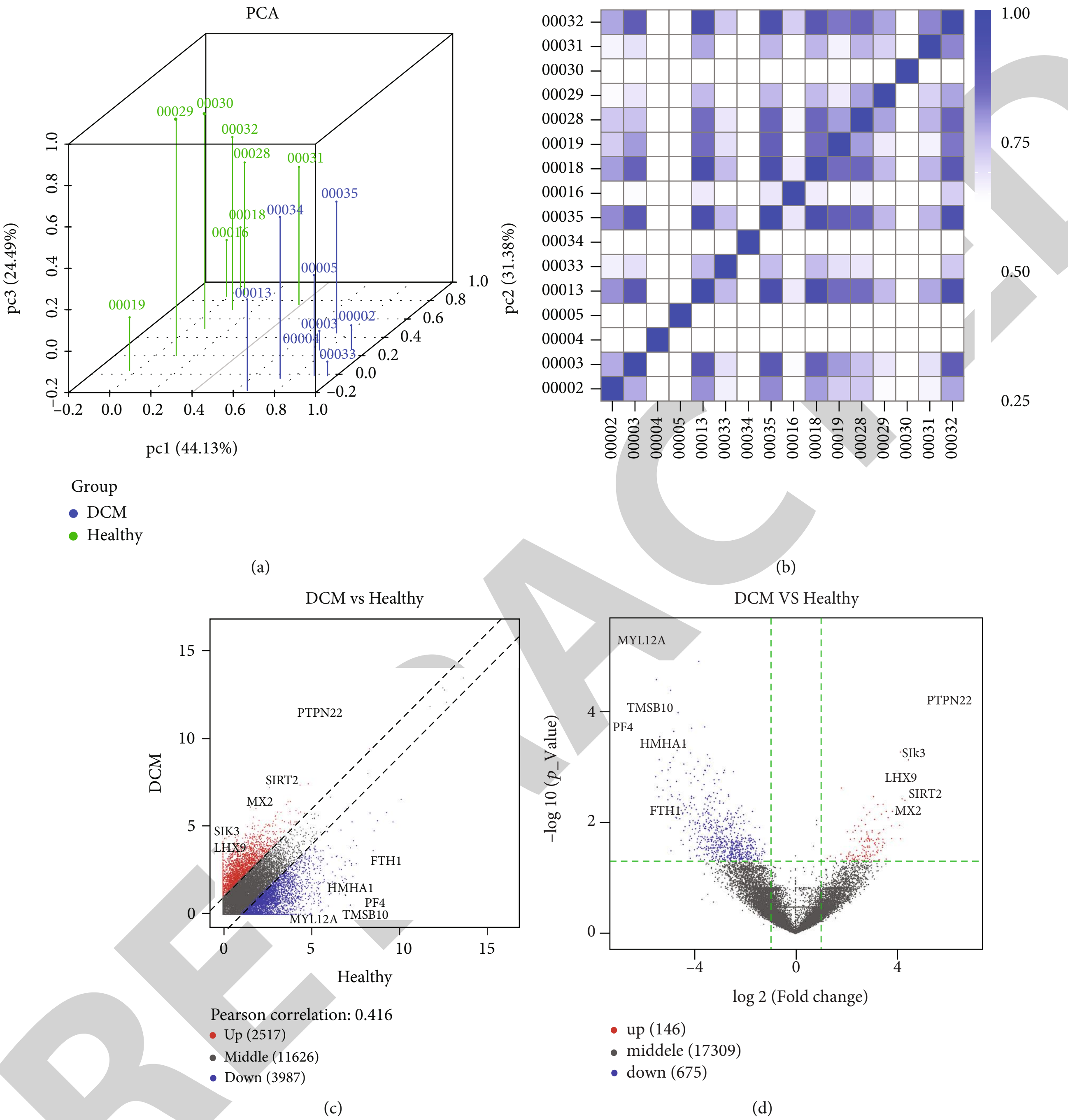


Figure 9: Continued.

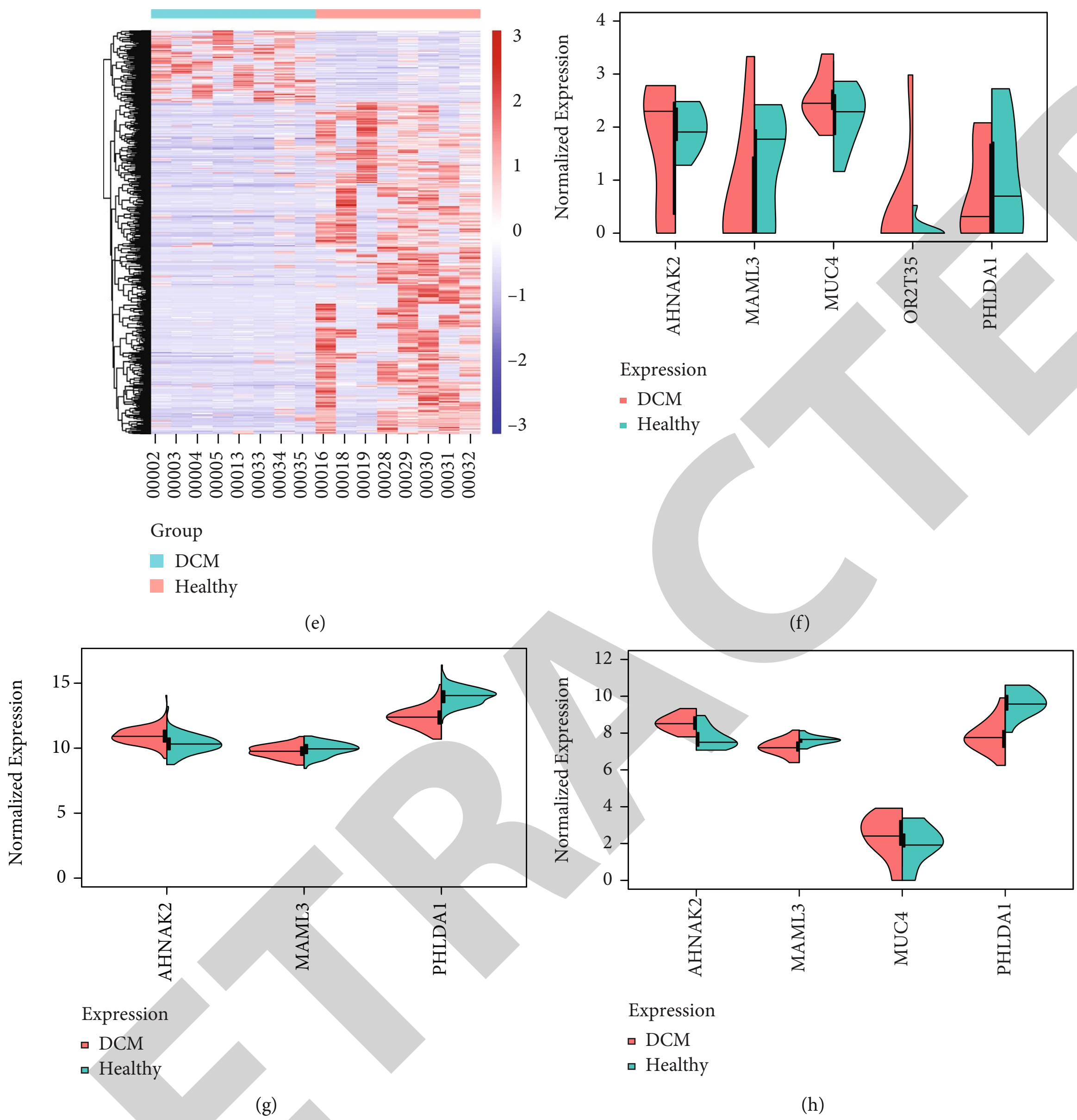


Figure 9: Validation of the expression of mutant genes in blood and left ventricle of DCM. (a) Principal component analysis for 8 DCM and control blood samples from the GSE101585 dataset. Blue: DCM samples and green: healthy samples. (b) Heat map visualizing the correlation between DCM and control blood groups. The intensity of the color is proportional to the correlation coefficient. (c) Scatter plots showing up- and downregulated genes between DCM and control blood groups. (d) Volcano plots depicting all DEGs between the two groups. (e) Heat map visualizing the expression patterns of DEGs between the two groups. Red: upregulation; blue: downregulation. (f) Differences in expression of AHNAK2, MAML3, MUC4, OR2T35, and PHLDA1 between DCM and control blood samples in the GSE101585 dataset. (g) Abnormal expression of AHNAK2, MAML3, and PHLDA1 between DCM and control left ventricle samples in the GSE141910 dataset. (h) Dysregulated expression of AHNAK2, MAML3, MUC4, and PHLDA1 between DCM and control left ventricle samples in the GSE126569 dataset.

enrichment analysis was utilized to probe into the biological functions of up- and downregulated genes. Upregulated genes were distinctly associated with extracellular matrix and collagen trimer. It has been well acknowledged that myocardial fibrosis is the main feature of DCM, involving changes in the extracellular matrix [18]. A retrospective study has found fibrosis of extracellular matrix is associated with the duration of DCM [19]. Cardiac fibrosis has a significant association with nonischemic DCM, thereby increasing its morbidity as well as mortality [20]. It has been found that changes in various genes may mediate pathological cardiac fibrosis, such as WWP2 [21]. Collagen-derived peptides have been considered circulating biomarkers for DCM, which could be mediated by different genes such as Galectin-3 [22]. Thus, these upregulated genes could be involved in regulating extracellular matrix and collagen formation, which should be further explored. Furthermore, we found that these upregulated genes were significantly enriched in immune-related pathways like Th17 cell differentiation, Th1 and Th2 cell differentiation, and cytokine and cytokine receptor, indicating that these genes could be involved in regulating immune response during the progression of DCM. As a recent study [23], Th1 and Th17 have been proposed as targets for the treatment of inflammatory DCM that is the main cause of heart failure among children as well as young adults [24]. Downregulated genes were significantly enriched in complement and coagulation cascades and phagosome, which were consistent with a previous study [25]. Based on the DEGs, we established a PPI network, composed of 171 up- and 136 downregulated genes. Among them, four hub genes were identified for DCM, including C3, GNB3, QSOX1, and APOB. Previously, C3 and APOB have been proposed as markers for stress cardiomyopathy [26]. QSOX1 exhibits high expression in myocardium tissues following acute heart failure [27]. Thus, these hub genes play a vital role in the biological processes, which may affect the expression of other genes in the PPI network.

DNA methylation is an important mechanism of epigenetic regulation [28]. Thus, identification of abnormal methylation is of clinical significance for DCM. Herein, 285 hyper- and 321 hypomethylated genes were screened for DCM. Hypermethylated genes were mainly enriched in PI3K-Akt pathway, while hypomethylated genes were primarily enriched in the calcium pathway. It has been confirmed that the PI3K-Akt pathway is in association with DCM development, which is activated or inactivated by different genes like PTEN [29]. There is an elevated myocyte calcium sensitivity for pediatric DCM in the late stage [30]. Imbalance of calcium homeostasis is closely related to DCM as well as heart failure [31]. DNA methylation may regulate gene expression. Herein, 20 differentially expressed and methylated genes were identified following integration of DEGs and DMGs. These genes had significant correlations with cell differentiation and protein digestion and absorption. More studies should be presented in further studies.

Genetic inheritance occurs in 30%-48% of patients [32]. We presented WES for 69 DCM patients in our cohort. Our results showed that MUC4 was the most frequent mutation gene which occupied 71% across 69 samples, followed by PHLDA1, AHNAK2, and MAML3. In the three independent datasets, we confirmed that PHLDA1 and MAML3 were highly expressed and AHNAK2 was lowly expressed in blood and left ventricle samples from DCM compared to control, indicating that the genetic mutation could lead to their abnormal expression. However, their expression and functions remain unclear in DCM.

5. Conclusion

Taken together, this study roundly expounded the molecular features and relevant biological functions for DCM from the transcriptome, genome, and epigenome perspectives, which may deepen the understanding for the pathogenesis of DCM. The key genes identified from different omics such as PHLDA1, MAML3, and AHNAK2 as potential therapeutic targets toward DCM will be further validated in our future studies.

Abbreviations

DCM:	Dilated cardiomyopathy
DEGs:	Differentially expressed genes
RNA-seq:	RNA sequencing
PPI:	Protein-protein interaction
DMGs:	Differentially methylated genes
SNVs:	Single-nucleotide variants
GEO:	Gene Expression Omnibus
GO:	Gene Ontology
KEGG:	Kyoto Encyclopedia of Genes and Genomes
WES:	Whole exome sequencing
SNP:	Single-nucleotide polymorphisms.

Data Availability

The datasets analyzed during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest

The authors declare no conflicts of interest.

Authors' Contributions

Li Liu, Jianjun Huang, and Yan Liu contributed equally to this work.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (81560076); Guangxi Natural Science Foundation (2018JJB140358); Middle-Aged and Young Teachers in Colleges and Universities in Guangxi Basic Ability Promotion Project (2017KY0514); Starting Research Projects for Introducing Dr. of Youjiang Medical University for Nationalities (2015bsky002); and the First Batch of High-level Talent Scientific Research Projects of the Affiliated Hospital of Youjiang Medical University for Nationalities in 2019 (R20196316).

References

[1] "Dilated cardiomyopathy," *Nature Reviews Disease Primers*, vol. 5, no. 1, 2019.

[2] A. R. Ednie, A. R. Parrish, M. J. Sonner, and E. S. Bennett, "Reduced hybrid/complex N-glycosylation disrupts cardiac electrical signaling and calcium handling in a model of dilated cardiomyopathy," *Journal of Molecular and Cellular Cardiology*, vol. 132, pp. 13–23, 2019.

[3] D. Reichart, C. Magnussen, T. Zeller, and S. Blankenberg, "Dilated cardiomyopathy: from epidemiologic to genetic phenotypes," *Journal of Internal Medicine*, vol. 286, no. 4, pp. 362–372, 2019.

[4] A. N. Rosenbaum, K. E. Agre, and N. L. Pereira, "Genetics of dilated cardiomyopathy: practical implications for heart failure management," *Nature Reviews. Cardiology*, vol. 17, no. 5, pp. 286–297, 2020.

[5] J. Yu, C. Zeng, and Y. Wang, "Epigenetics in dilated cardiomyopathy," *Current Opinion in Cardiology*, vol. 34, no. 3, pp. 260–269, 2019.

Hindawi
BioMed Research International
Volume 2024, Article ID 9838747, 1 page
https://doi.org/10.1155/2024/9838747

Retraction

Retracted: Multiomics Analysis of Genetics and Epigenetics Reveals Pathogenesis and Therapeutic Targets for Atrial Fibrillation

BioMed Research International

Received 12 March 2024; Accepted 12 March 2024; Published 20 March 2024

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope

(2) Discrepancies in the description of the research reported

(3) Discrepancies between the availability of data and the research described

(4) Inappropriate citations

(5) Incoherent, meaningless and/or irrelevant content included in the article

(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] L. Liu, J. Huang, B. Wei et al., "Multiomics Analysis of Genetics and Epigenetics Reveals Pathogenesis and Therapeutic Targets for Atrial Fibrillation," *BioMed Research International*, vol. 2021, Article ID 6644827, 36 pages, 2021.

Hindawi
BioMed Research International
Volume 2021, Article ID 6644827, 36 pages
https://doi.org/10.1155/2021/6644827

Research Article

Multiomics Analysis of Genetics and Epigenetics Reveals Pathogenesis and Therapeutic Targets for Atrial Fibrillation

Li Liu,[1] Jianjun Huang,[2] Baomin Wei,[3] Jianjiao Mo,[3] Qinjiang Wei,[3] Chengcai Chen,[4] Wei Yan,[3] Xiannan Huang,[3] Fengzhen He,[5] Lingling Qin,[6] Hehua Huang,[7] Xue Li,[8] and Xingshou Pan[3]

[1] *Department of Cardiology, Youjiang Medical University for Nationalities, Affiliated Hospital of Youjiang Medical University for Nationalities, Baise, 533000 Guangxi, China*
[2] *Department of Neurology, Youjiang Medical University for Nationalities, Affiliated Hospital of Youjiang Medical University for Nationalities, Baise, 533000 Guangxi, China*
[3] *Department of Cardiology, Affiliated Hospital of Youjiang Medical University for Nationalities, Baise, 533000 Guangxi, China*
[4] *Department of Ultrasound, Affiliated Hospital of Youjiang Medical University for Nationalities, Baise, 533000 Guangxi, China*
[5] *Graduate School of Youjiang Medical University for Nationalities, Baise, 533000 Guangxi, China*
[6] *Department of Medical Quality Management, Affiliated Hospital of Youjiang Medical University for Nationalities, Baise, 533000 Guangxi, China*
[7] *Department of Anatomy, Youjiang Medical University for Nationalities, Baise, 533000 Guangxi, China*
[8] *Department of Electrophysiology, Affiliated Hospital of Youjiang Medical University for Nationalities, Baise, 533000 Guangxi, China*

Correspondence should be addressed to Jianjun Huang; jianjun453@163.com and Xingshou Pan; pan8602@sina.com

Received 2 November 2020; Revised 14 January 2021; Accepted 9 February 2021; Published 26 March 2021

Academic Editor: Parameshachari B. D.

Objective. This study is aimed at understanding the molecular mechanisms and exploring potential therapeutic targets for atrial fibrillation (AF) by multiomics analysis. *Methods*. Transcriptomics and methylation data of AF patients were retrieved from the Gene Expression Omnibus (GEO). Differentially expressed genes (DEGs) and differentially methylated sites between AF and normal samples were screened. Then, highly expressed and hypomethylated and lowly expressed and hypermethylated genes were identified for AF. Weighted gene coexpression network analysis (WGCNA) was presented to construct AF-related coexpression networks. 52 AF blood samples were used for whole exome sequence. The mutation was visualized by the maftools package in R. Key genes were validated in AF using independent datasets. *Results*. DEGs were identified between AF and controls, which were enriched in neutrophil activation and regulation of actin cytoskeleton. RHOA, CCR2, CASP8, and SYNPO2L exhibited abnormal expression and methylation, which have been confirmed to be related to AF. PCDHA family genes had high methylation and low expression in AF. We constructed two AF-related coexpression modules. Single-nucleotide polymorphism (SNP) was the most common mutation type in AF, especially T > C. MUC4 was the most frequent mutation gene, followed by PHLDA1, AHNAK2, and MAML3. There was no statistical difference in expression of AHNAK2 and MAML3, for AF. PHLDA1 and MUC4 were confirmed to be abnormally expressed in AF. *Conclusion*. Our findings identified DEGs related to DNA methylation and mutation for AF, which may offer possible therapeutic targets and a new insight into the pathogenesis of AF from a multiomics perspective.

1. Introduction

Atrial fibrillation (AF) is a commonly diagnosed cardiac arrhythmia affecting 1% of the population globally, which is a major risk factor for stroke, heart failure, and premature death [1]. Drugs are the first choice for AF treatment. AF ablation only achieves a success rate of 60-70% [2]. The efficacy of currently available treatments is limited, which increases a major public medical burden and generates a large amount of medical expenses. Moreover, at the molecular levels, the mechanism of AF is incompletely understood. Epidemiological research shows that AF is a complex disease caused by genetic and environmental factors [3]. Due to the limited research on the role of biomarkers in the occurrence and development of AF and the management of clinical AF episodes, it is of importance to explore specific biomarkers of AF.

Multiomics analysis includes genomics (such as whole genome, single-nucleotide polymorphisms (SNP), and copy number alternation (CNA)), expression data (such as mRNA), proteomics, and epigenetics (such as methylation) [4]. With the development of next-generation sequencing (NGS) technology, abnormally expressed genes have been shown to be involved in the pathogenesis of AF [4]. DNA methylation, as one of the main epigenetic modifications, has been confirmed to be related to pathogenesis of AF [5]. DNA methylation occurs at the global and specific gene promoter level. Abnormal DNA methylation can affect the transcription and expression of key regulatory genes [6]. For example, the overall DNA methylation level of the AF group was significantly higher compared to controls [6]. Genome mutations are composed of single-nucleotide variants (SNVs), small insertions-deletions (indels), copy number alterations, and translocations [4]. In recent years, whole exome sequencing studies have identified multiple AF susceptibility gene loci [7]. As an example, a genome-wide association study has identified 104 AF-related genetic variants, which are involved in cardiac structural remodeling [7]. Nevertheless, these genes only partially explain the biological and genetic basis of AF. Only one study identified abnormally expressed genes (PSMC3, TINAG, and NUDT) regulated by methylation for AF based on multiomics analysis [5]. Herein, our study is aimed at comprehensively analyzing the genetics and epigenetics of AF, which could provide a new insight into underlying molecular mechanisms and provide therapeutic targets for AF.

2. Materials and Methods

2.1. Data Collection and Preprocessing. Microarray expression profile of left atrial (LA) myocardium from patients with AF and sinus rhythm (SR; each $n = 5$) was downloaded from the GSE14975 dataset in the Gene Expression Omnibus (GEO) repository (https://www.ncbi.nlm.nih.gov/gds/) [8]. Furthermore, we obtained the microarray expression profile of 14 AF (7 left AF and 7 right AF) and 12 SR (6 left SR and 6 right SR) samples from the GSE79768 dataset [9]. Methylation profiling data of 11 left atrium samples from 7 AF patients and 4 normal patients were retrieved from the GSE62727 dataset [10]. Microarray expression profile of 3 AF blood samples and 3 normal samples was retrieved from the GSE64904 dataset. normalizeBetweenArrays in the limma package was used to perform quartile normalization on the above microarray expression data [11]. Genes corresponding to each probe were annotated.

2.2. Differential Expression or Methylation Analysis. Differentially expressed genes (DEGs) between AF and SR samples were screened with the cutoff of false discovery rate (FDR) < 0.05 or 0.01 and $|\log_2 \text{fold change (FC)}| > 1$. Furthermore, differentially methylated sites were identified under the threshold of FDR < 0.05 and methylation difference > 0.15.

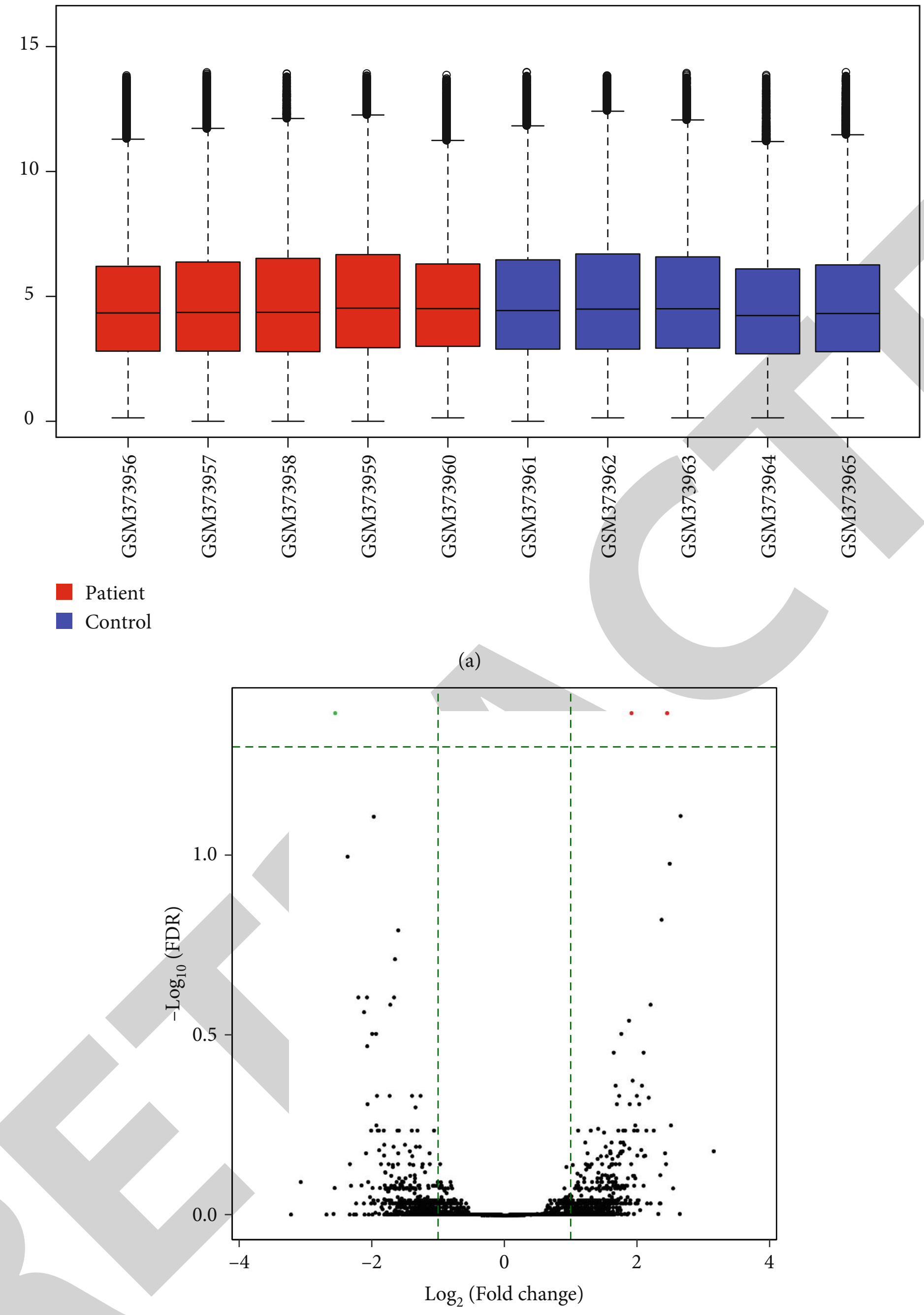
Table 1: Demographic characteristics of AF patients.

Characteristics	Number
Gender	
Female	13
Male	39
Age	66.20 ± 4.86
Hypertension	
Yes	22
No	30
Diabetes	
Yes	6
No	42
Smoking	
Yes	36
No	16

2.3. Functional Enrichment Analysis. Functional enrichment analysis of selected genes including Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) was presented using the clusterProfiler package in R [12]. GO included biological process (BP), cellular component (CC), and molecular function (MF). Adjusted p value < 0.05 was significantly enriched.

2.4. Weighted Gene Coexpression Network Analysis (WGCNA). Using the WGCNA package [13], coexpression analysis was presented based on the samples in the GSE79768 dataset. The 5000 genes with the largest expression variation were selected, and the samples were clustered based on the expression of these 5000 genes using the hclust package in R. To satisfy a scale-free network, soft threshold value was determined when independence degree > 0.85. Using the dynamic tree cutting, genes with similar expression patterns were merged into one module. The minimum number of genes in the module was 30. 400 genes were randomly selected from 5000 genes. The correlation in expression between these 400 genes was analyzed, and the results were visualized into a heat map. Then, we analyzed Pearson correlation between each module and clinical traits. In each module, correlation between gene significance (GS) and module membership (MM) was calculated.

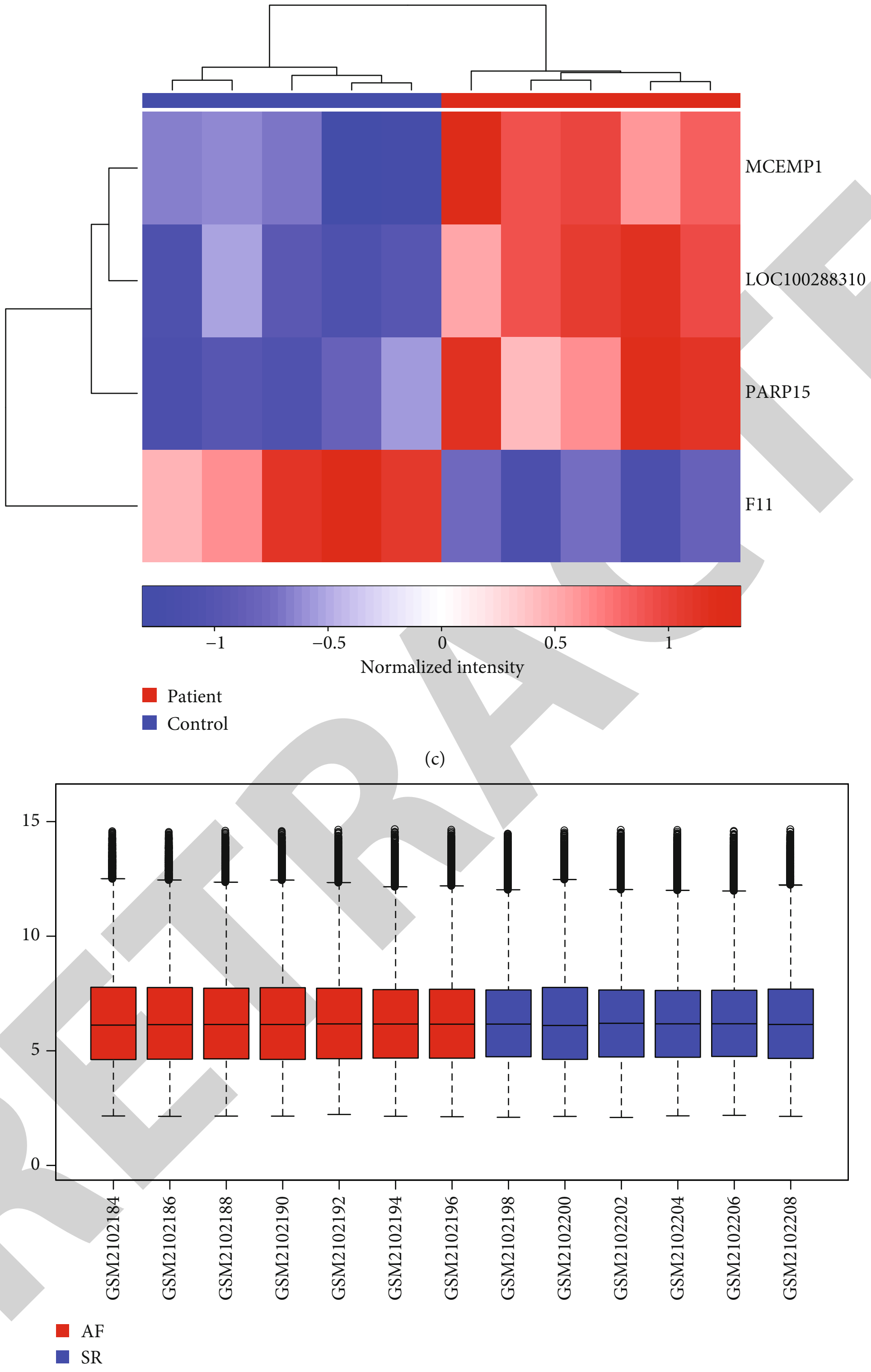
2.5. Protein-Protein Interaction (PPI) Network. Genes in coexpression modules were imported into the STRING online database (version 11.0; https://string-db.org/) [14]. PPI networks were visualized via the Cytoscape software [15] with the cutoff of 0.2 or 0.3. Core networks were constructed via the molecular complex detection (MCODE) [16]. The top ten hub genes were selected using the cytoHubba plugin in Cytoscape according to the maximal clique centrality (MCC) [17].



(a)

(b)

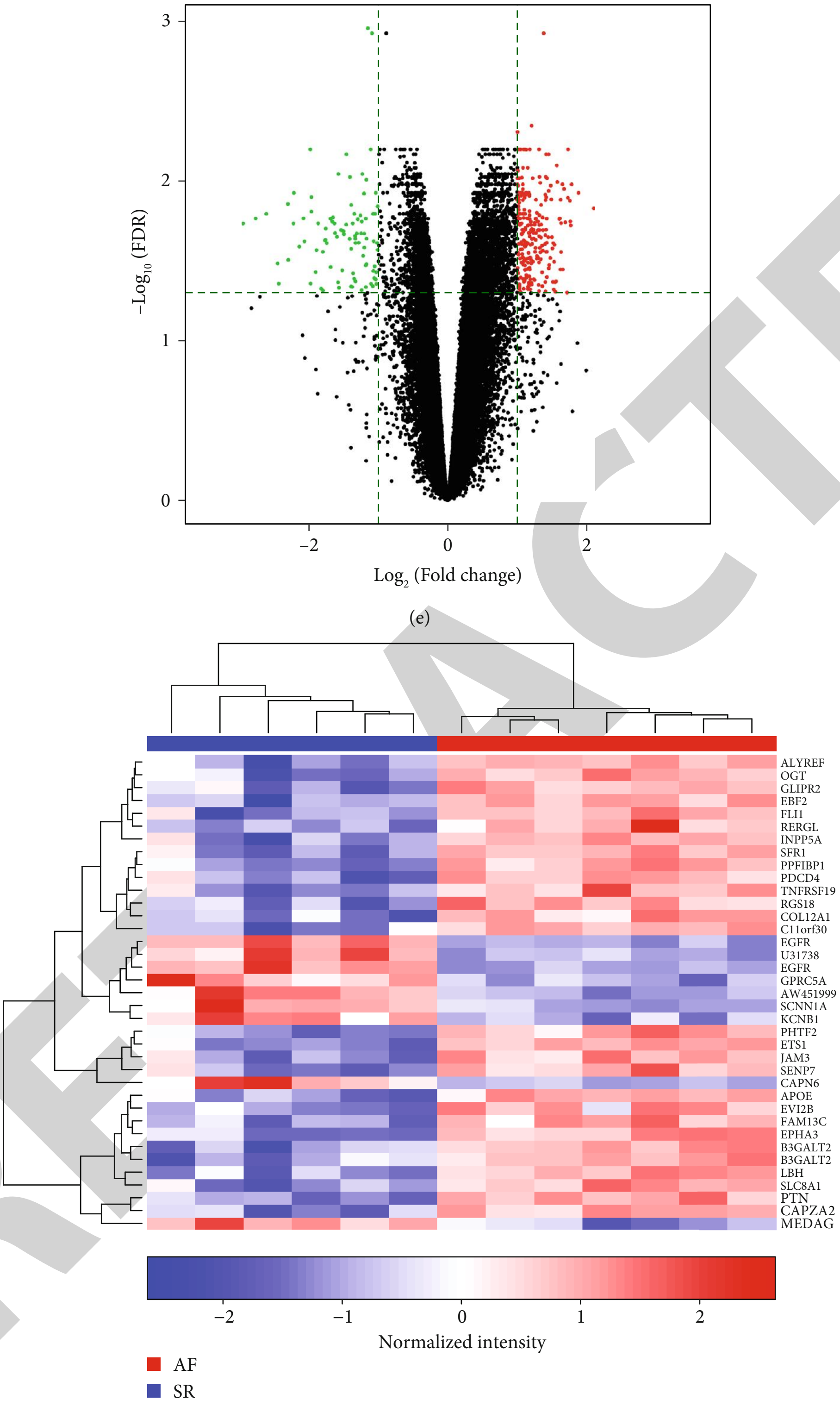
Figure 1: Continued.



(c)

(d)

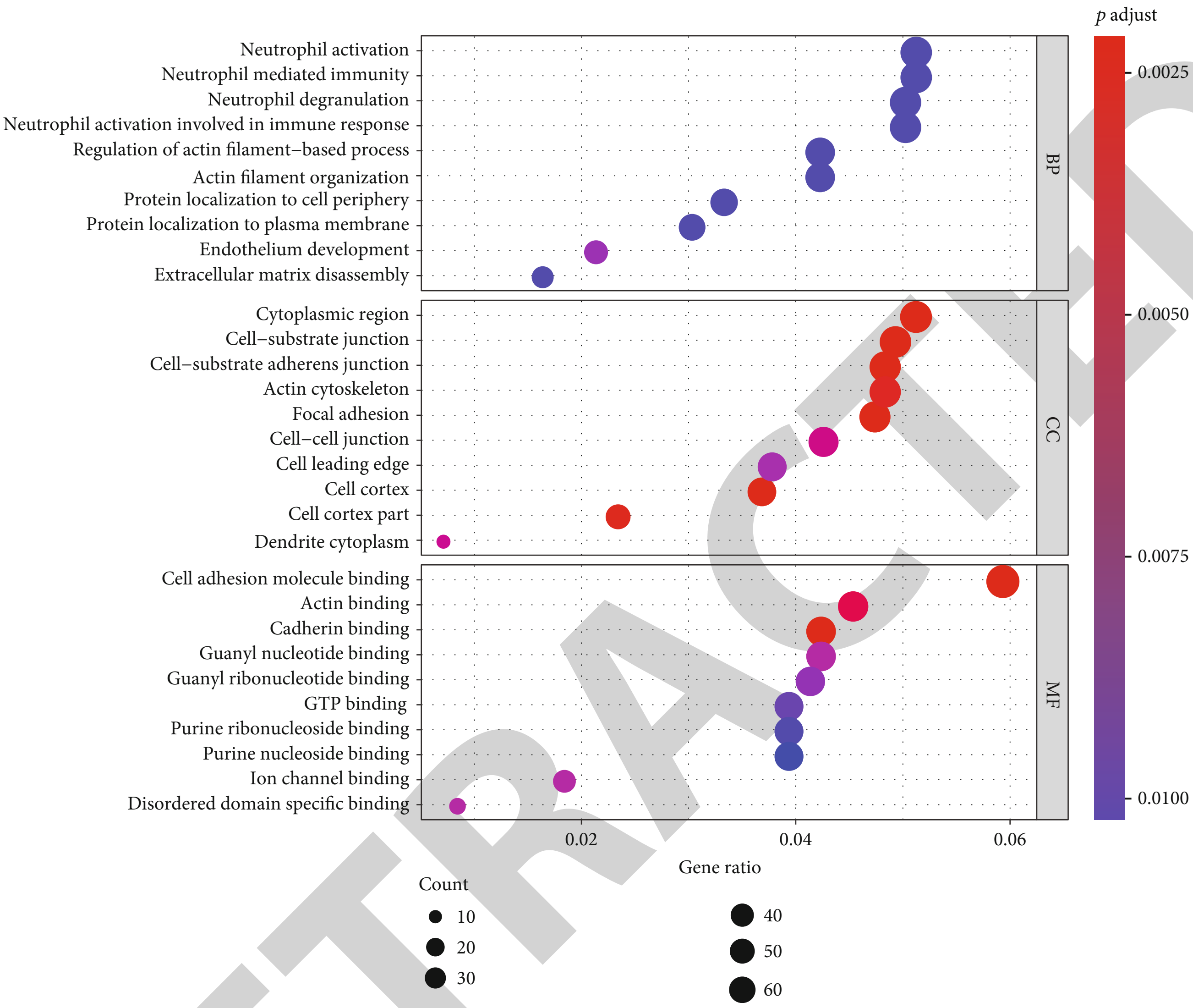
Figure 1: Continued.



(e)

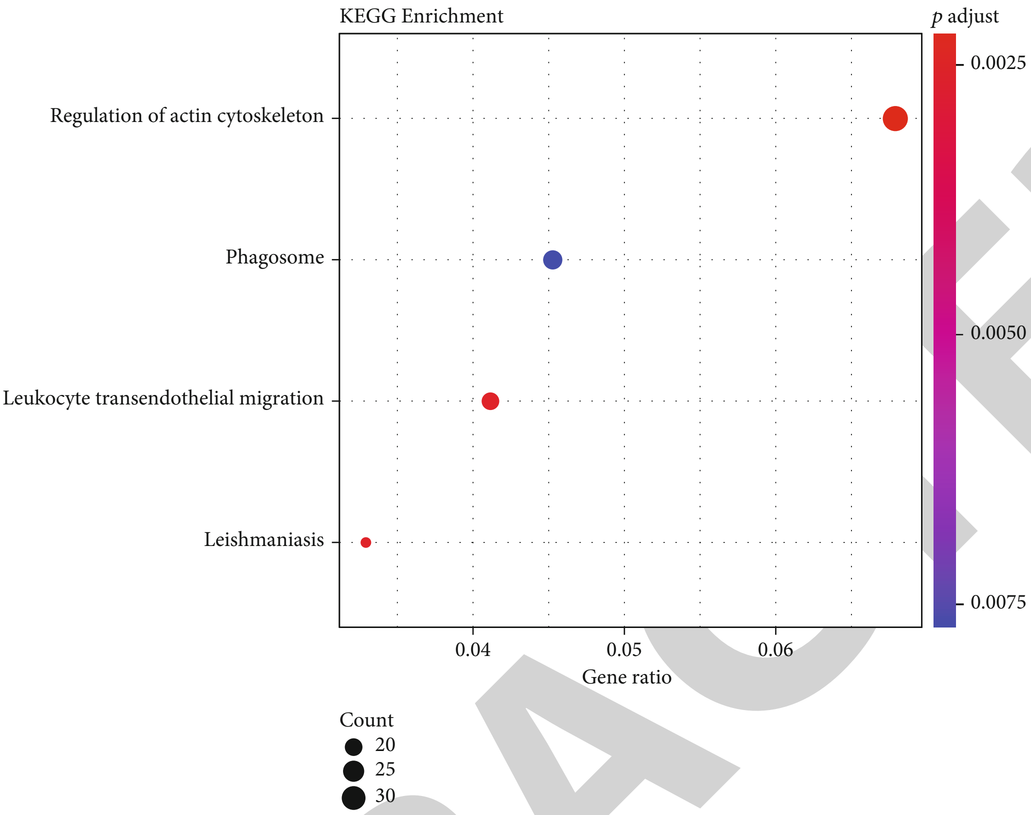
(f)

Figure 1: Continued.



(g)

Figure 1: Continued.



(h)

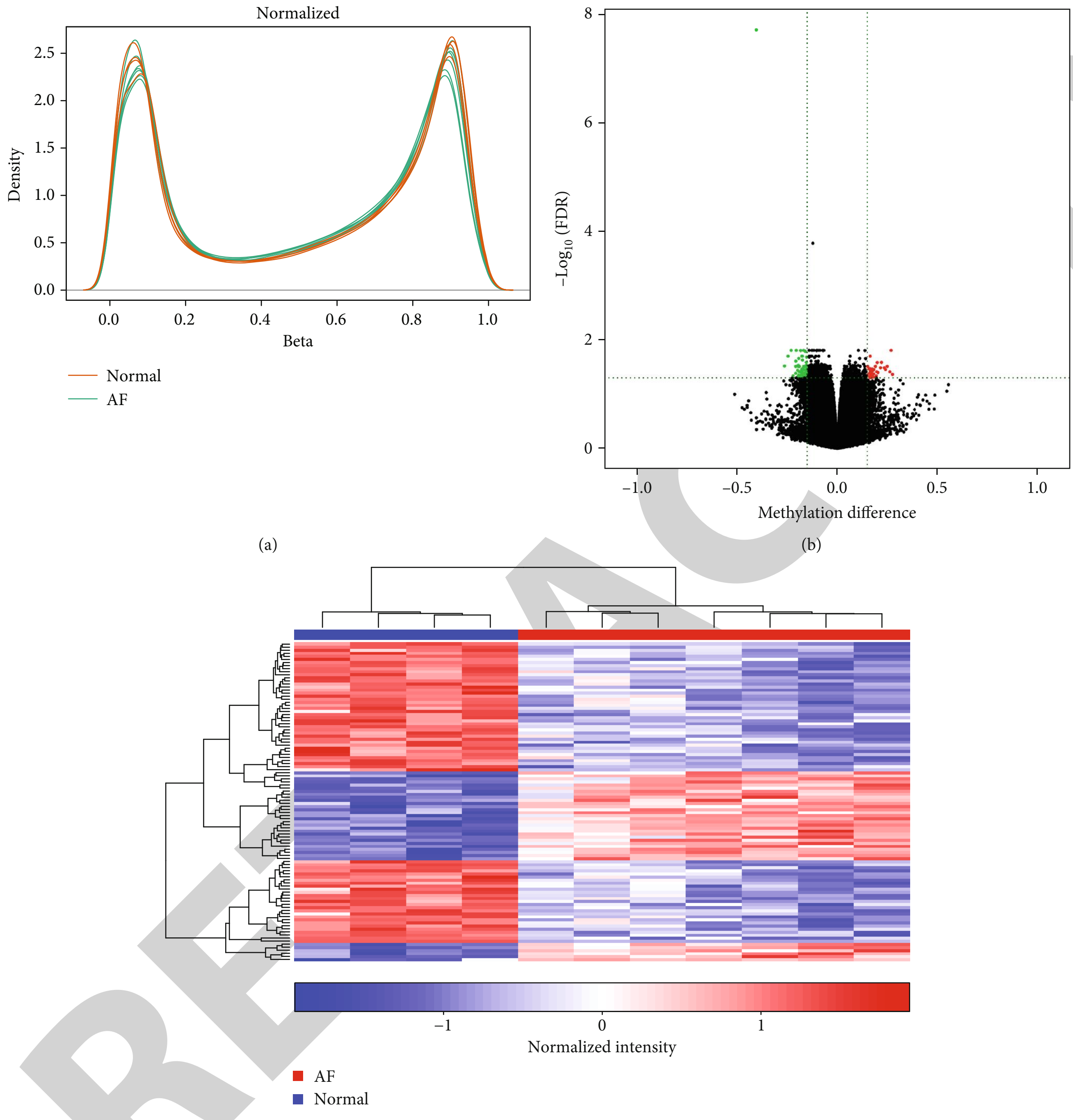
Figure 1: Screening DEGs for AF and functional enrichment analysis. (a) Box plots showing the expression levels of 5 AF and 5 SR samples from the GSE14975 dataset. Red indicates AF samples and blue indicates SR samples. (b, c) Violin plots and hierarchical clustering heat map depicting 4 DEGs between AF and SR samples. Red dot represents an upregulated gene and green dot represents a downregulated gene in AF. (d) The expression levels of 7 AF and 6 SR left atrium samples were shown from the GSE79768 dataset. Red suggests upregulation, while blue suggests downregulation in AF. (e) Screening 1433 DEGs between AF and SR groups. (f) Heat map showing the difference in expression patterns of 37 DEGs with FDR < 0.01 between the AF and SR groups. (g) GO enrichment analysis results composed of biological process (BP), cellular component (CC), and molecular function (MF) based on 1433 DEGs. (h) KEGG pathway annotation results according to 1433 DEGs.

2.6. Whole Exome Sequencing. Blood samples were obtained from 52 AF patients in the Affiliated Hospital of Youjiang Medical University for Nationalities. Whole exome sequencing was achieved by Wuhan Huada Medical Laboratory Co., Ltd. This study followed the guidelines of the Declaration of Helsinki and got the approval of the Ethics Committee of the Affiliated Hospital of Youjiang Medical University for Nationalities (YYFY-LL-2016-03). All participants provided a written informed consent. The demographic characteristics of AF patients are shown in Table 1. The mutation data of whole exome sequencing were filtered as follows: (1) the mutations with 1000G_EAS < 0.1, (2) homozygous mutations (Otherinfo = "hom"), and (3) the mutation type that had the greatest impact on the same gene in the same sample (impact was high, moderate, and low). Then, the selected mutation data were saved in the mutation annotation format (maf) format. The maftools package in R was utilized to count and visualize the maf file [18].

2.7. Statistical Analysis. All statistical analysis was presented by R language v4.0.2 (https://www.r-project.org/). p value < 0.05 was considered statistically significant.

3. Results

3.1. DEGs and Their Potential Functions in AF. In the GSE14975 dataset, box plot results showed that the median expression levels of 5 AF and 5 SR samples were basically at the same level (Figure 1(a)). Under the cutoff of FDR < 0.05 and $|\log_2\text{FC}| > 1$, 4 DEGs were identified between AF and normal samples (Figure 1(b)). Among them, MCEMP1, LOC100288310, and PARP15 were significantly upregulated and F11 was distinctly downregulated in AF compared to SR (Figure 1(b)). These DEGs could conspicuously distinguish AF from SR (Figure 1(c)). In the GSE79768 dataset, there was almost the consistent median expression level between 7 AF and 6 SR left atrium samples (Figure 1(d)). Totally, 1433 DEGs were screened for AF (Figure 1(e)). Among them, 37 DEGs with FDR < 0.01 were displayed, which could significantly distinguish AF from SR (Figure 1(f)). We further explored underlying biological functions of these 1433 DEGs. As shown in Figure 1(g), these DEGs were distinctly enriched in AF-related biological processes such as neutrophil activation, degranulation, and cell adhesion. KEGG enrichment analysis revealed that regulations of actin cytoskeleton,

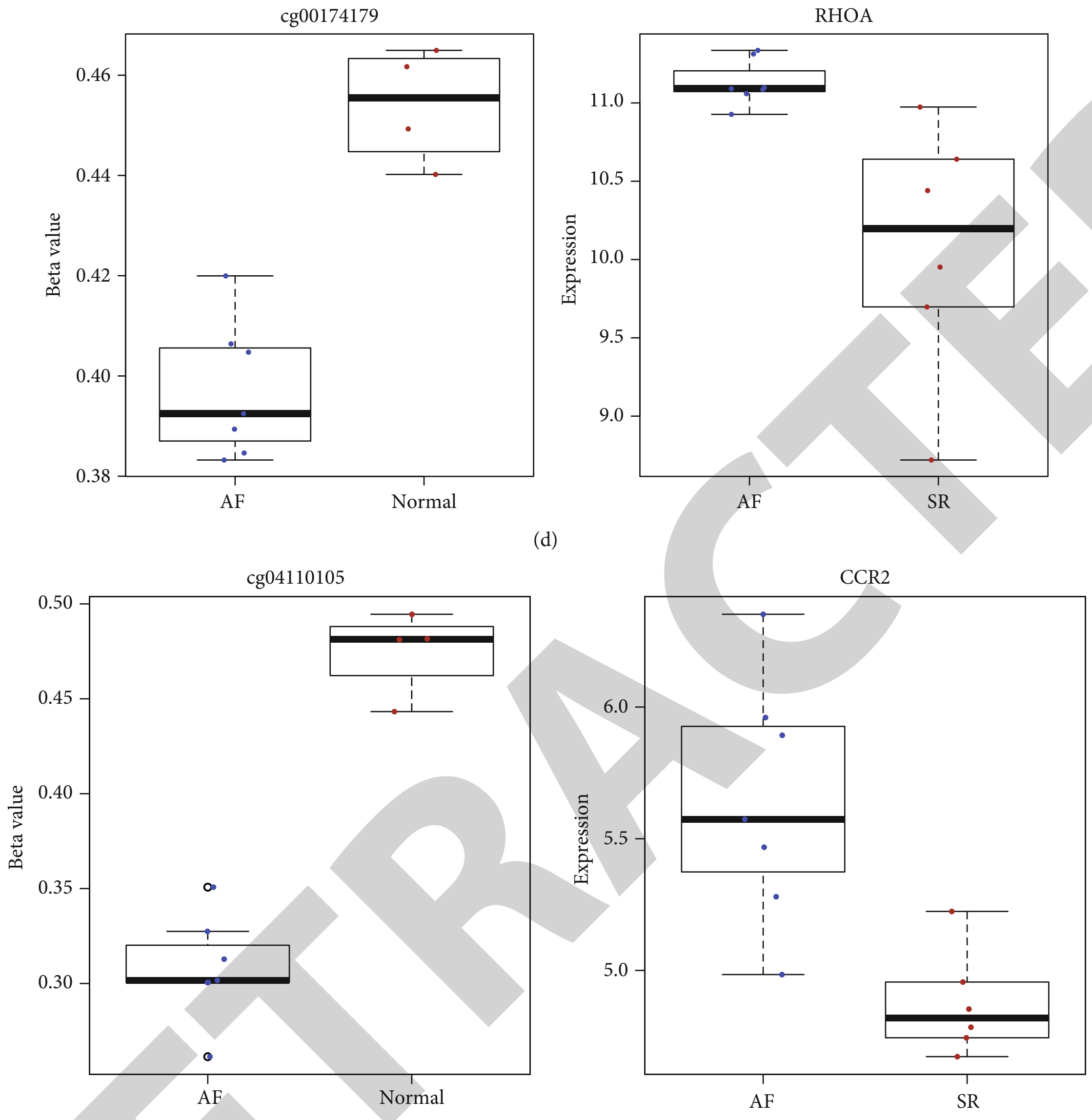


(a)

(b)

(c)

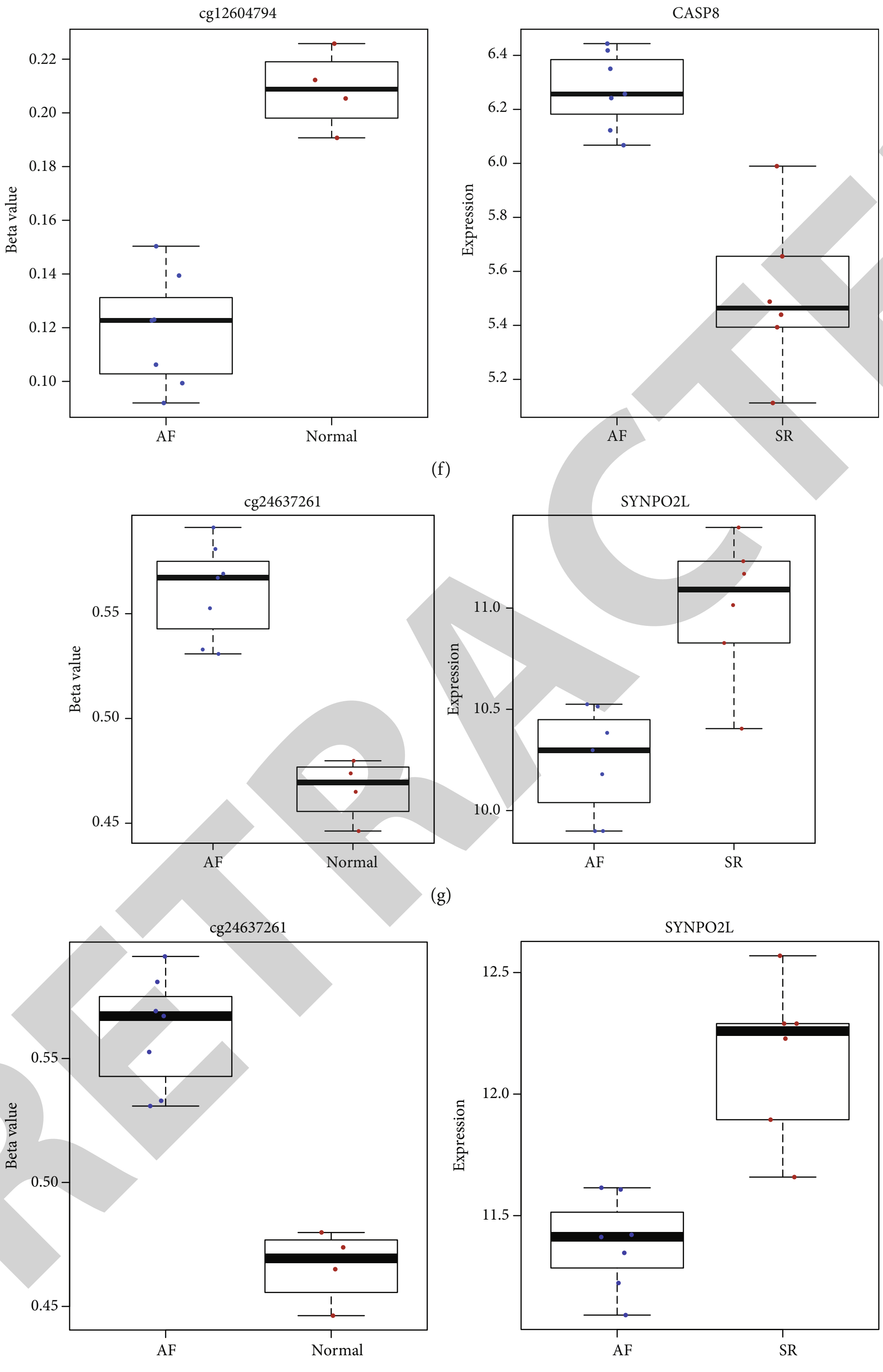
Figure 2: Continued.



(d)

(e)

FIGURE 2: Continued.

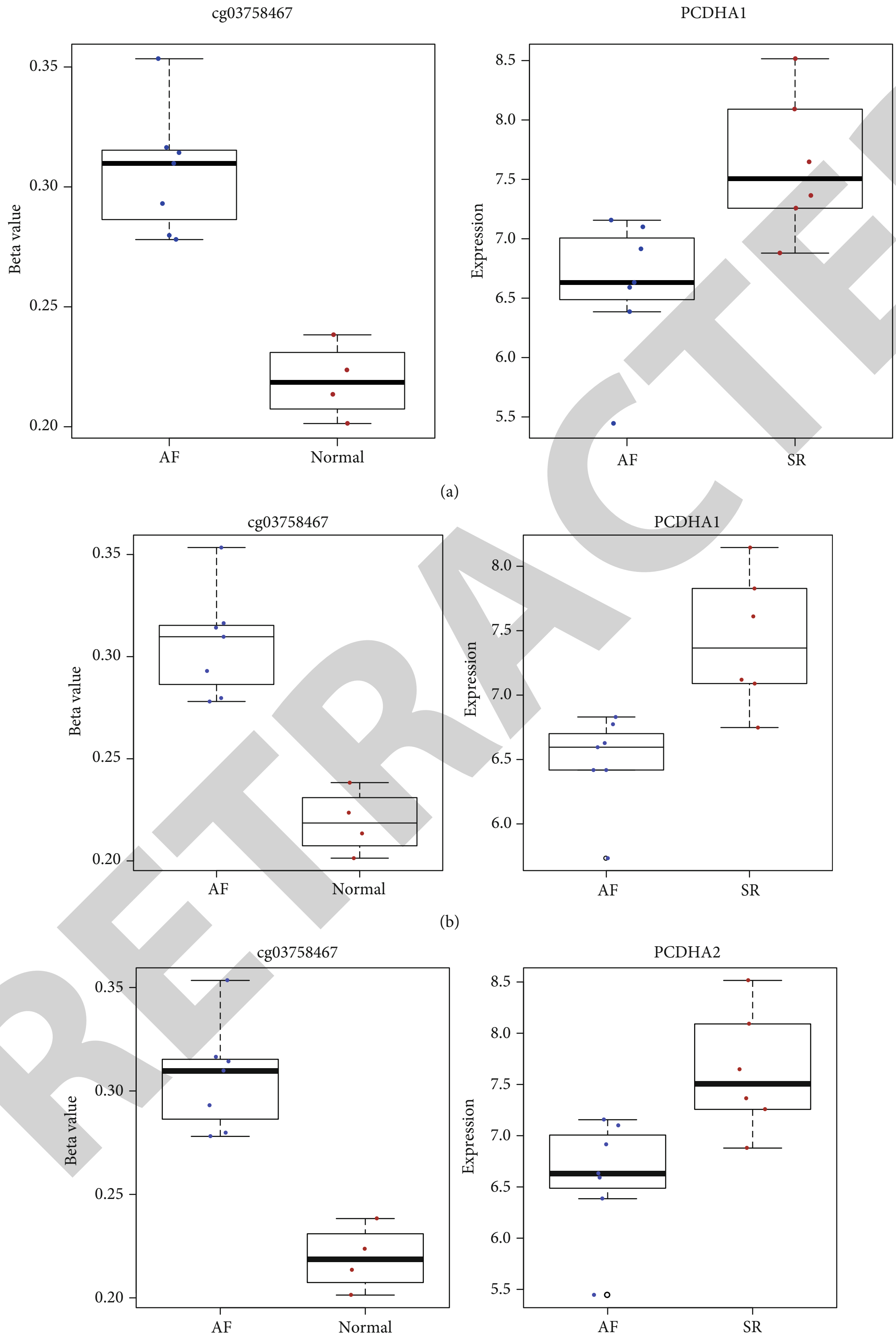


(f)

(g)

(h)

Figure 2: Identification of differentially expressed and methylated genes for AF. (a) The density plots showing β values from these 11 samples following normalization in the GSE62727 dataset. Red line indicates normal sample and green line indicates AF sample. (b) Violin plots showing 104 differentially methylated sites between the AF and normal groups. Green dots suggest hypomethylated sites and red dots suggest hypermethylated sites in AF. (c) Heat map demonstrating the methylation differences in differentially methylated sites between the two groups. Differentially expressed and methylated genes between the AF and normal groups, including RHOA (d), CCR2 (e), CASP8 (f), and SYNPO2L (g, h).

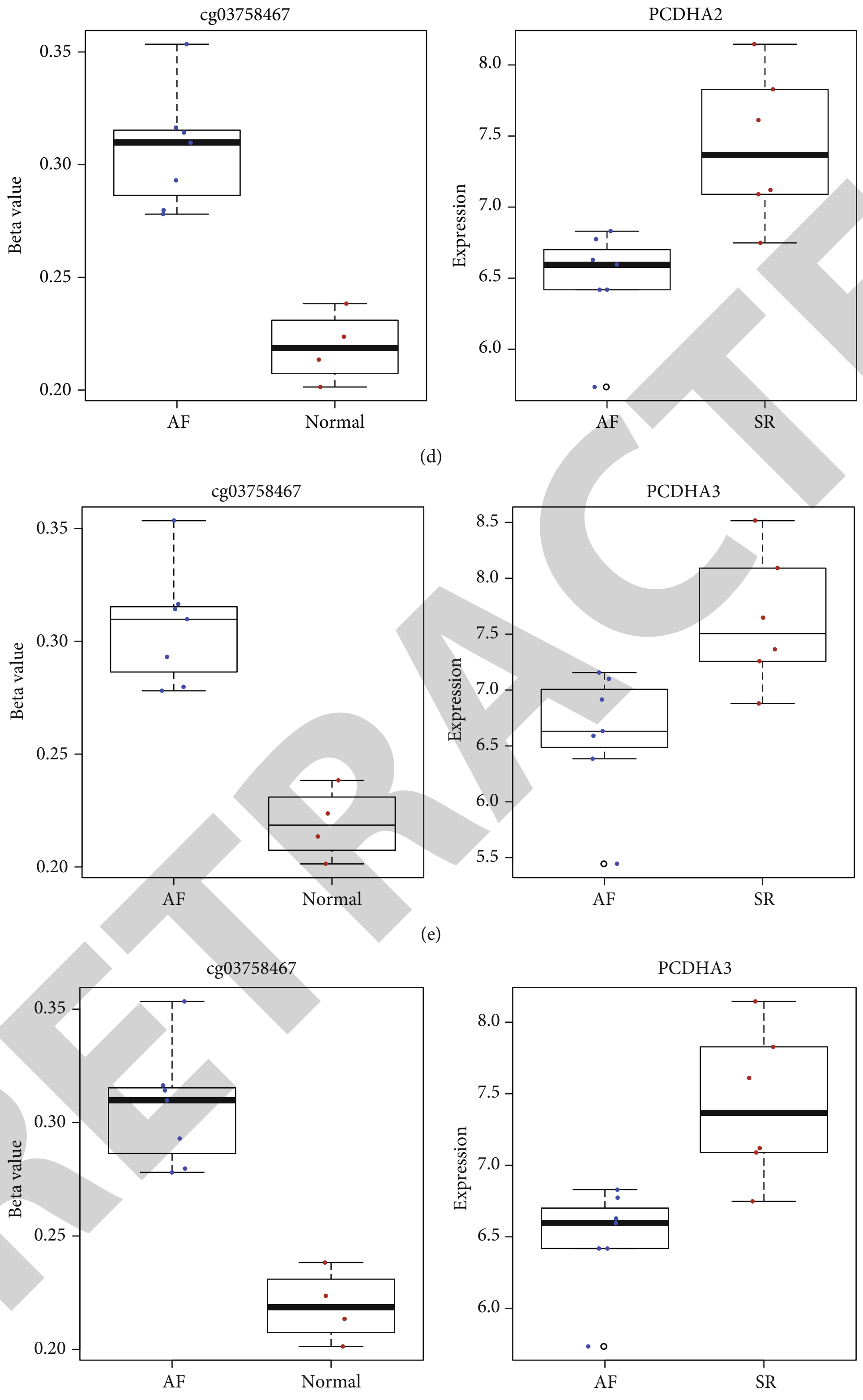


(a)

(b)

(c)

Figure 3: Continued.

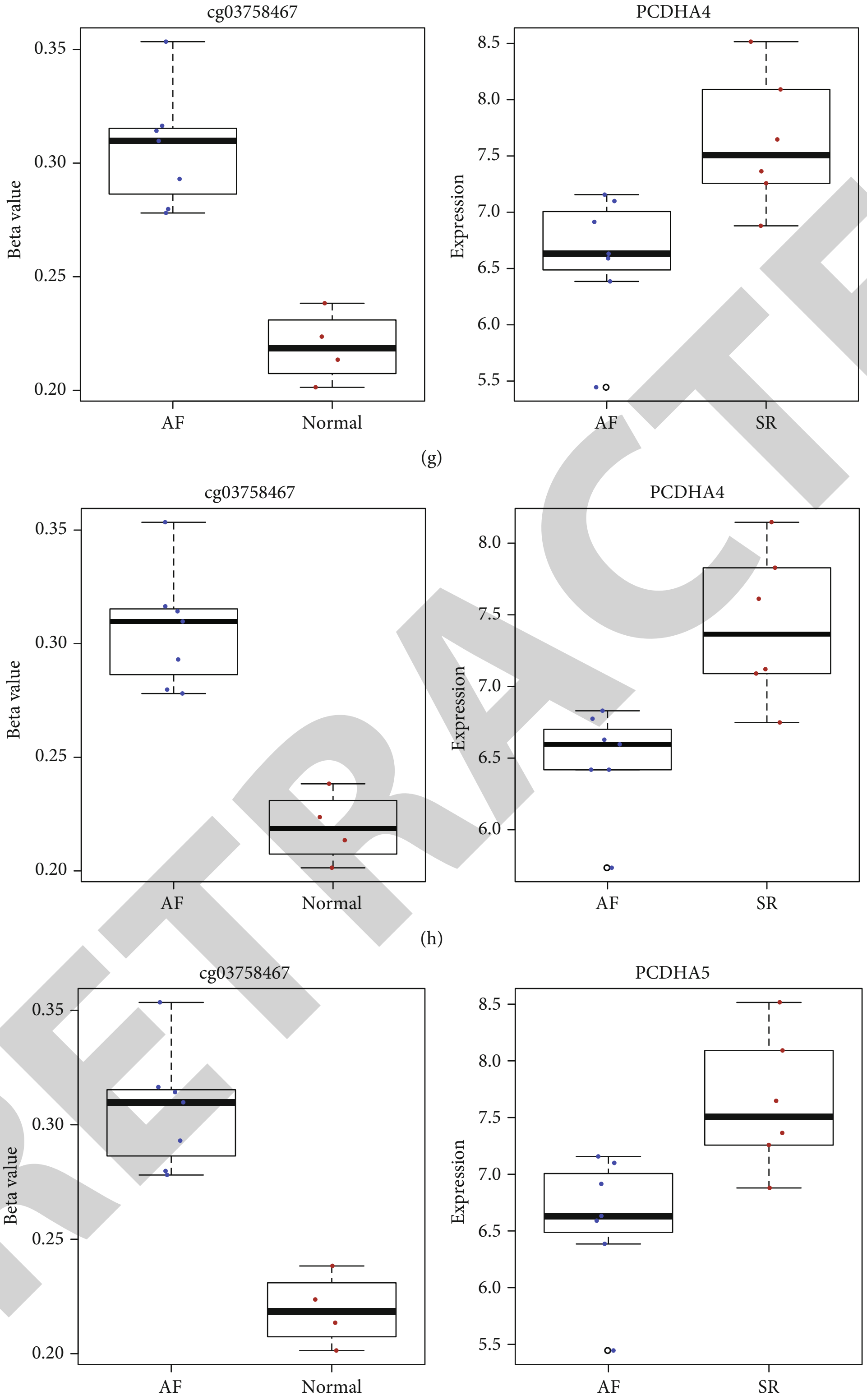


(d)

(e)

(f)

Figure 3: Continued.



(g)

(h)

(i)

Figure 3: Continued.

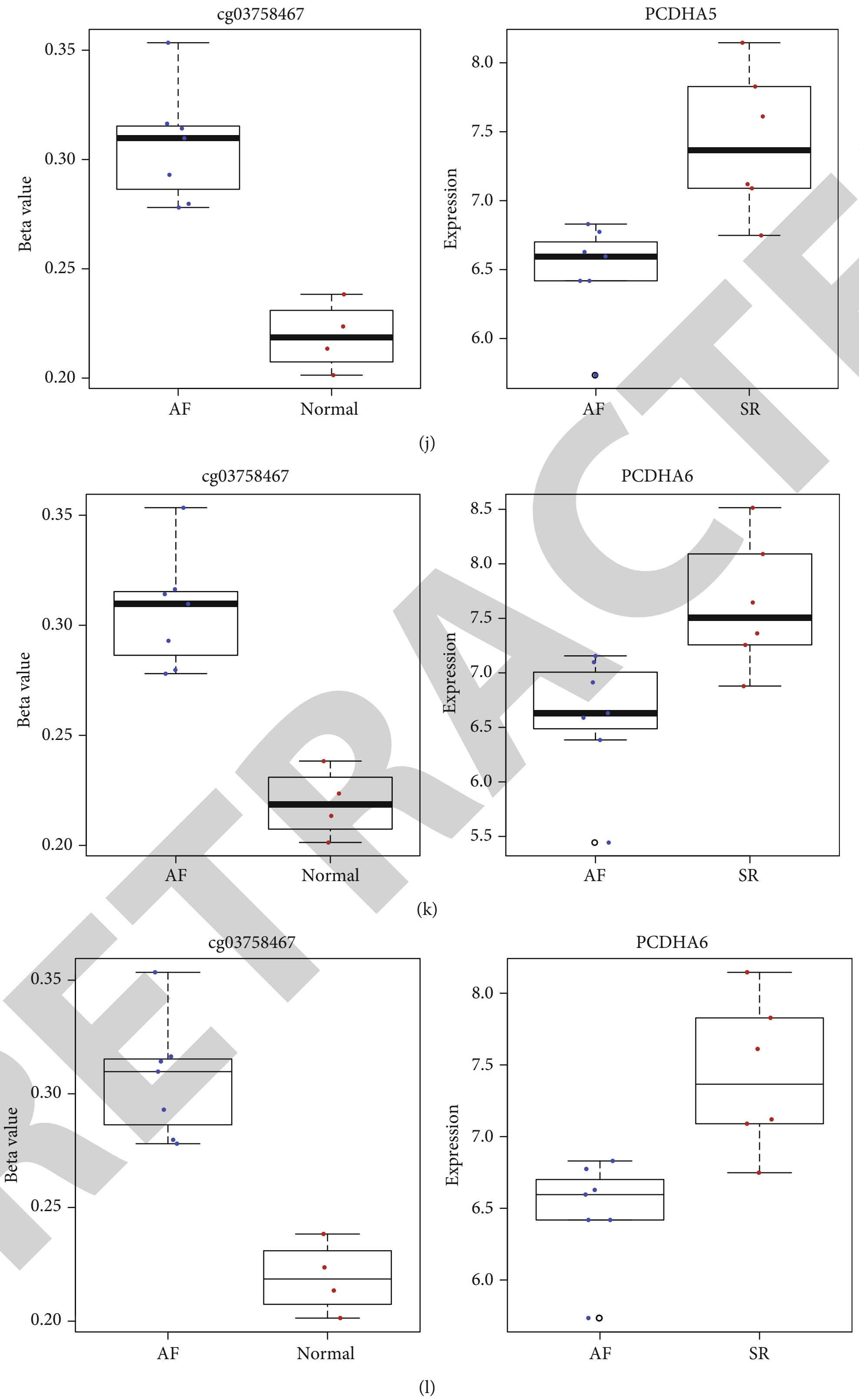
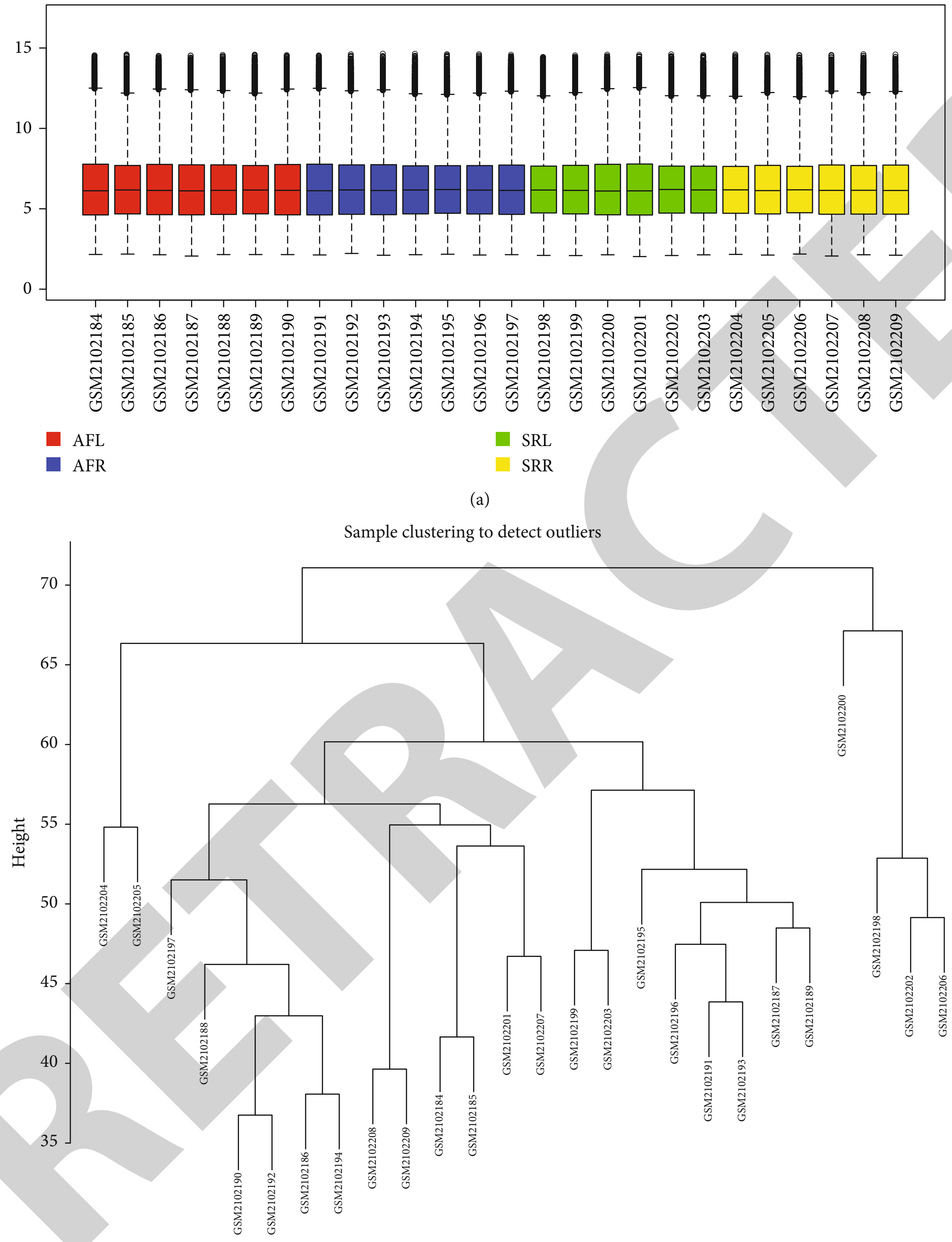


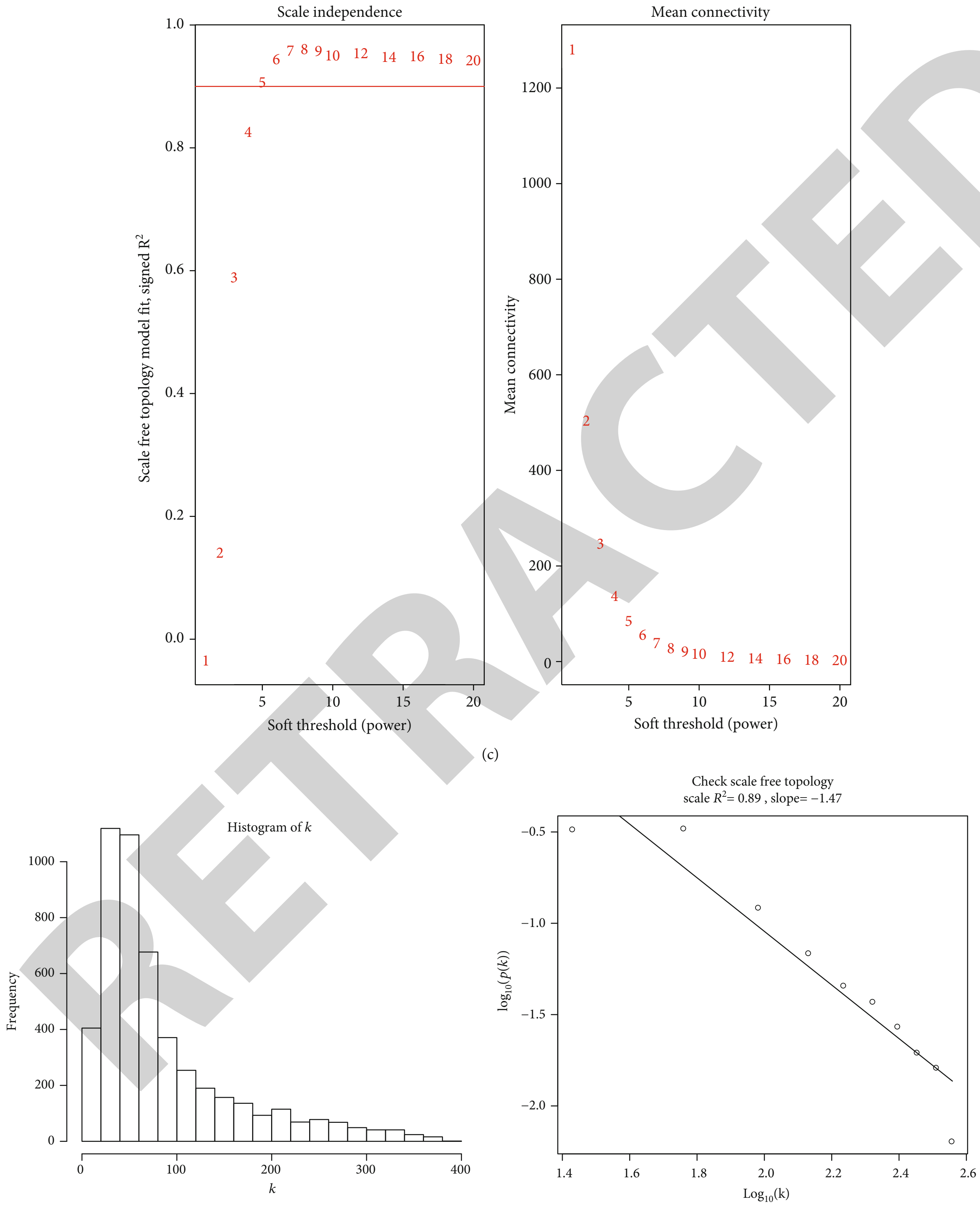
Figure 3: Differentially expressed and methylated PCDHA family genes in AF. (a, b) PCDHA1, (c, d) PCDHA2, (e, f) PCDHA3, (g, h) PCDHA4, (i, j) PCDHA5, and (k, l) PCDHA6.



(a)

(b)

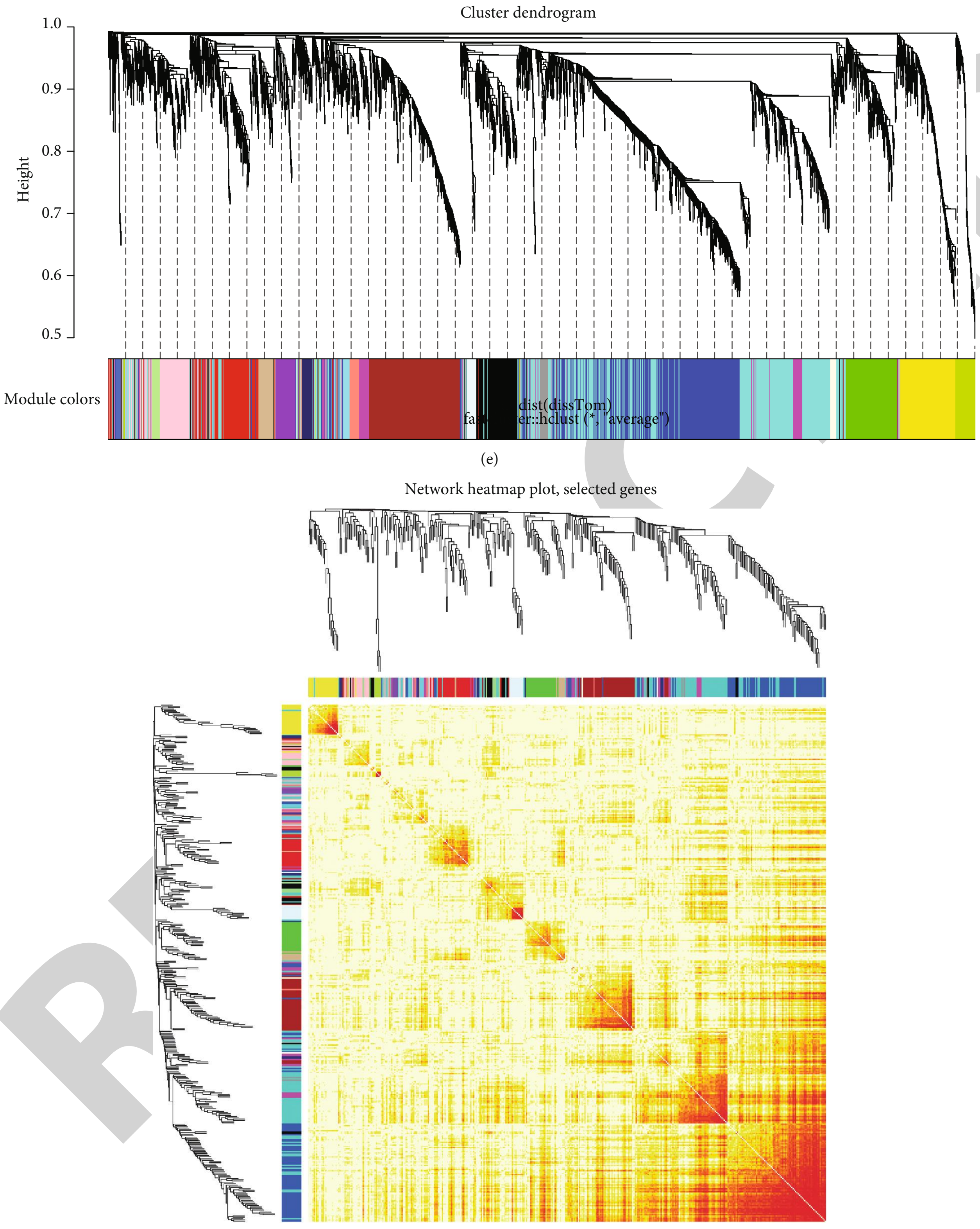
Figure 4: Continued.



(c)

(d)

Figure 4: Continued.



(e)

(f)

Figure 4: Construction of a coexpression network for AF via WGCNA. (a) Box plots depicting the expression levels in AF and SR samples from the GSE79768 dataset. (b) Sample clustering to detect outliers. (c) Scale independence and mean connectivity corresponding to difference soft threshold values. (d) Frequency of different k values and scale-free topology scale $R^2 = 0.89$ and slope $= -1.47$. (e) 21 coexpression modules were merged. Each module was marked by a certain color. (f) Heat map depicting the correlation between the expression of randomly selected 400 genes. The darker the color is, the stronger the correlation is.

phagosome, and leukocyte transendothelial migration were significantly enriched by these DEGs (Figure 1(h)).

3.2. Identification of Differentially Expressed and Methylated Genes for AF. We analyzed methylation expression profile of 7 AF and 4 normal left atrium samples from the GSE62727 dataset. Figure 2(a) depicts the density plots of β value from these 11 samples following normalization. With the threshold of $\text{FDR} < 0.05$ and methylation difference > 0.15, 104 differentially methylated sites were identified between AF and normal samples (Figure 2(b)). In Figure 2(c), differentially methylated sites can distinguish AF from normal samples. As shown in GO enrichment analysis results, genes corresponding to differentially methylated sites might be involved in regulation of hematopoietic stem cell migration. Following correlation analysis between methylation and transcriptome profiles, 28 differentially expressed and methylated genes were screened for AF. Among them, 5 genes have been reported to be involved in AF development. Among them, RHOA (Figure 2(d)), CCR2 (Figure 2(e)), and CASP8 (Figure 2(f)) were hypomethylated and highly expressed in AF than normal samples. Moreover, SYNPO2L (Figures 2(g) and 2(h)) was hypermethylated and lowly expressed in AF compared to controls.

3.3. Differentially Expressed and Methylated PCDHA Family Genes in AF. Among 28 differentially expressed and methylated genes, we found that PCDHA family genes were all hypermethylated and lowly expressed in AF compared to controls. PCDHA family genes had two hypermethylated sites between AF and SR samples, including PCDHA1 (Figures 3(a) and 3(b)), PCDHA2 (Figures 3(c) and 3(d)), PCDHA3 (Figures 3(e) and 3(f)), PCDHA4 (Figures 3(g) and 3(h)), PCDHA5 (Figures 3(i) and 3(j)), and PCDHA6 (Figures 3(k) and 3(l)).

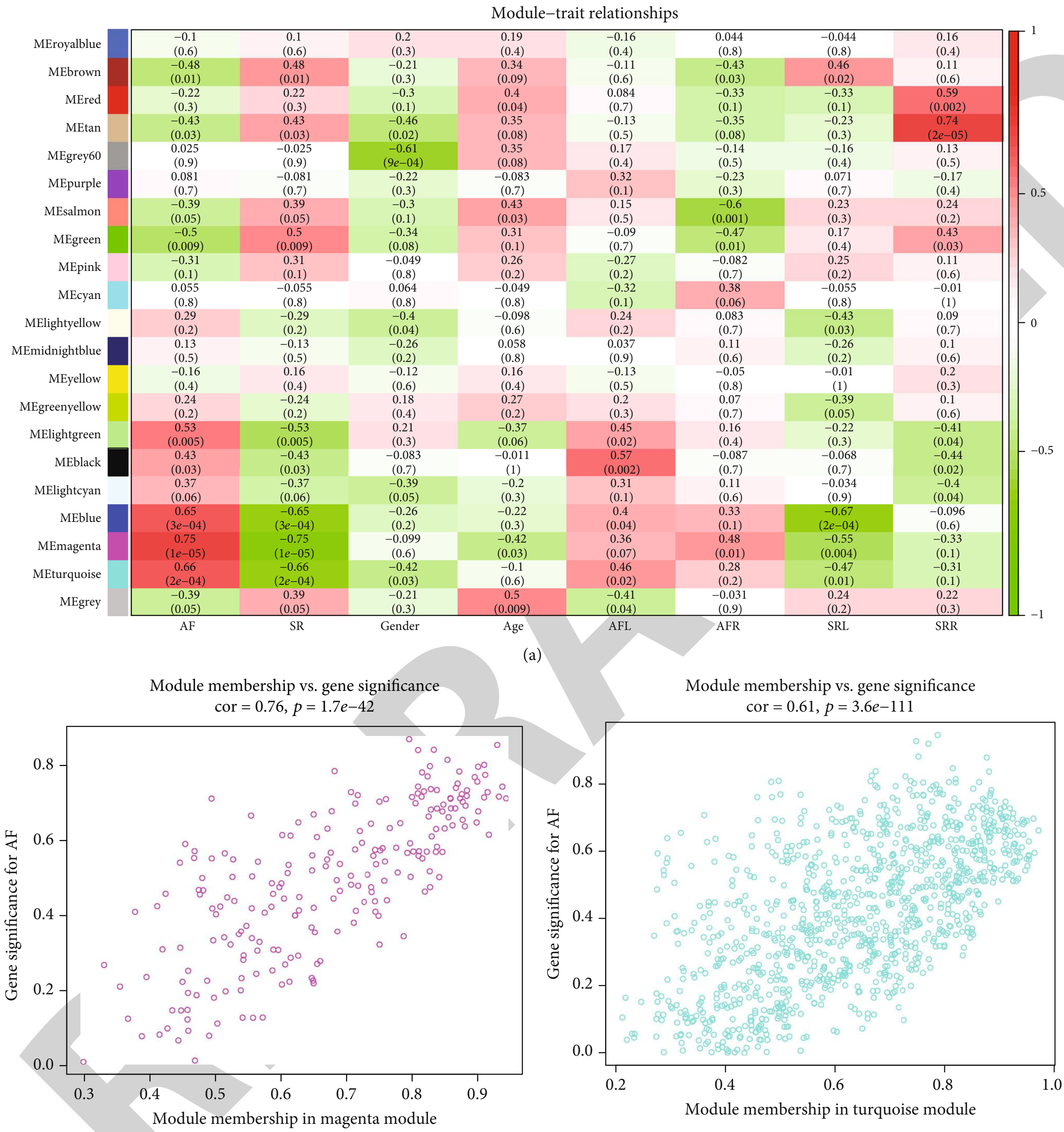
3.4. Construction of a Coexpression Network for AF. 14 AF (7 left AF and 7 right AF) and 12 SR (6 left SR and 6 right SR) samples from the GSE79768 dataset were employed for constructing a coexpression network for AF. After normalization, the expression levels in all samples tended to be the same (Figure 4(a)). According to the 5000 genes with the largest expression variation, the samples were clustered using the hclust package in R. As shown in Figure 4(b), there was no outlier. The biological interaction network must meet the scale free. In this study, when the soft threshold was 5, the independence degree was up to 0.89 (Figure 4(c)). Further analysis confirmed that the constructed coexpression network satisfied scale free when the soft threshold was 5 (Figure 4(d)). Finally, a total of 21 coexpression modules were identified for AF (Figure 4(e)). Each module was represented by a certain color. Table 2 lists the number of genes contained in each module. 400 genes were randomly selected from 5000 genes. Gene modules were determined based on the similarity of gene expression. The heat map depicted the high correlation between the expression of these 400 genes (Figure 4(f)).

3.5. Identification of AF-Related Coexpression Modules and Hub Genes. We further analyzed the correlation between 21 coexpression modules and different clinical traits. In Figure 5(a), magenta module was significantly correlated to AF ($r = 0.75$ and $p = 1e - 05$), SR ($r = -0.75$ and $p = 1e - 05$), age ($r = -0.42$ and $p = 0.03$), right AF (AFR; $r = 0.48$ and $p = 0.01$), and left SR (AFR; $r = -0.55$ and $p = 0.004$). Turquoise module had a significant correlation with AF ($r = 0.66$ and $p = 2e - 04$), SR ($r = -0.66$ and $p = 2e - 04$), gender ($r = -0.42$ and $p = 0.03$), left AF (AFL; $r = 0.46$ and $p = 0.02$), and left SR (SRL; $r = -0.47$ and $p = 0.01$). Thus, above two modules were significantly correlated to AF. Scatter plots showed that genes in magenta (Figure 5(b); $r = 0.76$ and $p = 1.7e - 42$) and turquoise (Figure 5(c); $r = 0.61$ and $p = 3.6e - 111$) modules were significantly related to AF. The cluster analysis results also indicated that magenta and turquoise modules were correlated with AF (Figure 5(d)). Genes in magenta module were significantly correlated with mesenchymal cell proliferation (Figure 5(e)). Furthermore, genes in turquoise module were distinctly enriched in fatty acid metabolic process (Figure 5(f)).

A PPI network composed of 92 nodes was constructed based on genes in magenta module with the cutoff value of 0.2 (Figure 5(g)). According to the PPI network, two core networks were constructed when score = 19 (Figure 5(h)) and 11.862 (Figure 5(i)). Using the cytoHubba plugin of Cytoscape, we identified the top ten hub genes for magenta module according to the maximal clique centrality (MCC; Figure 5(j)), including LSM5 (degree = 55), MRS2 (degree = 80), AIMP1 (degree = 91), ACTR6 (degree = 89), MFN1 (degree = 70), RWDD3 (degree = 73), CAPZA2 (degree = 70), C11orf30 (degree = 82), CCPG1 (degree = 77),

Table 2: The number of genes in each coexpression module.

Module	Number of genes
Black	268
Blue	868
Brown	569
Cyan	71
Green	299
Green yellow	111
Grey	84
Grey 60	49
Light cyan	59
Light green	48
Light yellow	42
Magenta	219
Midnight blue	61
Pink	241
Purple	119
Red	277
Royal blue	30
Salmon	76
Tan	96
Turquoise	1081
Yellow	332



Module–trait relationships

	AF	SR	Gender	Age	AFL	AFR	SRL	SRR
MEroyalblue	−0.1 (0.6)	0.1 (0.6)	0.2 (0.3)	0.19 (0.4)	−0.16 (0.4)	0.044 (0.8)	−0.044 (0.8)	0.16 (0.4)
MEbrown	−0.48 (0.01)	0.48 (0.01)	−0.21 (0.3)	0.34 (0.09)	−0.11 (0.6)	−0.43 (0.03)	0.46 (0.02)	0.11 (0.6)
MEred	−0.22 (0.3)	0.22 (0.3)	−0.3 (0.1)	0.4 (0.04)	0.084 (0.7)	−0.33 (0.1)	−0.33 (0.1)	0.59 (0.002)
MEtan	−0.43 (0.03)	0.43 (0.03)	−0.46 (0.02)	0.35 (0.08)	−0.13 (0.5)	−0.35 (0.08)	−0.23 (0.3)	0.74 (2*e*−05)
MEgrey60	0.025 (0.9)	−0.025 (0.9)	−0.61 (9*e*−04)	0.35 (0.08)	0.17 (0.4)	−0.14 (0.5)	−0.16 (0.4)	0.13 (0.5)
MEpurple	0.081 (0.7)	−0.081 (0.7)	−0.22 (0.3)	−0.083 (0.7)	0.32 (0.1)	−0.23 (0.3)	0.071 (0.7)	−0.17 (0.4)
MEsalmon	−0.39 (0.05)	0.39 (0.05)	−0.3 (0.1)	0.43 (0.03)	0.15 (0.5)	−0.6 (0.001)	0.23 (0.3)	0.24 (0.2)
MEgreen	−0.5 (0.009)	0.5 (0.009)	−0.34 (0.08)	0.31 (0.1)	−0.09 (0.7)	−0.47 (0.01)	0.17 (0.4)	0.43 (0.03)
MEpink	−0.31 (0.1)	0.31 (0.1)	−0.049 (0.8)	0.26 (0.2)	−0.27 (0.2)	−0.082 (0.7)	0.25 (0.2)	0.11 (0.6)
MEcyan	0.055 (0.8)	−0.055 (0.8)	0.064 (0.8)	−0.049 (0.8)	−0.32 (0.1)	0.38 (0.06)	−0.055 (0.8)	−0.01 (1)
MElightyellow	0.29 (0.2)	−0.29 (0.2)	−0.4 (0.04)	−0.098 (0.6)	0.24 (0.2)	0.083 (0.7)	−0.43 (0.03)	0.09 (0.7)
MEmidnightblue	0.13 (0.5)	−0.13 (0.5)	−0.26 (0.2)	0.058 (0.8)	0.037 (0.9)	0.11 (0.6)	−0.26 (0.2)	0.1 (0.6)
MEyellow	−0.16 (0.4)	0.16 (0.4)	−0.12 (0.6)	0.16 (0.4)	−0.13 (0.5)	−0.05 (0.8)	−0.01 (1)	0.2 (0.3)
MEgreenyellow	0.24 (0.2)	−0.24 (0.2)	0.18 (0.4)	0.27 (0.2)	0.2 (0.3)	0.07 (0.7)	−0.39 (0.05)	0.1 (0.6)
MElightgreen	0.53 (0.005)	−0.53 (0.005)	0.21 (0.3)	−0.37 (0.06)	0.45 (0.02)	0.16 (0.4)	−0.22 (0.3)	−0.41 (0.04)
MEblack	0.43 (0.03)	−0.43 (0.03)	−0.083 (0.7)	−0.011 (1)	0.57 (0.002)	−0.087 (0.7)	−0.068 (0.7)	−0.44 (0.02)
MElightcyan	0.37 (0.06)	−0.37 (0.06)	−0.39 (0.05)	−0.2 (0.3)	0.31 (0.1)	0.11 (0.6)	−0.034 (0.9)	−0.4 (0.04)
MEblue	0.65 (3*e*−04)	−0.65 (3*e*−04)	−0.26 (0.2)	−0.22 (0.3)	0.4 (0.04)	0.33 (0.1)	−0.67 (2*e*−04)	−0.096 (0.6)
MEmagenta	0.75 (1*e*−05)	−0.75 (1*e*−05)	−0.099 (0.6)	−0.42 (0.03)	0.36 (0.07)	0.48 (0.01)	−0.55 (0.004)	−0.33 (0.1)
MEturquoise	0.66 (2*e*−04)	−0.66 (2*e*−04)	−0.42 (0.03)	−0.1 (0.6)	0.46 (0.02)	0.28 (0.2)	−0.47 (0.01)	−0.31 (0.1)
MEgrey	−0.39 (0.05)	0.39 (0.05)	−0.21 (0.3)	0.5 (0.009)	−0.41 (0.04)	−0.031 (0.9)	0.24 (0.2)	0.22 (0.3)

(a)

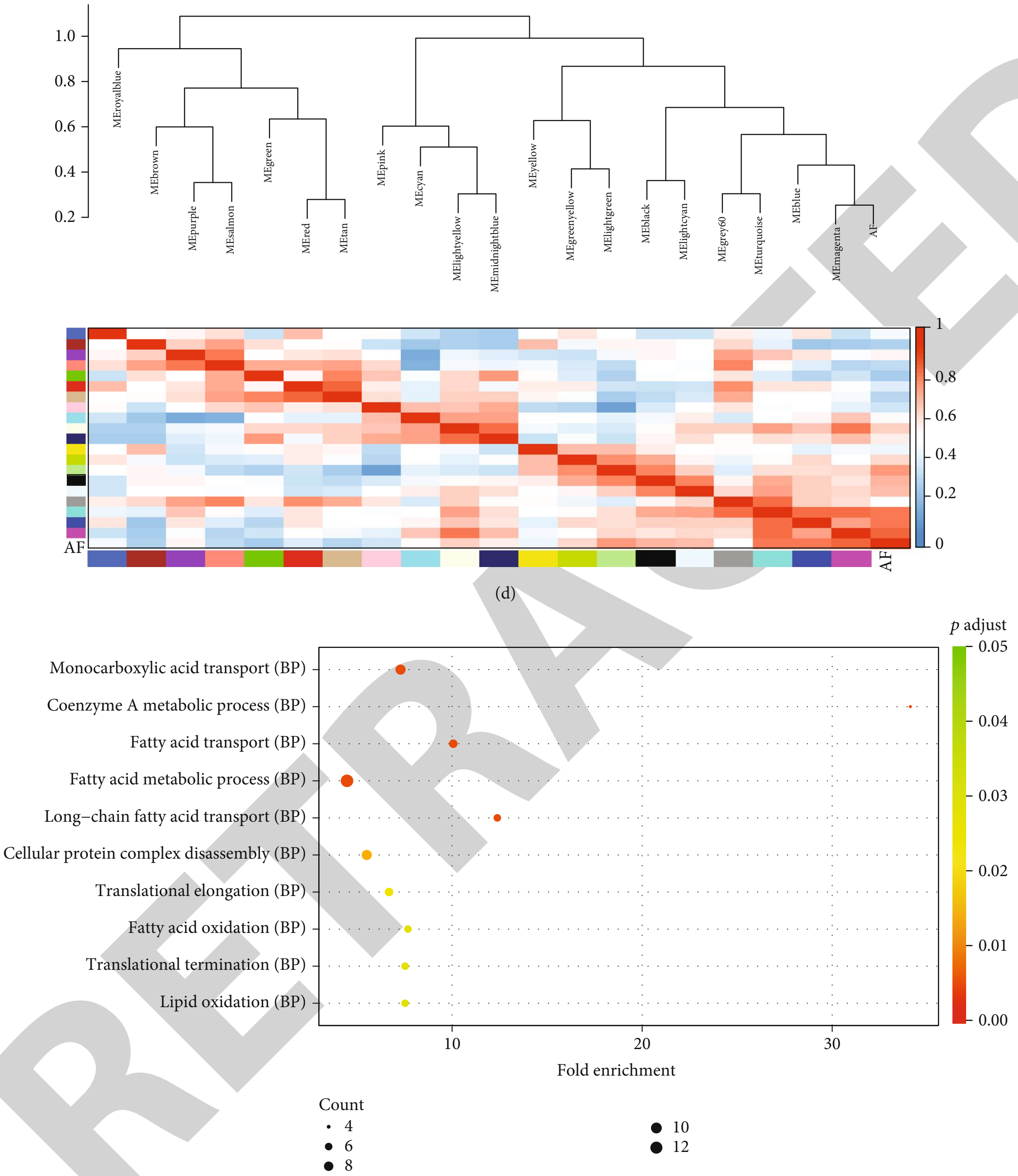
Module membership vs. gene significance
cor = 0.76, *p* = 1.7*e*−42

(b)

Module membership vs. gene significance
cor = 0.61, *p* = 3.6*e*−111

(c)

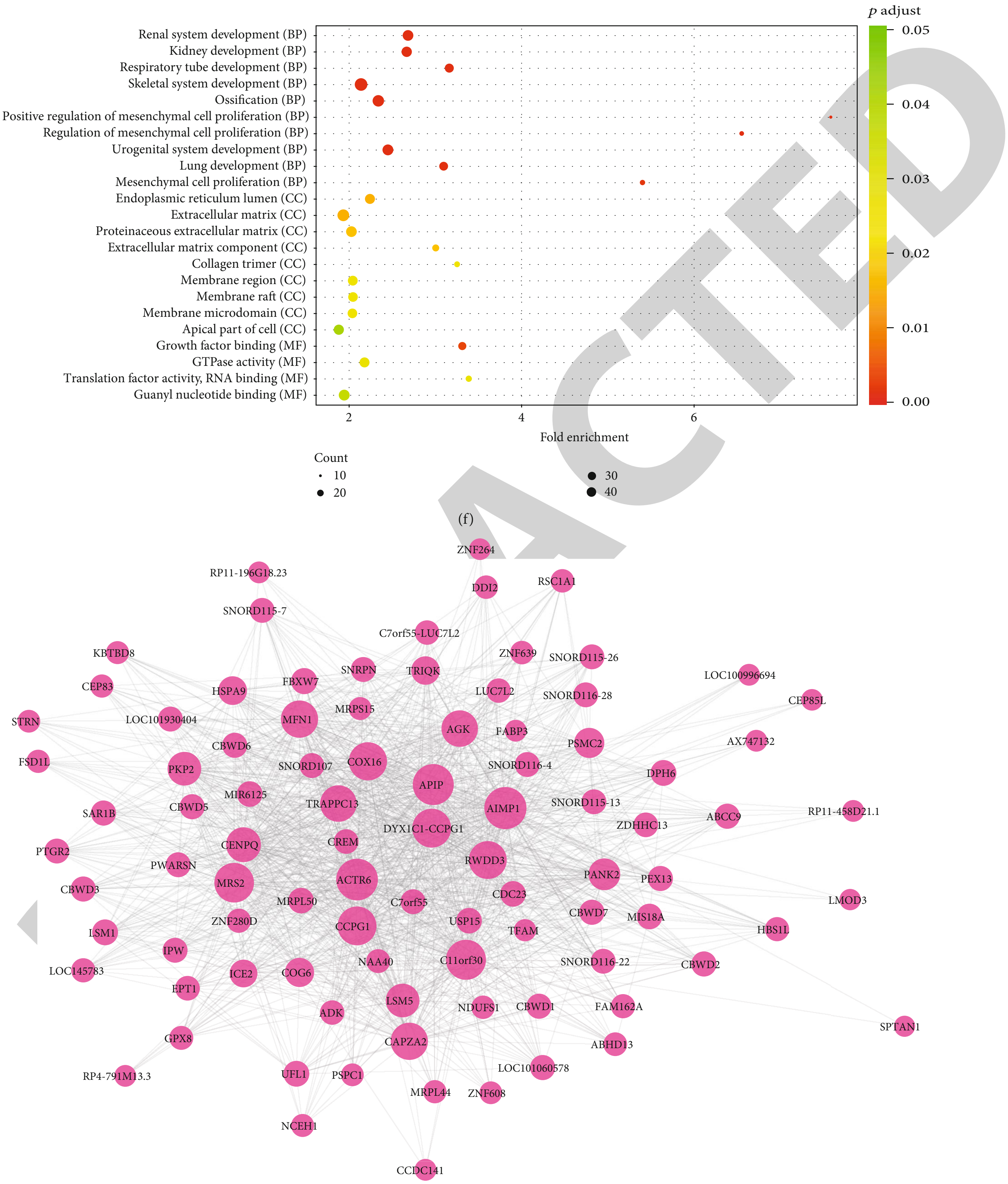
Figure 5: Continued.



(d)

(e)

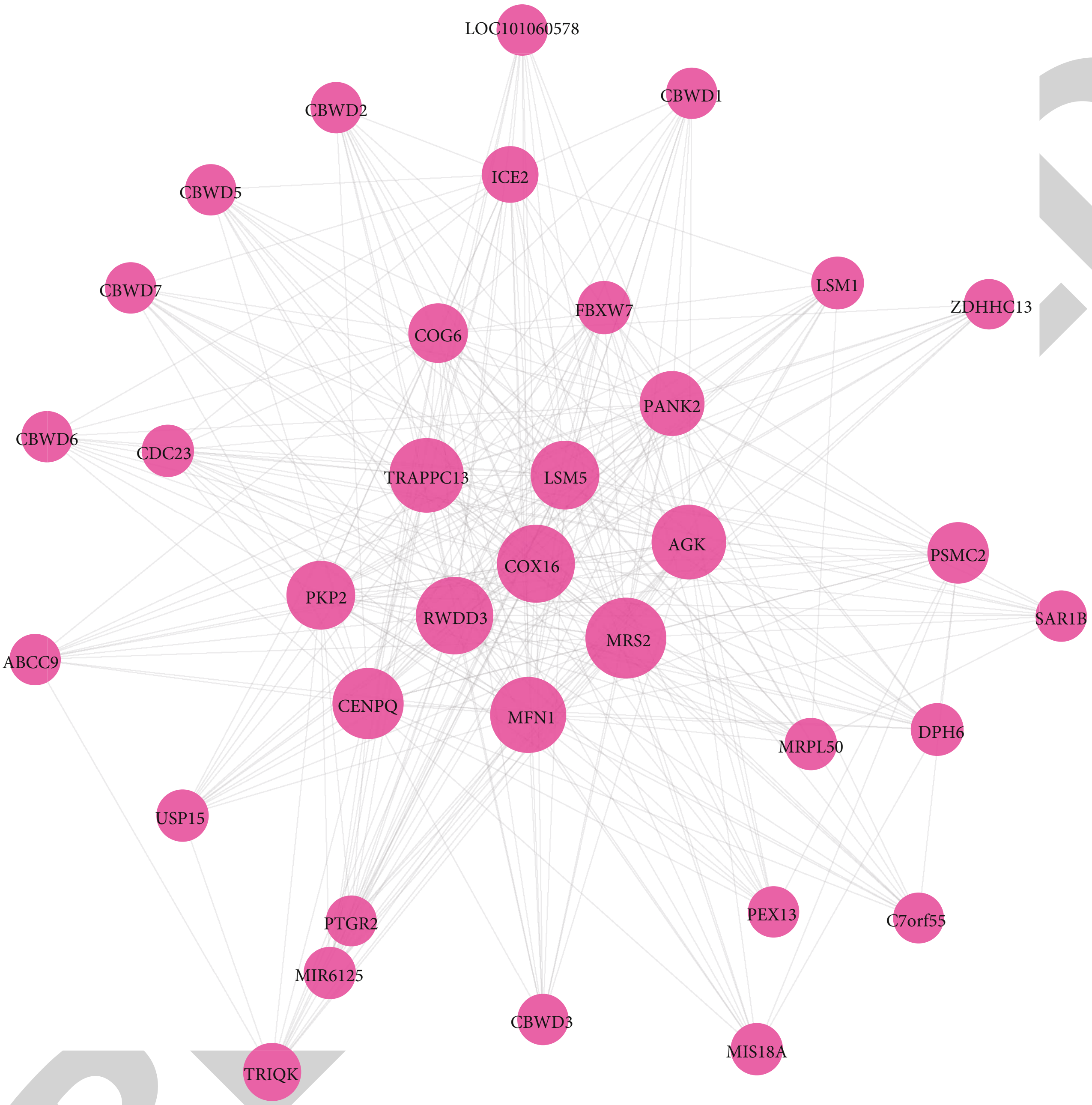
Figure 5: Continued.



(f)

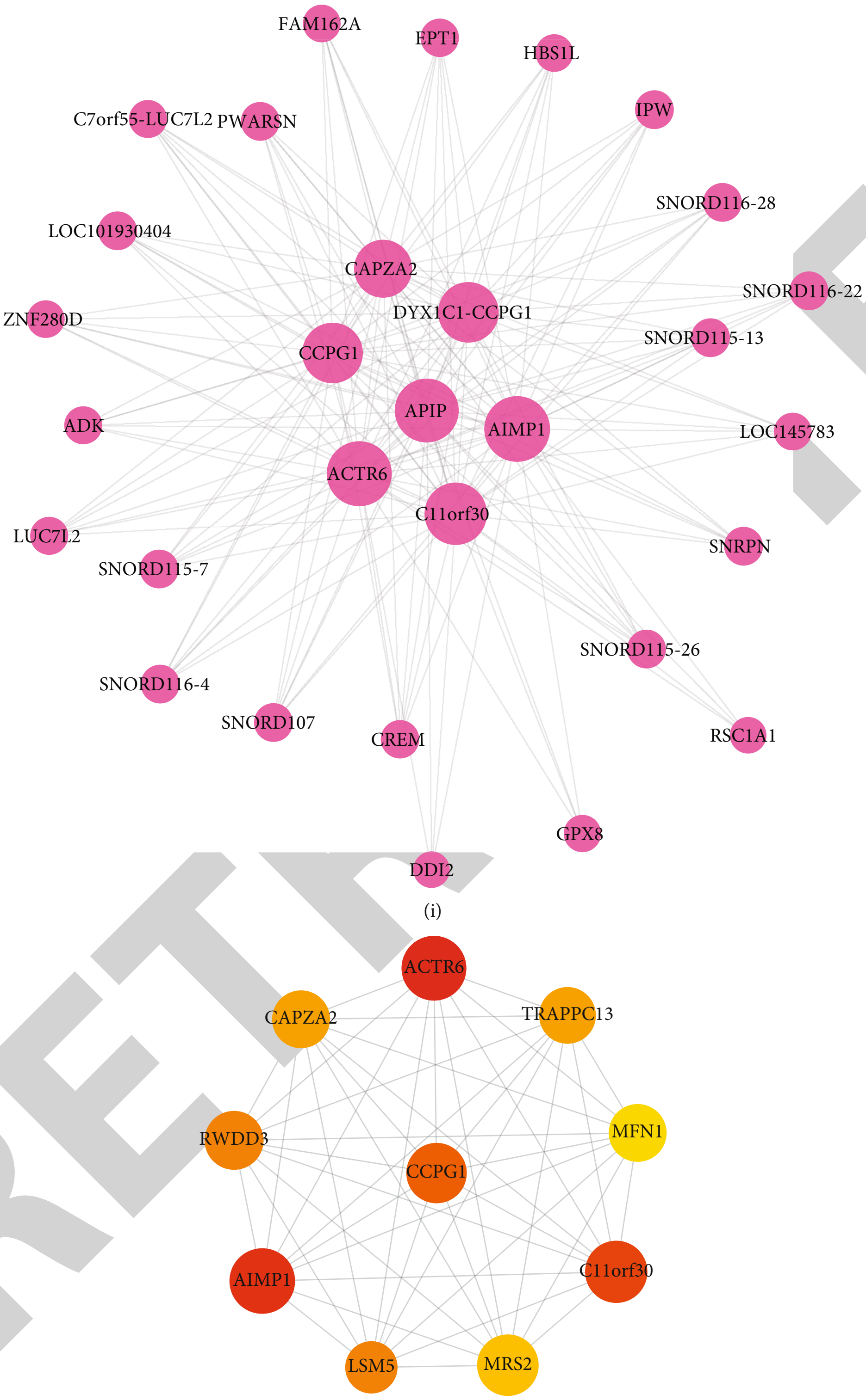
(g)

Figure 5: Continued.



(h)

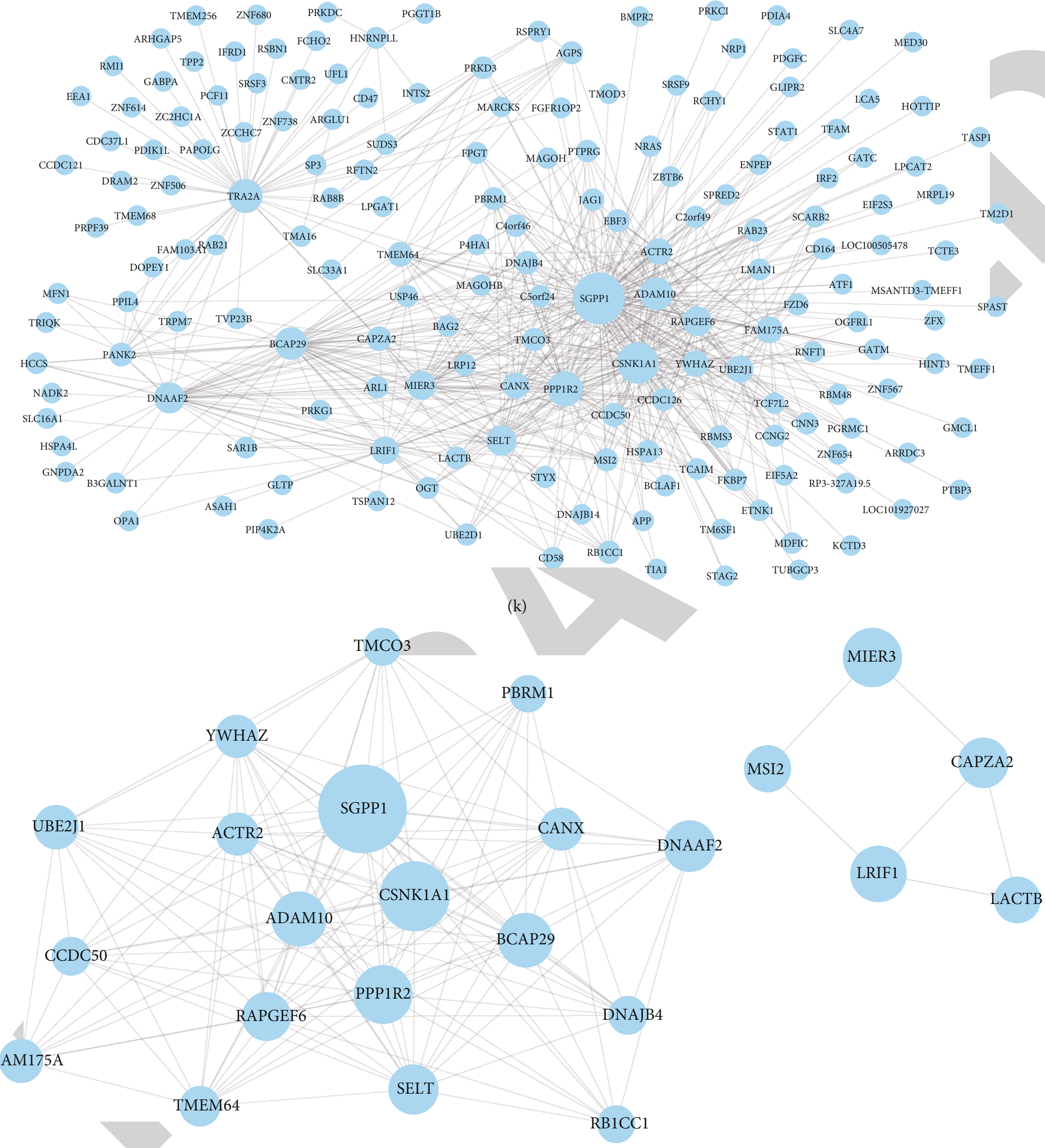
Figure 5: Continued.



(i)

(j)

Figure 5: Continued.

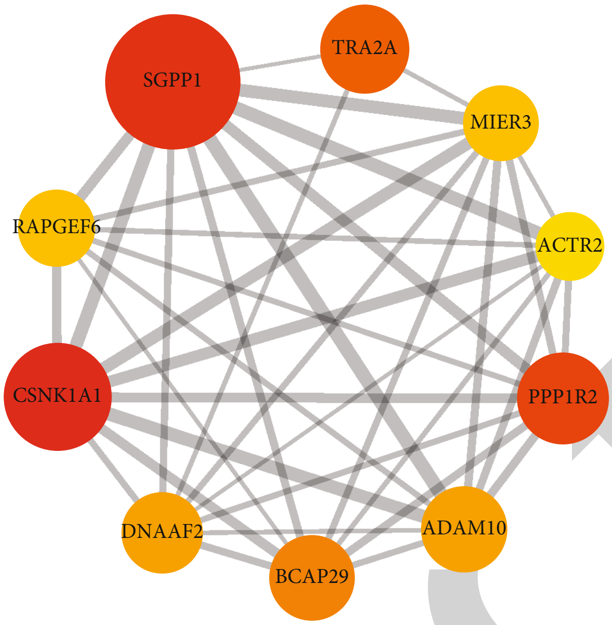


(k)

(l)

(m)

Figure 5: Continued.



(n)

Figure 5: Identification of AF-related coexpression modules and hub genes. (a) Heat map showing the relationships between 21 modules and different clinical traits. Red cell suggests positive correlation between coexpression module and clinical trait. Green suggests negative correlation between coexpression module and clinical trait. The darker the color is, the stronger the correlation is. (b, c) Scatter plots depicting the correlation between molecular membership in magenta or turquoise modules and gene significance for AF. (d) Heat map showing the correlation between AF and 21 coexpression modules. (e, f) GO enrichment analysis results of genes in magenta or turquoise modules. (g) A PPI network based on genes in magenta module. (h, i) Two core networks for genes in magenta module. (j) The top ten hub genes for magenta module according to MCC. (k) A PPI network based on genes in turquoise module. (l, m) Two core networks for genes in turquoise module. (n) The top ten hub genes for turquoise module according to MCC.

and TRAPPC13 (degree = 67). In Figure 5(k), there was a PPI network including 184 nodes on the basis of genes in turquoise module under the cutoff value of 0.3. When score = 14.667 (Figure 5(l)) and 3 (Figure 5(m)), two core networks were built for turquoise module. According to the MCC, the top ten hub genes were identified for turquoise module (Figure 5(n)), including ACTR2 (degree = 22), MIER3 (degree = 32), CSNK1A1 (degree = 78), BCAP29 (degree = 46), PPP1R2 (degree = 55), ADAM10 (degree = 47), RAPGEF6 (degree = 34), DNAAF2 (degree = 40), TRA2A (degree = 51), and SGPP1 ($n = 117$).

3.6. Whole Exome Sequencing Reveals Landscape of Mutation in AF. As shown in Figure 6(a) and Table 3, missense mutation and nonsense mutation were the top two variant classifications. Furthermore, SNP was the most common type of mutations, followed by insert and deletion (Figure 6(b)). Among all single-nucleotide variant (SNV) classifications, C > T was the most frequent mutation type, followed by T > C (Figure 6(c)). Furthermore, we counted the mutation frequencies of each sample and the median value of mutation was 66, as shown in Figure 6(d). In Figure 6(e), missense mutation was the most common mutation frequency, followed by nonsense mutation. Figure 6(f) displays the top ten mutated genes including MUC4 (71%), PHLDA1 (77%), AHNAK2 (52%), MAML3 (44%), OR2T35 (37%), SHROOM2 (25%), SAGE1 (19%), OPN1LW (19%), FLNA (19%), and FUNDC1 (19%) in AF.

PHLDA1 (in frame deletion; 77%), MUC4 (missense mutation; 71%), AHNAK2 (missense mutation; 52%), MAML3 (frame shift deletion; 44%), and OR2T35 (missense mutation; 37%) were the top five genes with mutation frequency among 52 AF samples (Figure 7(a)). Figure 7(b) displays the top 30 mutually exclusive and cooccurring genes in AF. PHLDA1 and MUC4 exhibited the highest mutation frequencies in AF (Figure 7(c)).

3.7. Validation of Key Genes in AF. The microarray expression profiles from the GSE64904 dataset including 3 AF and 3 SR samples were used for validation of key genes in AF. Firstly, the expression profiles of all samples were normalized (Figures 8(a) and 8(b)). PCA results confirmed that there was a distinct difference between AF and SR samples (Figure 8(c)). Heat map visualized the correlation between AF and SR samples based on the gene expression profiles (Figure 8(d)). Under the cutoff of adjusted $p < 0.05$ and FC > 2, 85 genes were upregulated and 73 were downregulated in AF samples compared to SR samples (Figures 8(e) and 8(f)). As shown in Figure 8(g), these genes could significantly distinguish AF from normal samples. Figure 8(h) separately visualized the top 20 upregulated and downregulated genes between AF and SR samples. However, there was no statistical difference in expression of AHNAK2, MAML3, MUC4, and PHLDA1 between AF and SR samples (Figure 8(i)). In the GSE14975 dataset, PHLDA1 expression was significantly upregulated in AF samples than normal samples (Figure 8(j)). In the

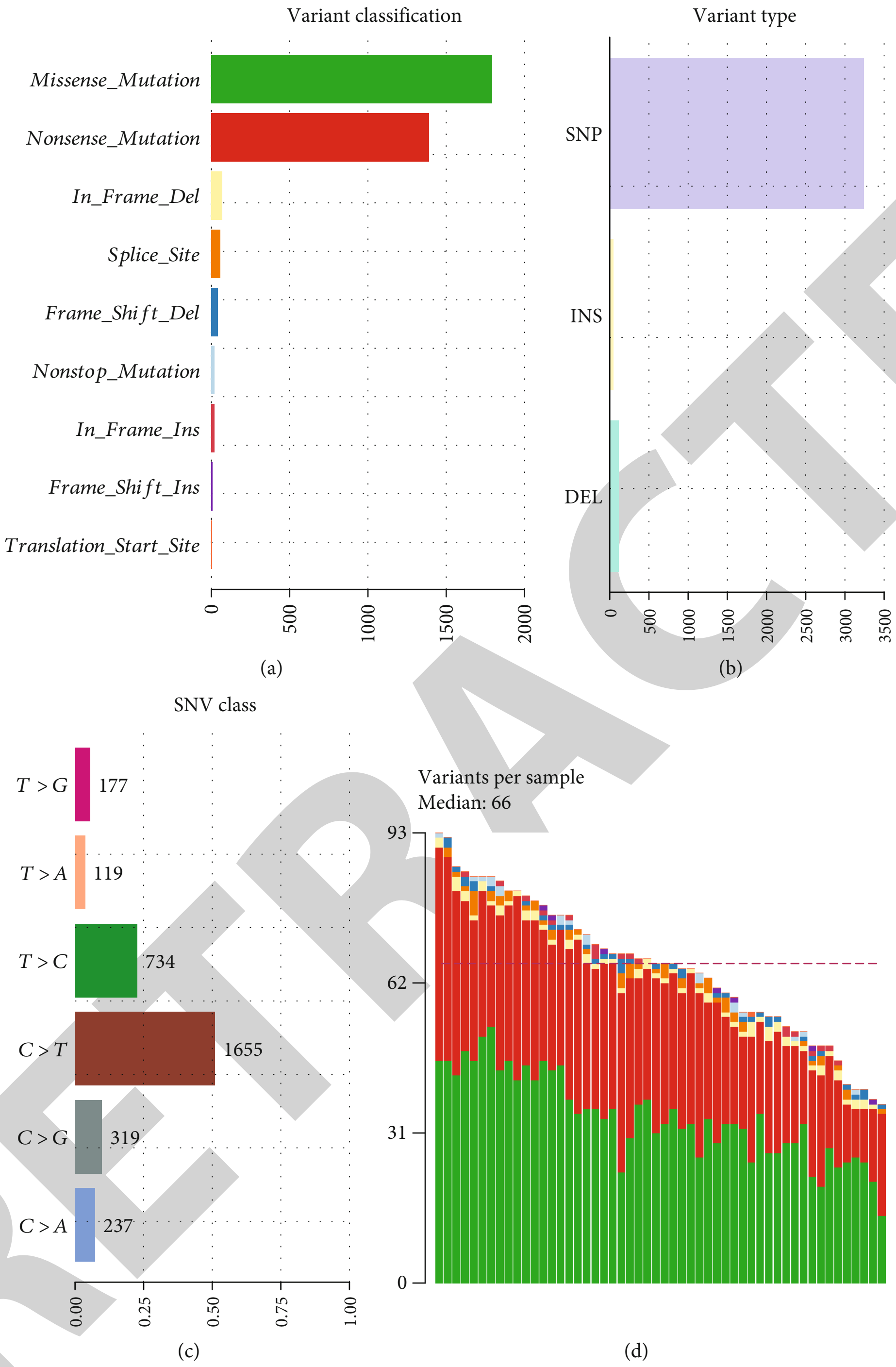


Figure 6: Continued.

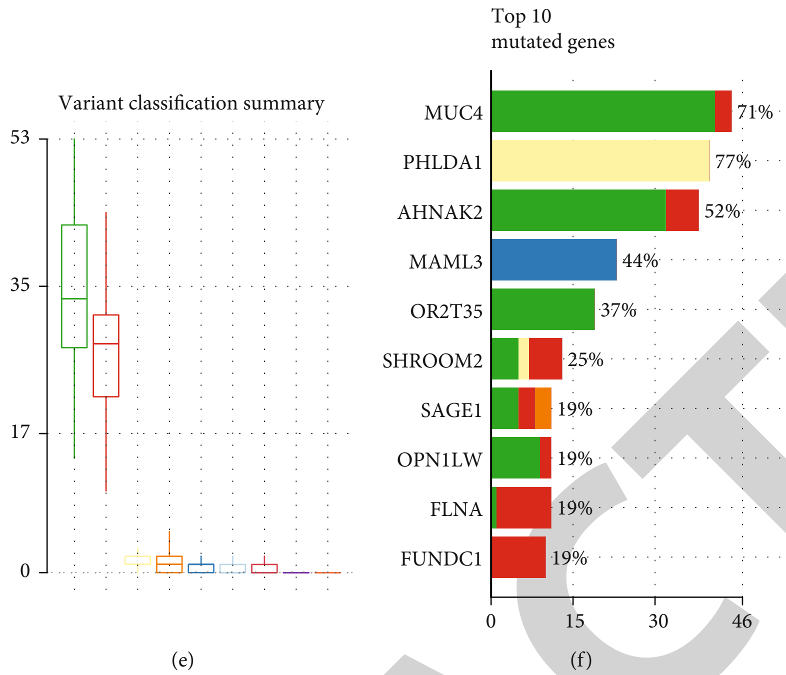


Figure 6: Landscape of mutation in AF by whole exome sequencing. (a) The mutation frequency in AF. (b) Variant types. (c) SNV classifications. (d) The distribution of variants in each AF sample. (e) The summary of variant classification. (f) The top 10 mutated genes in AF samples.

Table 3: Mutation types in AF and SR samples.

ID	Summary	Mean	Median
Frame shift deletion	41	0.788	1
Frame shift insert	7	0.135	0
In frame deletion	70	1.346	1
In frame insert	19	0.365	0
Missense mutation	1792	34.462	33.5
Nonsense mutation	1391	26.75	28
Nonstop mutation	20	0.385	0
Splice site	58	1.115	1
Translation start site	1	0.019	0
Total	3399	65.365	66

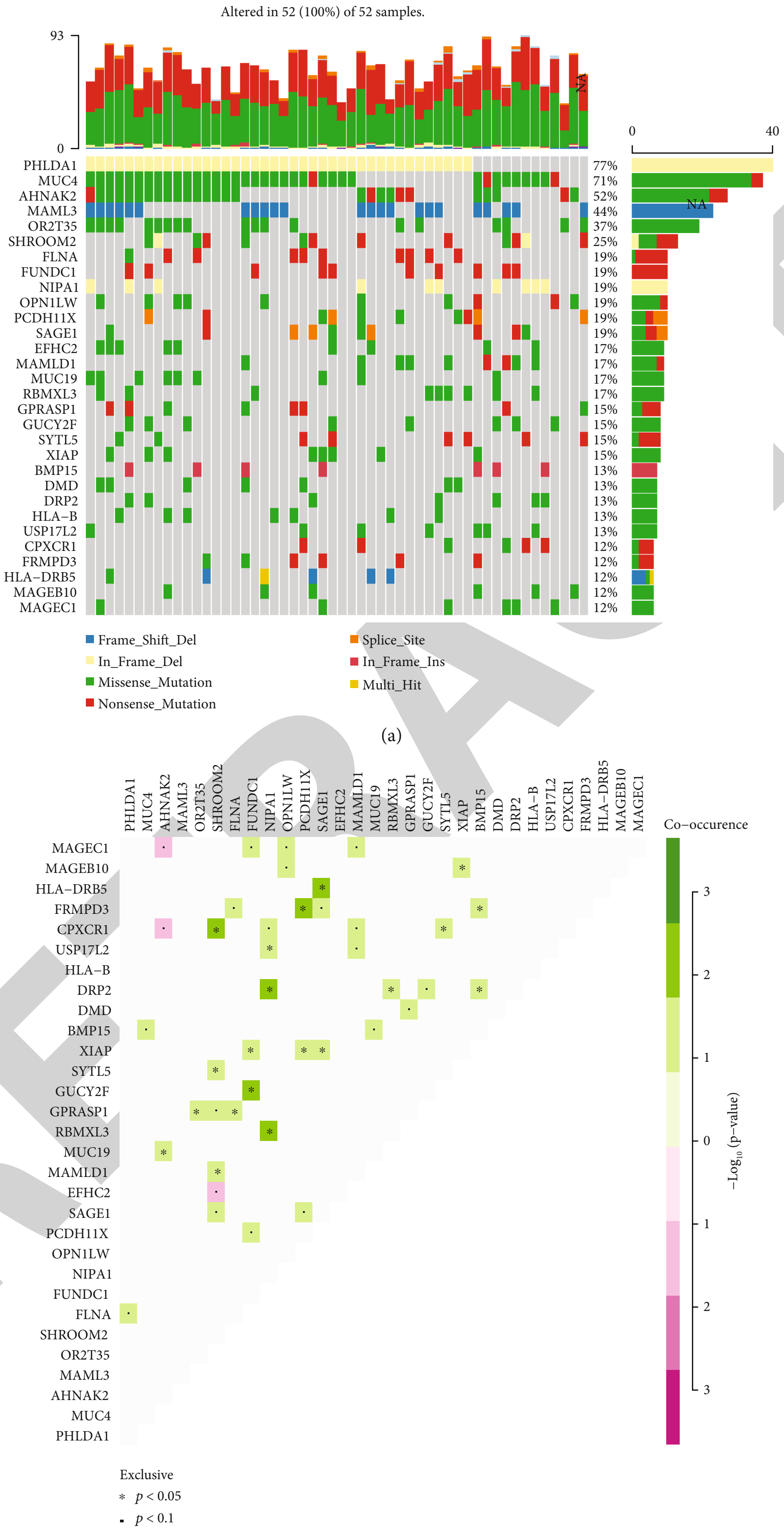
GSE79768 dataset, MUC4 expression was distinctly downregulated in AF compared to SR samples (Figure 8(k)).

4. Discussion

AF is a common cardiovascular disease. The underlying mechanisms of AF remain largely unclear. Therefore, it is essential for elucidating the underlying mechanism of AF development. This study explored pathogenesis and therapeutic targets for AF through multiomics analysis of genetics and epigenetics.

Abnormal expression is widely involved in the progression of AF. Thus, we identified DEGs between AF and normal samples in different datasets. In the GSE14975 dataset, 4 DEGs were screened for AF compared to normal samples, including 3 upregulated genes (MCEMP1, LOC100288310, and PARP15) and 1 downregulated gene (F11). However, there is no study concerning all of them in AF. In the GSE79768 dataset, 1433 DEGs were screened for AF. Functional enrichment analysis demonstrated that these DEGs were distinctly enriched in AF-related biological processes such as neutrophil activation, degranulation, and cell adhesion. It has been found that myocardial inflammatory infiltration may be a cause of AF, including neutrophil and inflammation markers [19]. Plasma vascular cell adhesion molecule-1 can predict the risk of postoperative AF [20]. In a population-based cohort study, vascular cell adhesion molecule-1 is in association with new-onset AF [21]. Combining previous studies, these DEGs could be involved in AF development via mediating key biological processes. Our KEGG enrichment analysis revealed that these DEGs were associated with regulation of actin cytoskeleton, phagosome, and leukocyte transendothelial migration. As previous studies, it has been found that several genes could regulate the cytoskeleton arrangement of cardiomyocytes in AF [22]. Atrial autophagic flux could be activated in response to AF [23].

Limited evidence suggests that abnormal DNA methylation may be related to the pathogenesis of AF. In this study, we comprehensively analyzed gene expression and DNA methylation profiles. As a result, we identified 28 differentially expressed and methylated genes for AF. As a recent study, Liu et al. identified abnormally expressed PSMC3, TINAG, and NUDT regulated by methylation for AF [5]. Among 28 differentially expressed and methylated genes, 5 have been reported to be related with AF. RHOA, CCR2, and CASP8 were hypomethylated and



(a)

(b)

Figure 7: Continued.

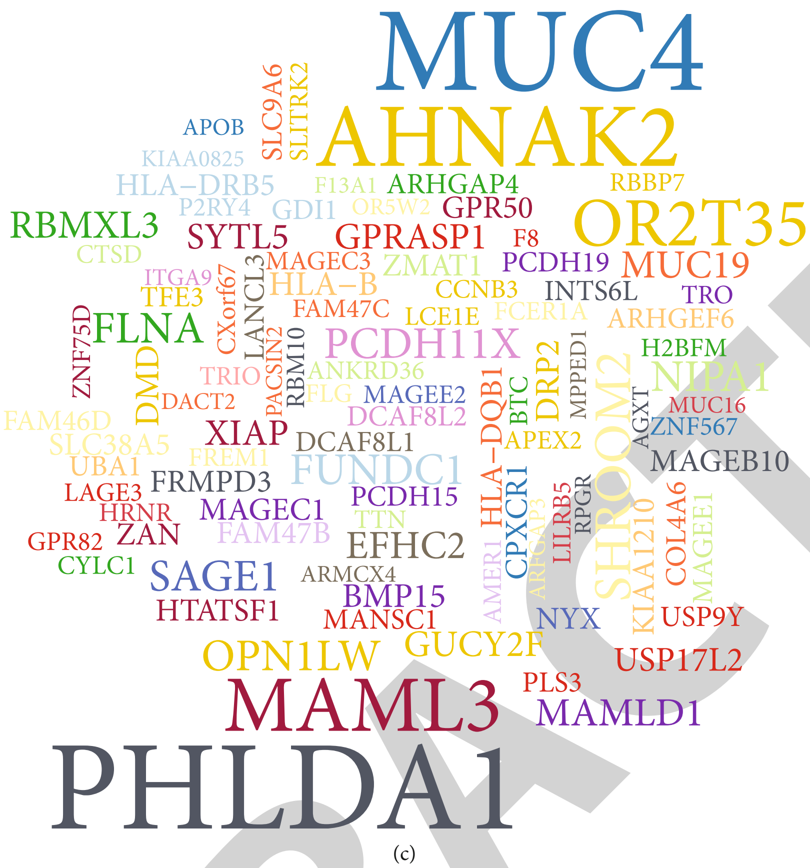
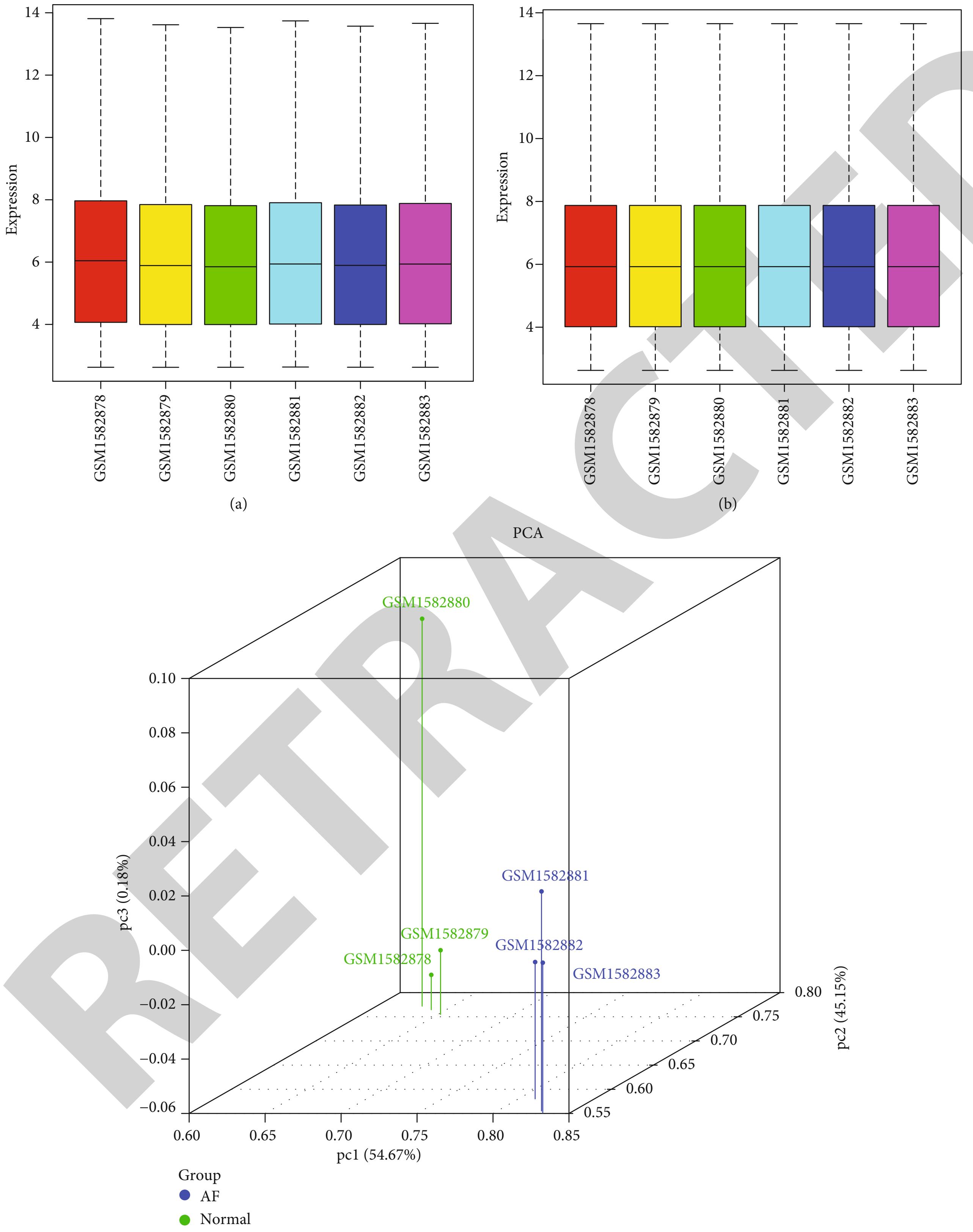


Figure 7: The most common mutated genes in AF. (a) Oncoplot showing the somatic landscape of 52 AF samples. Genes are sorted by mutation frequency. (b) Triangular matrix displaying the top 30 mutually exclusive and cooccurring genes in AF. Green is indicative of tendency to cooccurrence, while pink is indicative of tendency to exclusiveness. (c) The top 100 genes with mutation frequency. The larger the font is, the higher the mutation frequency is.

highly expressed in AF compared to normal samples. Moreover, SYNPO2L was hypermethylated and lowly expressed in AF than controls. High RHOA expression has been confirmed in leukocytes of AF patients compared to controls [24]. A recent study, CCR2 has been identified as a key gene associated with AF progression [25]. CASP8 is associated with recurrence of arrhythmia after catheter ablation of AF [26]. Intriguingly, we found that PCDHA family genes were all hypermethylated and lowly expressed in AF compared to controls, which might become underlying biomarkers for AF.

WGCNA has been widely applied to explore complex biological processes by construction of gene coexpression networks and functional key modules associated with clinical features, which could provide comprehensive insights into specific diseases or conditions [27]. In this study, WGCNA was used to identify potential mechanisms and biomarkers or therapeutic targets for AF using microarray expression profiles. Totally, 21 coexpression modules were constructed for AF. Among them, two coexpression modules (magenta and turquoise) were significantly associated with AF. Recently, Li et al. identify AF-related coexpression modules and hub genes via WGCNA [27]. Functional enrichment analysis revealed that genes in the two modules were involved in various key biological processes. For example, genes in the magenta module could participate in the proliferation of mesenchymal cells. Interstitial fibrosis plays a key role during AF progression. Fibroblast cells are differentiated from proliferative cardiac mesenchymal progenitor cells [28]. Thus, these genes might be associated with pathophysiological processes of AF. Our data suggested that genes in the turquoise were involved in fatty acid metabolic process. As previous studies, serum fatty acid binding proteins have been considered as potential biomarkers for AF [29]. Fatty acid metabolism-related genes are distinctly correlated to autophagy among patients with chronic AF [30]. Hence, it is of importance to further probe into the functions of these genes in the fatty acid metabolic process.

Previous studies on the mechanism of AF focused on specific pathophysiological functions, and relatively few studies have established a comprehensive regulatory network. Based on magenta and turquoise modules, we separately constructed PPI networks for AF, indicating that there were complex interactions between them. Hub genes usually play a core role in the PPI networks. Herein, we identified ten hub genes for magenta- (LSM5, MRS2, AIMP1, ACTR6, MFN1, RWDD3, CAPZA2, C11orf30, CCPG1, and TRAPPC13) and turquoise-related (ACTR2, MIER3, CSNK1A1, BCAP29, PPP1R2, ADAM10, RAPGEF6,

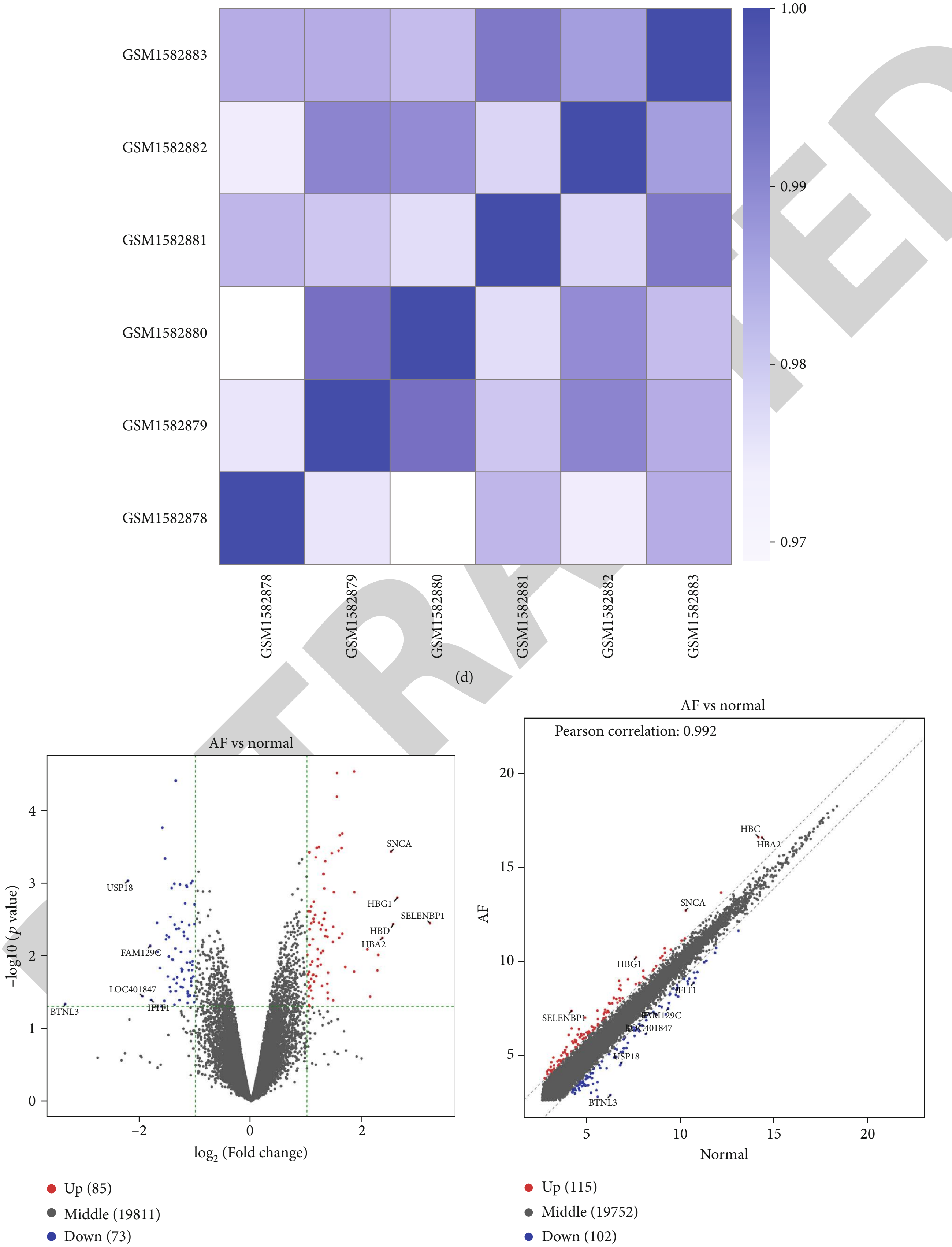


(a)

(b)

(c)

Figure 8: Continued.

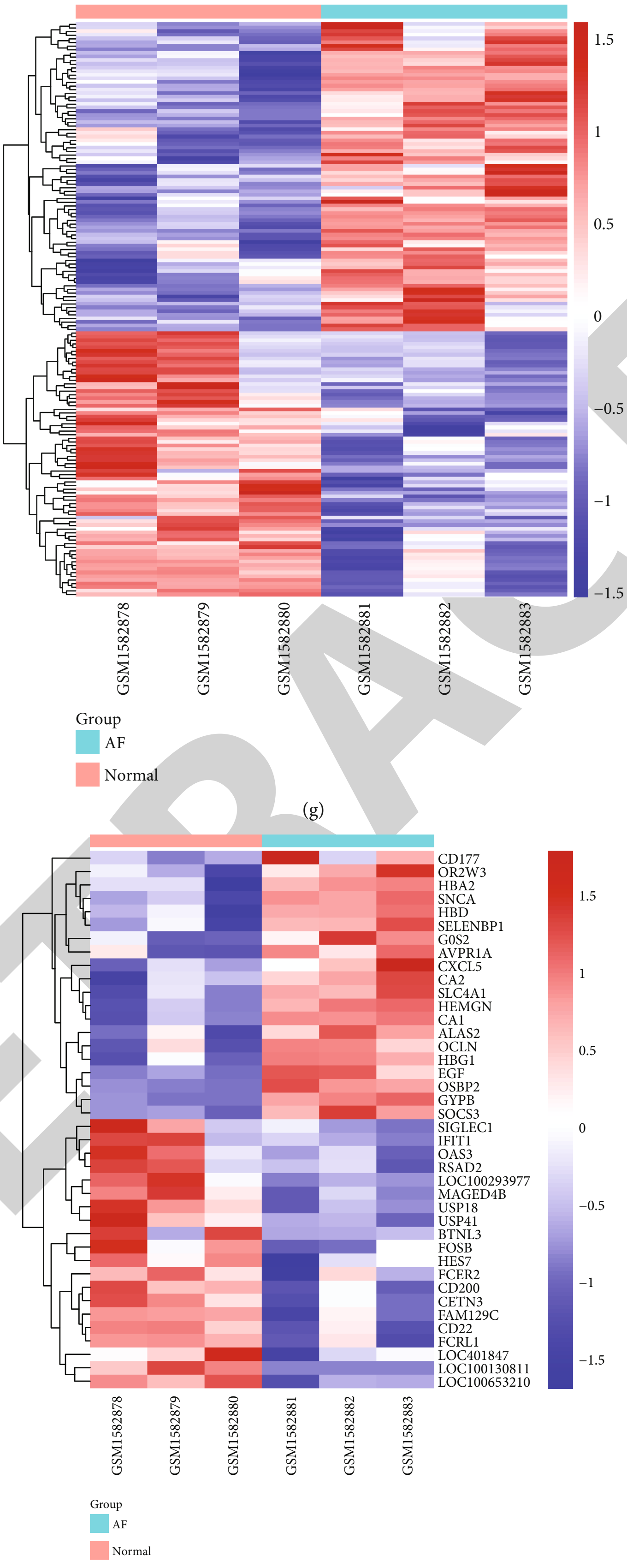


(d)

(e)

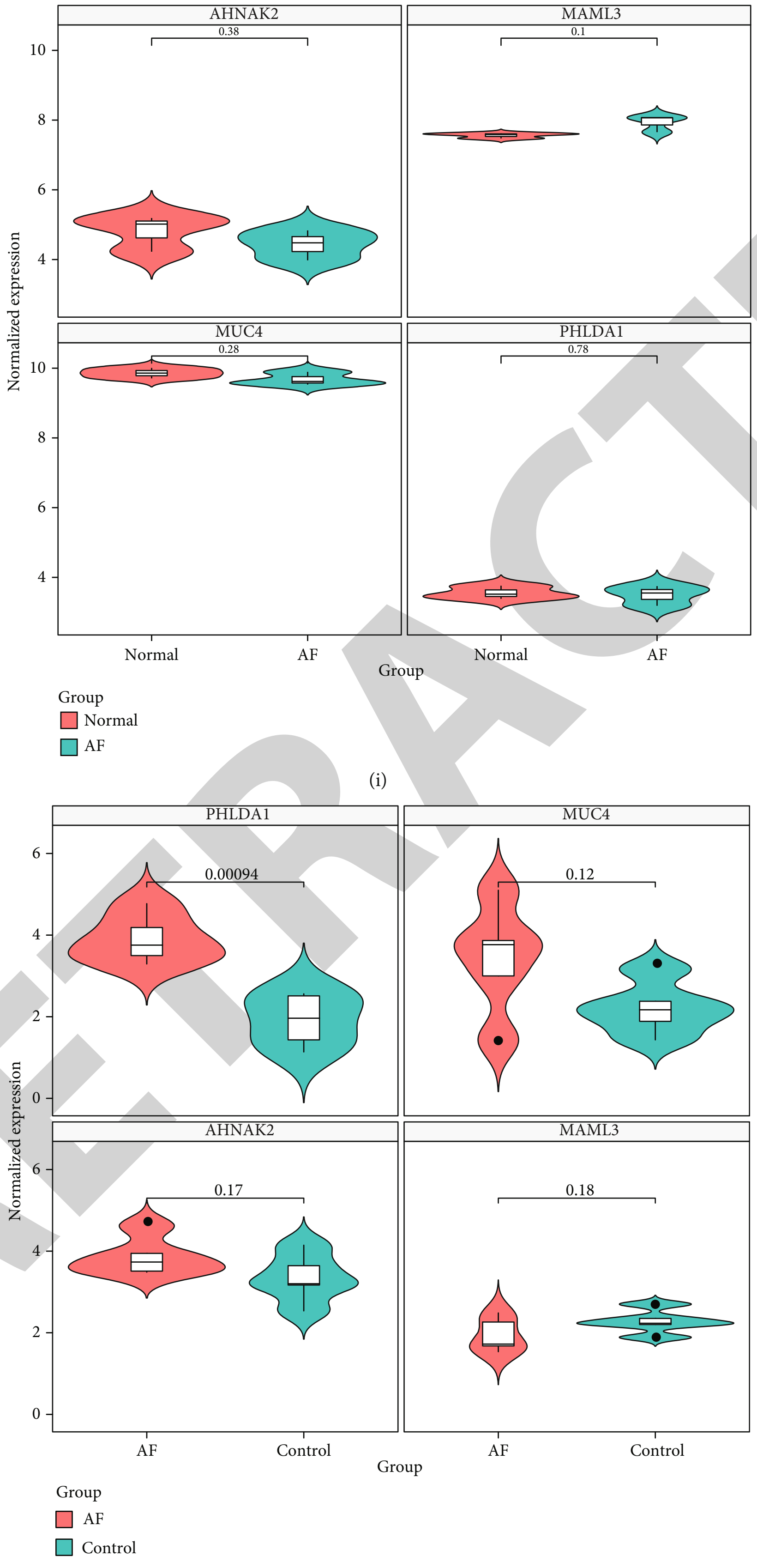
(f)

Figure 8: Continued.



(h)

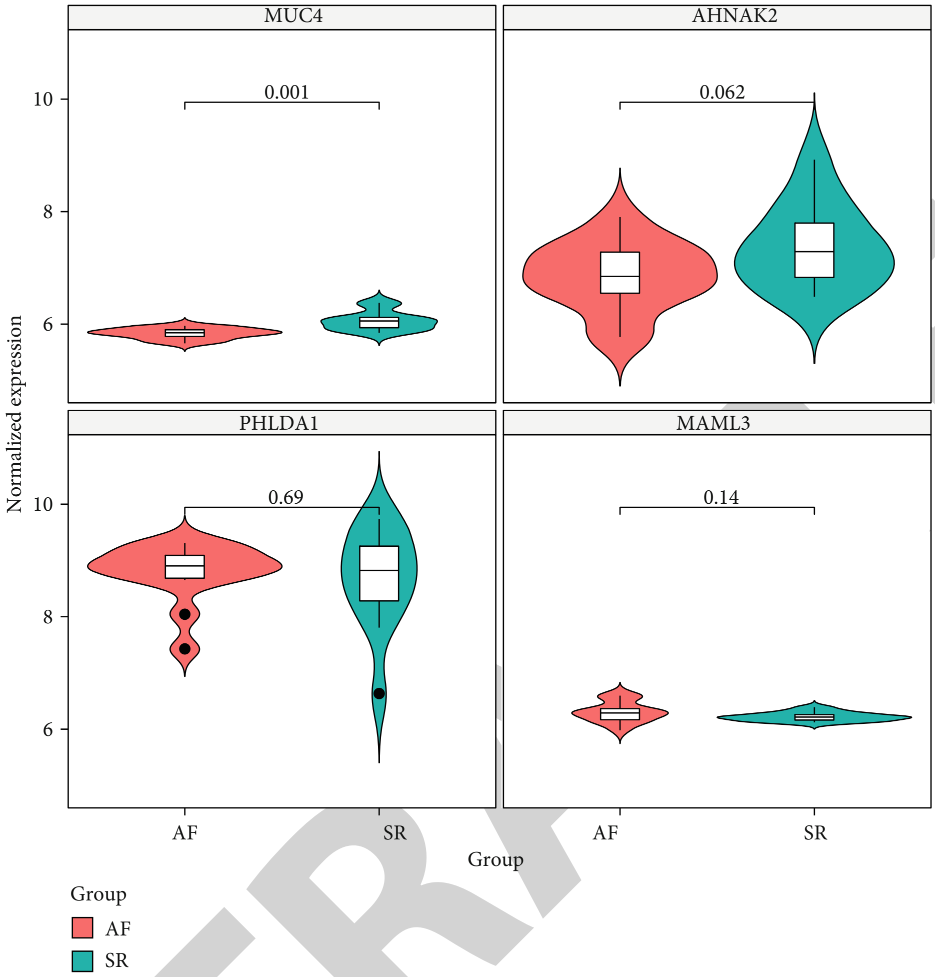
Figure 8: Continued.



(i)

(j)

Figure 8: Continued.



(k)

Figure 8: Validation of key genes in AF. (a, b) Box plots showing the expression levels in AF and SR samples before and after normalization. (c) PCA results confirming the difference between AF and SR samples. (d) Heat map visualizing the correlation between different samples based on the gene expression profile. (e, f) Violin plots and scatter plots showing DEGs between the two groups. (g) Hierarchical clustering analysis results demonstrating the expression patterns of DEGs between the two groups. (h) The top 40 DEGs in AF and SR samples. The expression levels of AHNAK2, MAML3, MUC4, and PHLDA1 were detected between AF and SR samples in the GSE64904 (i), GSE14975 (j), and GSE79768 (k).

DNAAF2, TRA2A, and SGPP1) PPI networks. Among them, high ADAM10 expression has been confirmed to be in relationship with AF [31]. Nevertheless, most of them remain unclear in AF.

SNPs have been widely found on different AF susceptibility loci [32]. Herein, Whole exome sequencing was performed for 52 AF samples. Our data suggested that SNP (especially C > T and T > C) was the most mutation type for AF, which was consistent with previous studies [33]. MUC4, PHLDA1, AHNAK2, and MAML3 were the most frequently four mutated genes for AF. Their abnormal expression was validated in independent datasets. Nevertheless, at present, no studies have reported their mutations in AF.

Collectively, this study expounded pathogenesis and underlying molecular mechanism for AF. Moreover, we provided promising therapeutic targets for AF, which could be worth further exploring in future studies.

5. Conclusion

Through multiomics analysis of genetics and epigenetics, we identified abnormal expressed and methylated genes in multiple datasets. Key coexpression modules were constructed, and hub genes were screened for AF. Furthermore, whole exome sequence revealed mutated genes such as PHLDA1 and MUC4 in AF. Taken together, our study provided possible therapeutic targets and a new insight into the pathogenesis of AF.

Abbreviations

AF:	Atrial fibrillation
GEO:	Gene Expression Omnibus
DEGs:	Differentially expressed genes
WGCNA:	Weighted gene coexpression network analysis
SNP:	Single-nucleotide polymorphism

CNA:	Copy number alternation
SNVs:	Single-nucleotide variants
indels:	Insertions-deletions
LA:	Left atrial
SR:	Sinus rhythm
FDR:	False discovery rate
FC:	Fold change
GO:	Gene Ontology
KEGG:	Kyoto Encyclopedia of Genes and Genomes
BP:	Biological process
CC:	Cellular component
MF:	Molecular function
GS:	Gene significance
MM:	Module membership
PPI:	Protein-protein interaction
MCODE:	Molecular complex detection
MCC:	Maximal clique centrality
maf:	Mutation annotation format.

Data Availability

The datasets analyzed during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest

The authors declare no conflicts of interest.

Authors' Contributions

Li Liu and Jianjun Huang contributed equally to this work.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (81560076), the Guangxi Natural Science Foundation (2018JJB140358), the Middle-aged and Young Teachers in Colleges and Universities in Guangxi Basic Ability Promotion Project (2017KY0514), the Starting Research Projects for Introducing Dr of Youjiang Medical University for Nationalities (2015bsky002), the First Batch of High-level Talent Scientific Research Projects of the Affiliated Hospital of Youjiang Medical University for Nationalities in 2019 (R20196316).

References

[1] X. Huang, Y. Li, J. Zhang, X. Wang, Z. Li, and G. Li, "The molecular genetic basis of atrial fibrillation," *Human Genetics*, vol. 139, no. 12, pp. 1485–1498, 2020.

[2] D. Tousoulis, "Biomarkers in atrial fibrillation; from pathophysiology to diagnosis and treatment," *Current Medicinal Chemistry*, vol. 26, no. 5, pp. 762–764, 2019.

[3] H. Tao, K. H. Shi, J. J. Yang, and J. Li, "Epigenetic mechanisms in atrial fibrillation: new insights and future directions," *Trends in Cardiovascular Medicine*, vol. 26, no. 4, pp. 306–318, 2016.

[4] W. Zhu, L. Xie, J. Han, and X. Guo, "The application of deep learning in cancer prognosis prediction," *Cancers (Basel)*, vol. 12, no. 3, p. 603, 2020.

[5] B. Liu, X. Shi, K. Ding et al., "The joint analysis of multi-omics data revealed the methylation-expression regulations in atrial fibrillation," *Frontiers in Bioengineering and Biotechnology*, vol. 8, p. 187, 2020.

[6] K. Shen, T. Tu, Z. Yuan et al., "DNA methylation dysregulations in valvular atrial fibrillation," *Clinical Cardiology*, vol. 40, no. 9, pp. 686–691, 2017.

[7] A. F. van Ouwerkerk, F. M. Bosada, K. van Duijvenboden et al., "Identification of atrial fibrillation associated genes and functional non-coding variants," *Nature Communications*, vol. 10, no. 1, p. 4755, 2019.

[8] O. Adam, D. Lavall, K. Theobald et al., "Rac1-induced connective tissue growth factor regulates connexin 43 and N-cadherin expression in atrial fibrillation," *Journal of the American College of Cardiology*, vol. 55, no. 5, pp. 469–480, 2010.

[9] F. C. Tsai, Y. C. Lin, S. H. Chang et al., "Differential left-to-right atria gene expression ratio in human sinus rhythm and atrial fibrillation: implications for arrhythmogenesis and thrombogenesis," *International Journal of Cardiology*, vol. 222, pp. 104–112, 2016.

[10] J. Zhou, J. Gao, Y. Liu et al., "Human atrium transcript analysis of permanent atrial fibrillation," *International Heart Journal*, vol. 55, no. 1, pp. 71–77, 2014.

[11] M. E. Ritchie, B. Phipson, D. I. Wu et al., "limma powers differential expression analyses for RNA-sequencing and microarray studies," *Nucleic Acids Research*, vol. 43, no. 7, article e47, 2015.

[12] G. Yu, L. G. Wang, Y. Han, and Q. Y. He, "clusterProfiler: an R package for comparing biological themes among gene clusters," *OMICS*, vol. 16, no. 5, pp. 284–287, 2012.

[13] P. Langfelder and S. Horvath, "WGCNA: an R package for weighted correlation network analysis," *BMC Bioinformatics*, vol. 9, no. 1, p. 559, 2008.

[14] D. Szklarczyk, A. L. Gable, D. Lyon et al., "STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets," *Nucleic Acids Research*, vol. 47, no. D1, pp. D607–d613, 2019.

[15] S. Paul, M. Andrew, O. O. Baliga Nitin et al., "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome Research*, vol. 13, no. 11, pp. 2498–2504, 2003.

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

[17] C. H. Chin, S. H. Chen, H. H. Wu, C. W. Ho, M. T. Ko, and C. Y. Lin, "cytoHubba: identifying hub objects and subnetworks from complex interactome," *BMC Systems Biology*, vol. 8, Supplement 4, p. S11, 2014.

[18] A. Mayakonda, D. C. Lin, Y. Assenov, C. Plass, and H. P. Koeffler, "Maftools: efficient and comprehensive analysis of somatic variants in cancer," *Genome Research*, vol. 28, no. 11, pp. 1747–1756, 2018.

[19] M. L. Fontes, J. P. Mathew, H. M. Rinder, D. Zelterman, B. R. Smith, and C. S. Rinder, "Atrial fibrillation after cardiac surgery/cardiopulmonary bypass is associated with monocyte activation," *Anesthesia & Analgesia*, vol. 101, no. 1, pp. 17–23, 2005.

[20] H. Verdejo, J. Roldan, L. Garcia et al., "Systemic vascular cell adhesion molecule-1 predicts the occurrence of postoperative atrial fibrillation," *International Journal of Cardiology*, vol. 150, no. 3, pp. 270–276, 2011.

Hindawi
BioMed Research International
Volume 2024, Article ID 9762690, 1 page
https://doi.org/10.1155/2024/9762690

Retraction

Retracted: Weighted Gene Coexpression Network Analysis Uncovers Critical Genes and Pathways for Multiple Brain Regions in Parkinson's Disease

BioMed Research International

Received 12 March 2024; Accepted 12 March 2024; Published 20 March 2024

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope

(2) Discrepancies in the description of the research reported

(3) Discrepancies between the availability of data and the research described

(4) Inappropriate citations

(5) Incoherent, meaningless and/or irrelevant content included in the article

(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] J. Huang, L. Liu, L. Qin, H. Huang, and X. Li, "Weighted Gene Coexpression Network Analysis Uncovers Critical Genes and Pathways for Multiple Brain Regions in Parkinson's Disease," *BioMed Research International*, vol. 2021, Article ID 6616434, 46 pages, 2021.

Hindawi
BioMed Research International
Volume 2021, Article ID 6616434, 46 pages
https://doi.org/10.1155/2021/6616434

Research Article

Weighted Gene Coexpression Network Analysis Uncovers Critical Genes and Pathways for Multiple Brain Regions in Parkinson's Disease

Jianjun Huang,[1] Li Liu,[2] Lingling Qin,[3] Hehua Huang,[4] and Xue Li[5]

[1]*Department of Neurology, Youjiang Medical University for Nationalities, Affiliated Hospital of Youjiang Medical University for Nationalities, Baise, 533000 Guangxi, China*
[2]*Department of Cardiology, Youjiang Medical University for Nationalities, Affiliated Hospital of Youjiang Medical University for Nationalities, Baise, 533000 Guangxi, China*
[3]*Department of Medical Quality Management, Affiliated Hospital of Youjiang Medical University for Nationalities, Baise, 533000 Guangxi, China*
[4]*Department of Anatomy, Youjiang Medical University for Nationalities, Baise, 533000 Guangxi, China*
[5]*Department of Electrophysiology, Affiliated Hospital of Youjiang Medical University for Nationalities, Baise, 533000 Guangxi, China*

Correspondence should be addressed to Jianjun Huang; jianjun453@163.com

Received 4 December 2020; Revised 21 January 2021; Accepted 8 February 2021; Published 15 March 2021

Academic Editor: Parameshachari B. D.

Objective. In this study, we aimed to identify critical genes and pathways for multiple brain regions in Parkinson's disease (PD) by weighted gene coexpression network analysis (WGCNA). *Methods*. From the GEO database, differentially expressed genes (DEGs) were separately identified between the substantia nigra, putamen, prefrontal cortex area, and cingulate gyrus of PD and normal samples with the screening criteria of p value < 0.05 and $|\log_2 \text{fold change (FC)}| > 0.585$. Then, a coexpression network was presented by the WGCNA package. Gene modules related to PD were constructed. Then, PD-related DEGs were used for construction of PPI networks. Hub genes were determined by the cytoHubba plug-in. Functional enrichment analysis was then performed. *Results*. DEGs were identified for the substantia nigra (17 upregulated and 52 downregulated genes), putamen (317 upregulated and 317 downregulated genes), prefrontal cortex area (39 upregulated and 72 downregulated genes), and cingulate gyrus (116 upregulated and 292 downregulated genes) of PD compared to normal samples. Gene modules were separately built for the four brain regions of PD. PPI networks revealed hub genes for the substantia nigra (SLC6A3, SLC18A2, and TH), putamen (BMP4 and SNAP25), prefrontal cortex area (SNAP25), and cingulate gyrus (CTGF, CDH1, and COL5A1) of PD. These DEGs in multiple brain regions were involved in distinct biological functions and pathways. GSEA showed that these DEGs were all significantly enriched in electron transport chain, proteasome degradation, and synaptic vesicle pathway. *Conclusion*. Our findings revealed critical genes and pathways for multiple brain regions in PD, which deepened the understanding of PD-related molecular mechanisms.

1. Introduction

Parkinson's disease (PD) is the second most common neurodegenerative disease related to the loss of dopaminergic neurons globally [1]. It is characterized by tremor and slow movement, affecting approximately 7 million people worldwide, most of whom are elderly [2]. Male and age are independent risk factors of PD [3]. Due to its complex pathogenesis, symptomatic treatment is mainly applied such as the replacement of dopamine [4]. Molecular biomarkers have been proven as promising clinical tools for PD diagnosis [5]. Thus, it is an urgent need to uncover new strategies for early diagnosis and therapeutic intervention to improve the quality of life of the affected population.

Understanding the mechanisms of PD at the molecular levels is valuable for clinical treatment. With the widespread

Table 1: Dataset information from the GEO database.

Location	Accession	Platform	Type	Number
Substantia nigra	GSE20292	GPL96	Microarray	18 control vs. 11 PD
Substantia nigra	GSE7621	GPL570	Microarray	9 control vs. 16 PD
Putamen	GSE20291	GPL96	Microarray	20 control vs. 15 PD
Prefrontal cortex area	GSE20168	GPL96	Microarray	15 control vs. 14 PD
Prefrontal cortex area	GSE68719	GPL11154	RNA-seq	44 control vs. 29 PD
Cingulate gyrus	GSE110716	GPL11153	RNA-seq	8 control vs. 8 PD

use of microarray and RNA-seq technologies, genes related to PD have been widely identified, which help decipher the complex pathogenesis of PD, thereby promoting the development of effective drug targets and preventing the occurrence of PD at an early stage [6–8]. Gene coexpression networks are widely used for function prediction and identification of genes modules in a set of samples including PD [9]. As a method of bioinformatics research, WGCNA is usually applied to reveal the correlation between genes in different samples [10–12]. However, the candidate biomarkers for clinical gene therapy of PD are unclear. In this study, the microarray and RNA-seq datasets from GEO were used to identify DEGs between multiple brain regions of PD and normal samples. Then, through WGCNA, PD-related key modules were constructed. Further functional enrichment analysis was carried out to evaluate the potential functions of genes in key modules.

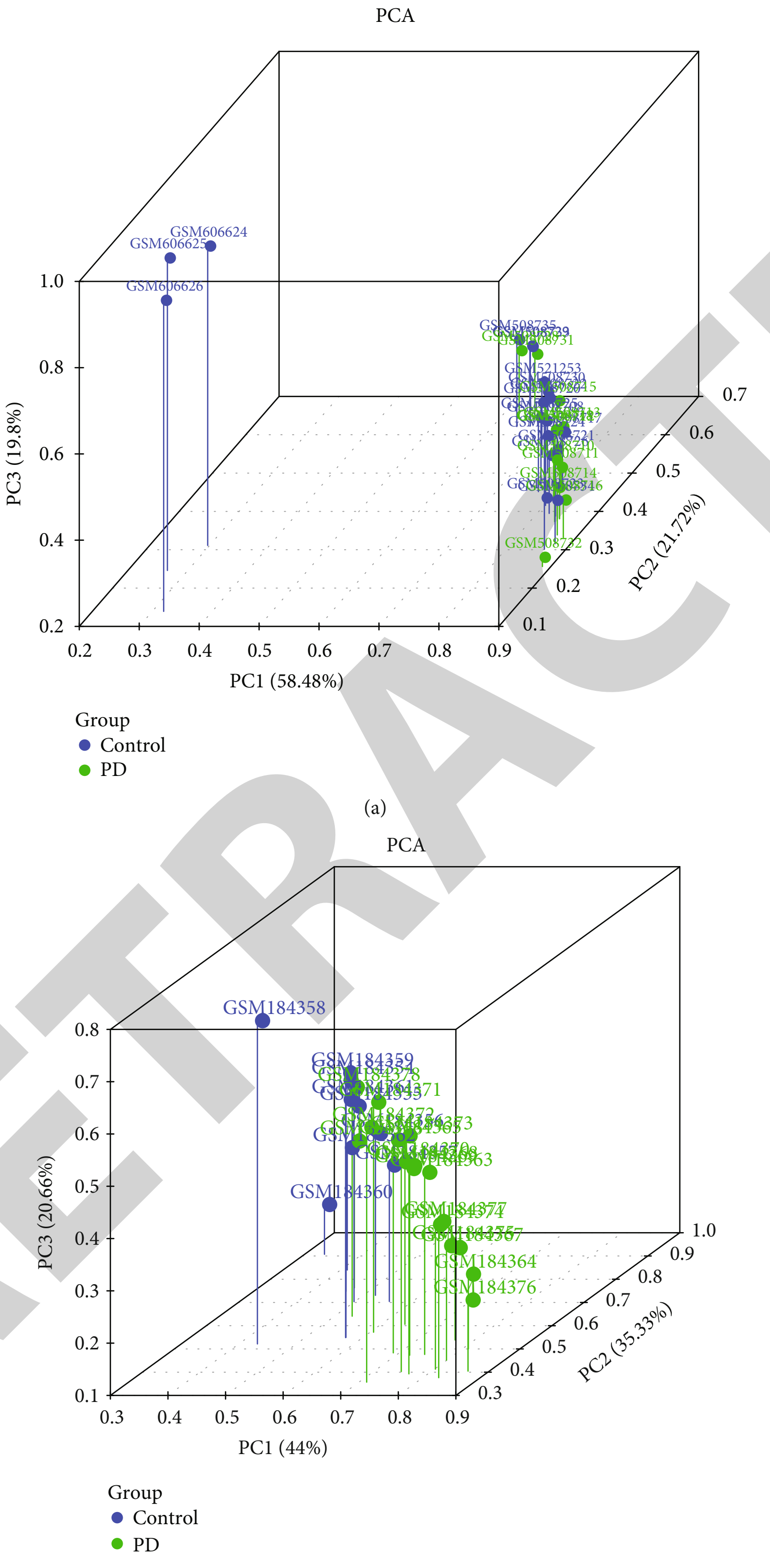
2. Materials and Methods

2.1. Data Collection and Preprocessing. Expression profiles of PD (GSE20292, GSE7621, GSE20291, GSE20168, GSE68719, and GSE110716) were downloaded from the Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) database (Table 1). The GSE20291 dataset included 11 PD substantia nigra samples and 18 normal samples on the GPL96 (Affymetrix Human Genome U133A Array) platform. The GSE7621 dataset was composed of 16 PD substantia nigra samples and 9 control samples on the GPL570 (Affymetrix Human Genome U133 Plus 2.0 Array) platform. The GSE20291 dataset covered 15 PD putamen samples and 20 control samples on the GPL96 (Affymetrix Human Genome U133A Array) platform. The GSE20168 dataset included 14 prefrontal cortex PD samples and 15 normal samples on the platform of GPL96 (Affymetrix Human Genome U133A Array). There were 29 prefrontal cortex samples and 44 normal samples in the GSE68719 dataset on the GPL11154 (Illumina HiSeq 2000 (Homo sapiens)) platform. The GSE110716 dataset was composed of 8 cingulate gyrus PD samples and 8 normal samples on the platform of GPL15433 (Illumina HiSeq 1000 (Homo sapiens)). Raw data was standardized by $\log_2$ conversion. Principal component analysis (PCA) was presented to detect and remove outliers and to find samples with high similarity. Furthermore, the correlation of gene expression levels between samples was analyzed.

2.2. Differential Expression Analysis. Microarray expression data were used for differential expression analysis between the PD group and the control group using the limma package [13]. Before analyzing the expression differences, the probes were annotated. For the case where multiple probes corresponded to the same gene, the average value of multiple probes was taken as the expression value of the gene. For the case where there were multiple datasets at the same site, DEGs of multiple datasets were overlapped as the final significant DEGs for downstream analysis. The high-throughput sequencing data were utilized for DEGs between the PD group and the control group by the edgeR package [14]. The screening threshold for a significant difference in gene expression was adjusted p value < 0.05 and $|\log_2\text{fold change (FC)}| > 0.585$.

2.3. Weighted Gene Coexpression Network Analysis (WGCNA). In this study, the WGCNA package was used to realize WGCNA [15]. Through the goodGeneSample function, a hierarchical clustering tree was constructed for all samples and outliers of which node height was significantly higher than other samples were removed. The gene coexpression similarity matrix was composed of the absolute value of the Pearson correlation coefficient between two genes. The correlation matrix was constructed as follows: $S = [S_{i,j}]$ (i and j indicate the i^{th} and j^{th} gene). Soft threshold β value was then calculated according to the following formula: $a_{i,j} = \text{power}\,(S_{i,j}, \beta) = |S_{i,j}|^{\beta}$ ($a_{i,j}$ indicates the adjacency function between the i^{th} and j^{th} genes). To follow the principle of nonscale network, $R^2 > 0.8$ was set. After determining the soft threshold β through the pickSoftThreshold function, the correlation matrix $S = [S_{i,j}]$ was converted into adjacency matrix $A = [A_{i,j}]$ by the pickSoftThreshold function. Topological overlap measure (TOM) was performed to calculate the degree of association between genes as follows: $\text{TOM}_{IJ} = (\sum_u a_{iu} a_{uj} + a_{ij})/(\min\,(ki, kj) + 1 - a_{ij})$ (a_{ij} is $[0, 1]$). Gene modules were divided based on the high topological overlap between genes in the modules. The dynamic cutting tree algorithm was used to calculate gene modules.

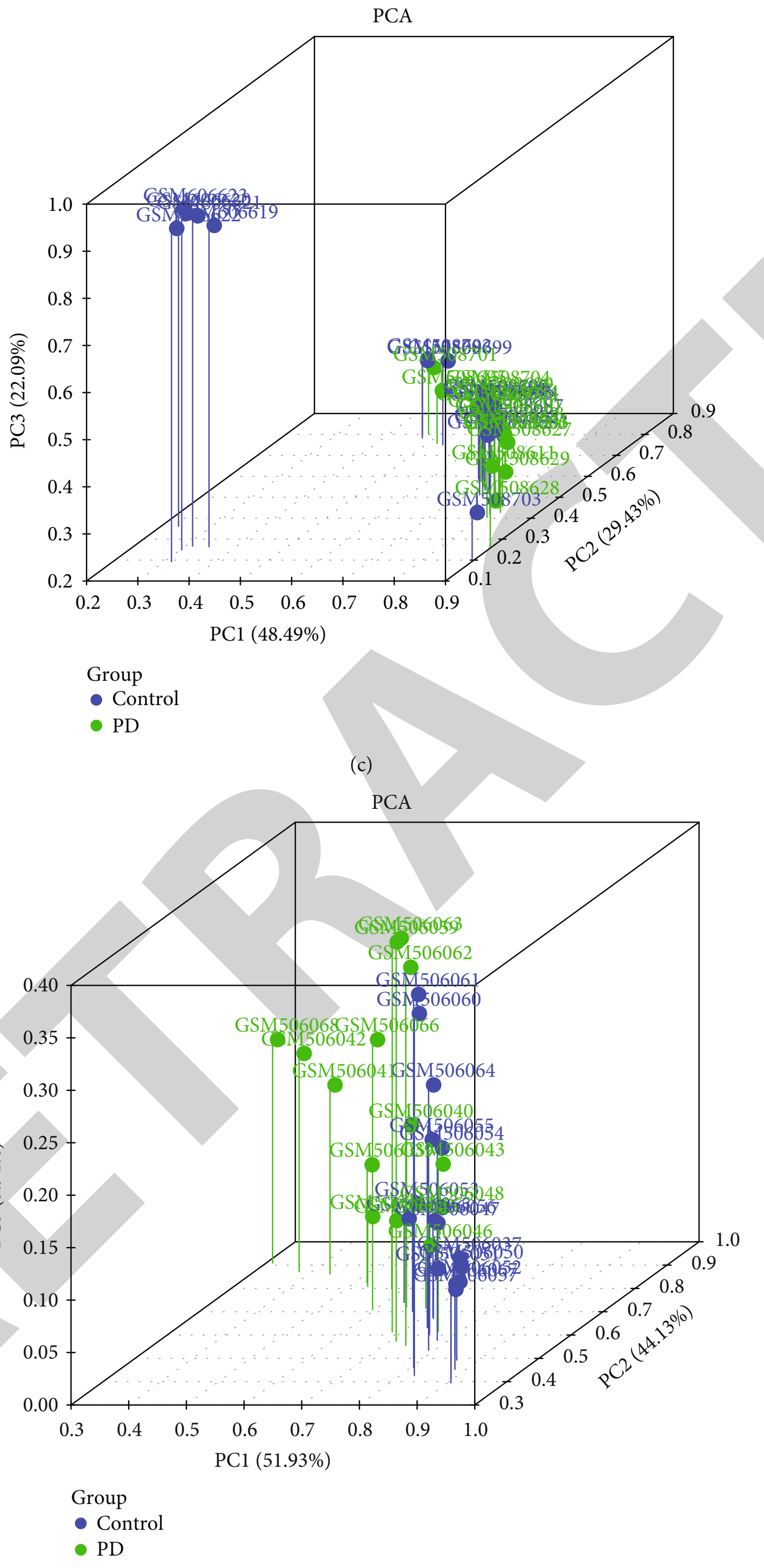
2.4. Protein-Protein Interaction (PPI) Network. PPI of the target gene list was analyzed using the STRING (https://string-db.org/) online database [16]. The confidence of protein interaction was set as combined score > 0.4. Then, the Cytoscape software was utilized to visualize the PPI network [17]. By the cytoHubba plug-in [16], the degree of connectivity



(a)

(b)

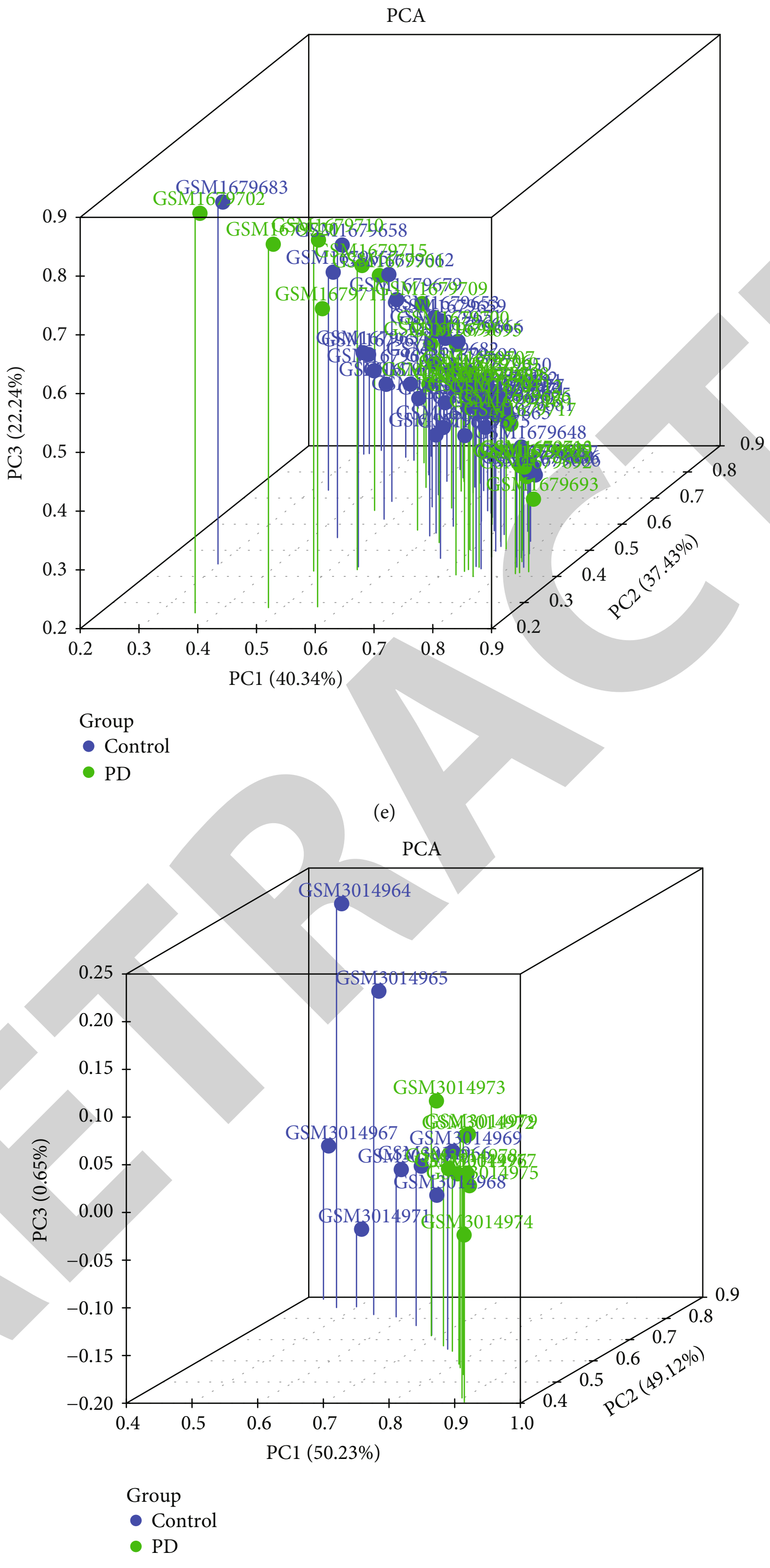
Figure 1: Continued.



(c)

(d)

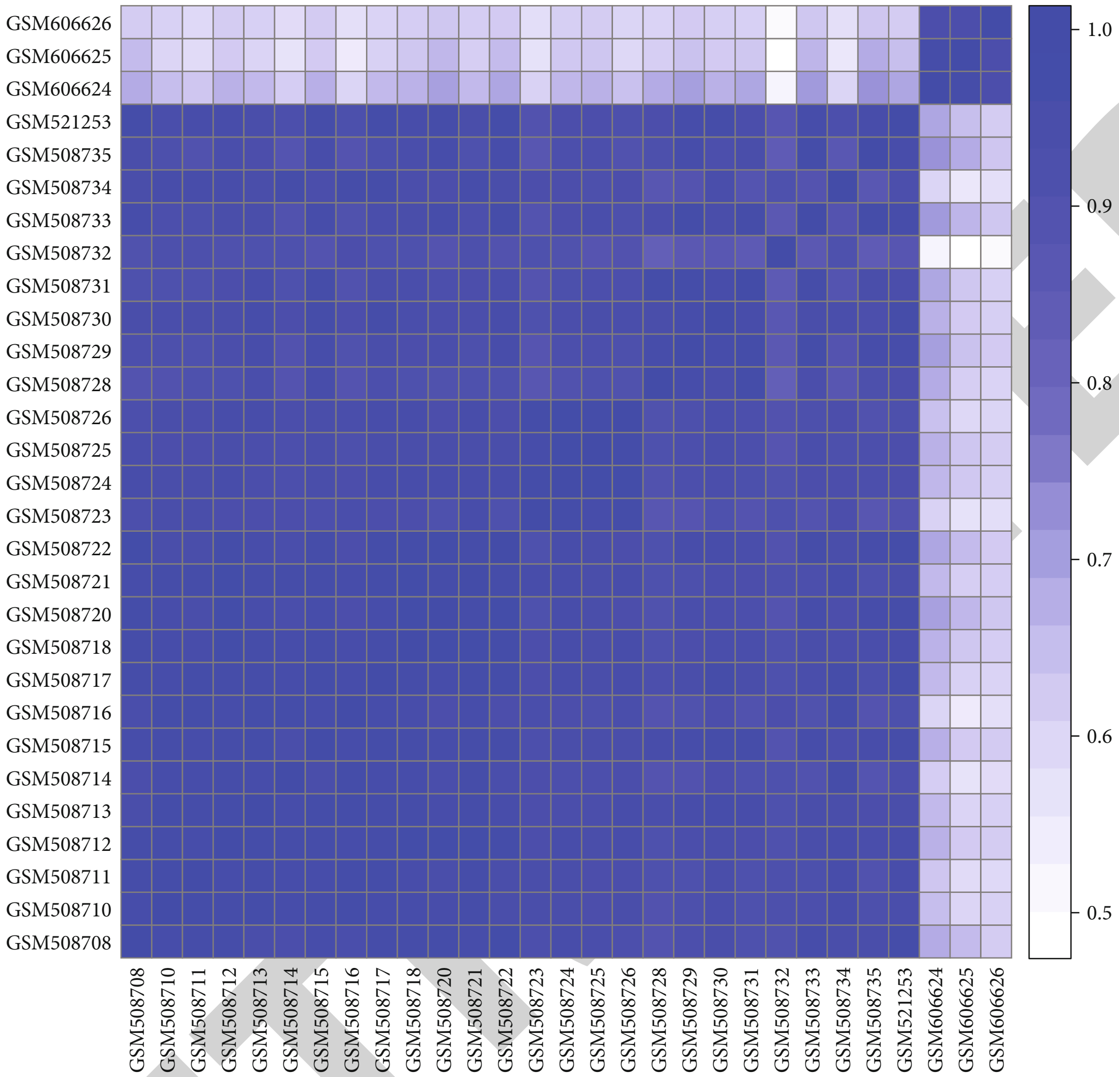
Figure 1: Continued.



(e)

(f)

Figure 1: Principal component analysis for multiple brain regions of PD samples and normal samples: (a, b) substantia nigra (GSE20292 and GSE7621); (c) putamen (GSE20291); (d, e) prefrontal cortex area (GSE20168 and GSE68719); (f) cingulate gyrus (GSE110716). There are three principal components (PC1, PC2, and PC3). Blue indicates control normal samples, and green indicates PD samples.



(a)

Figure 2: Continued.

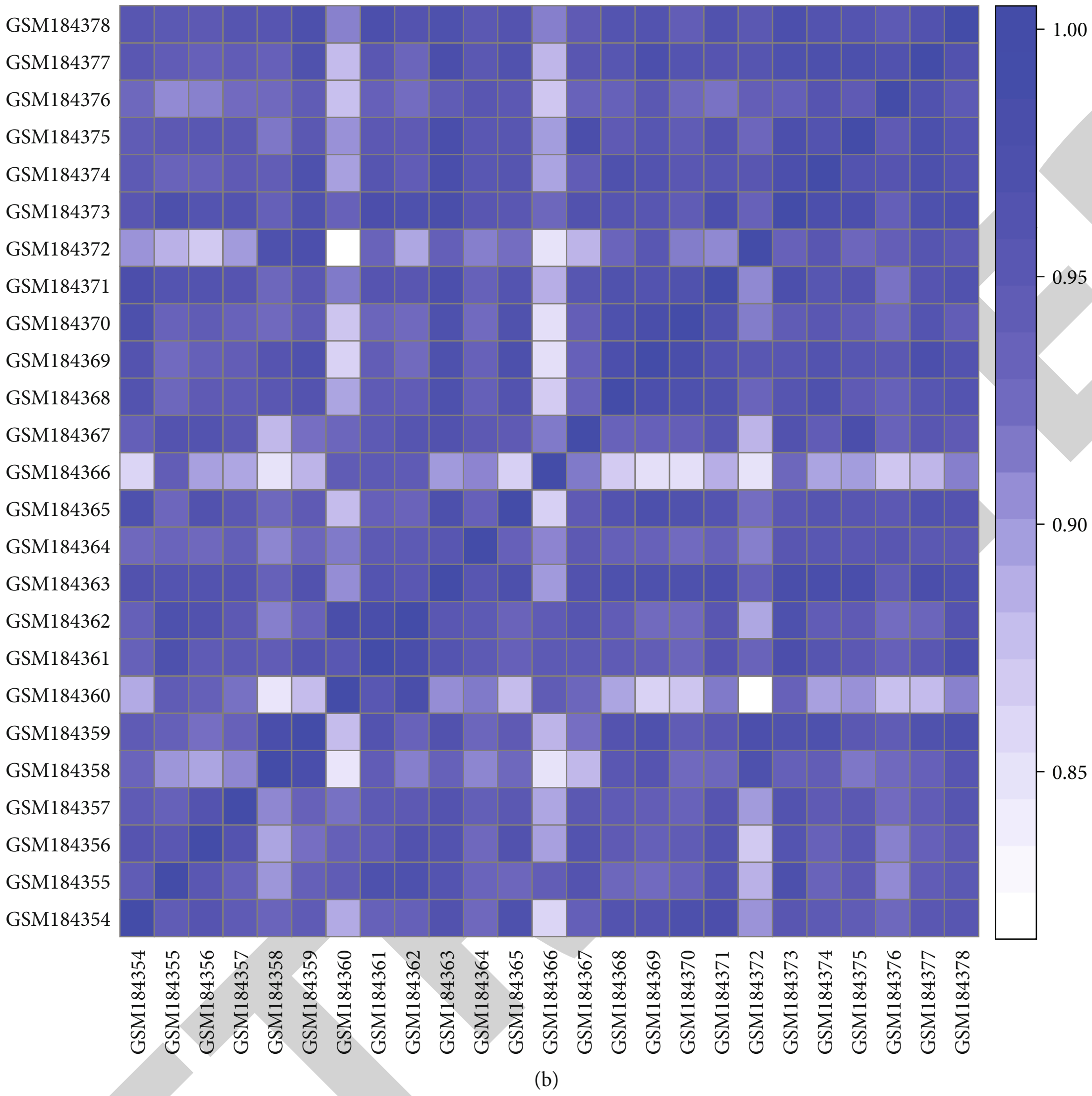
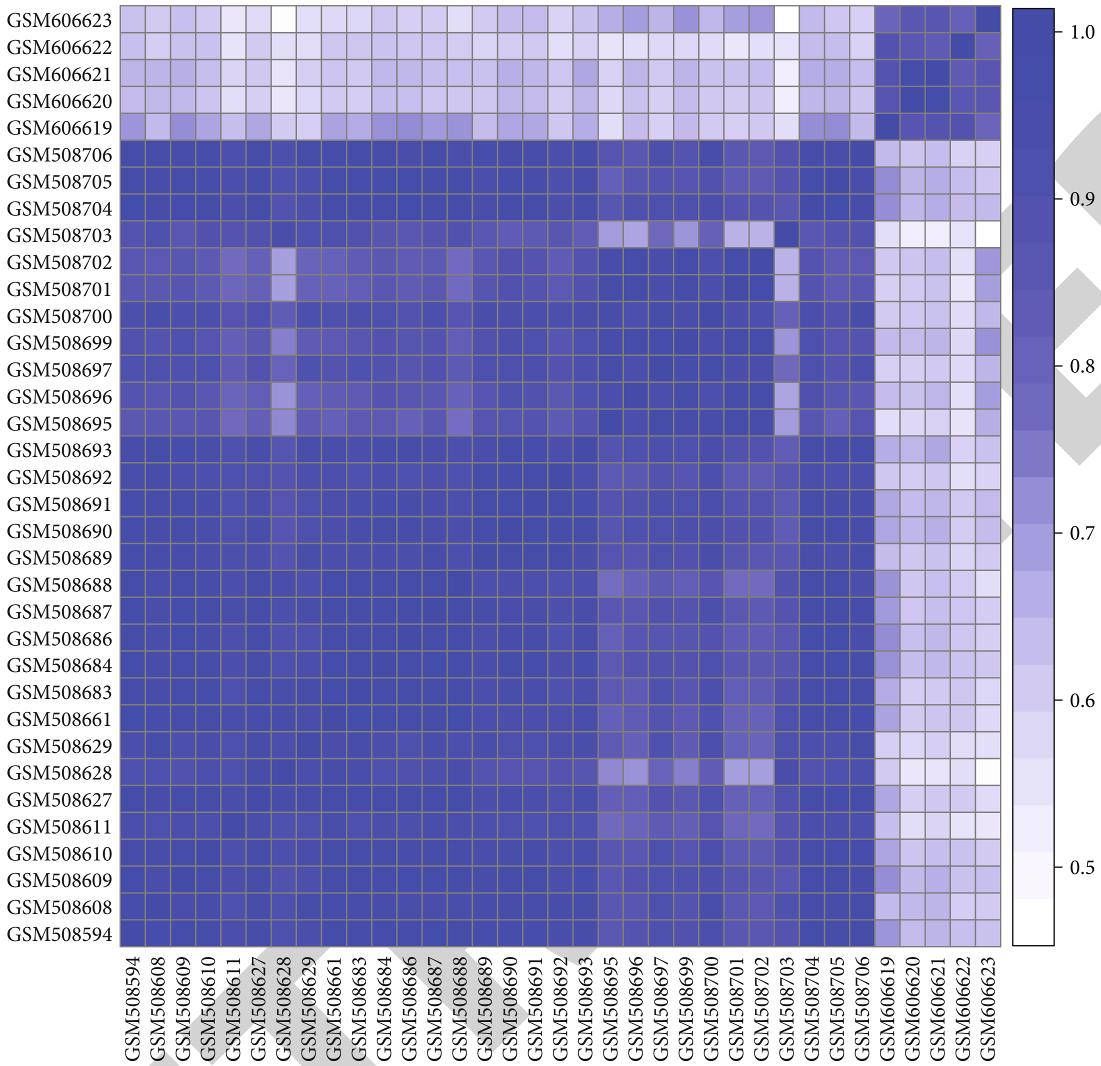
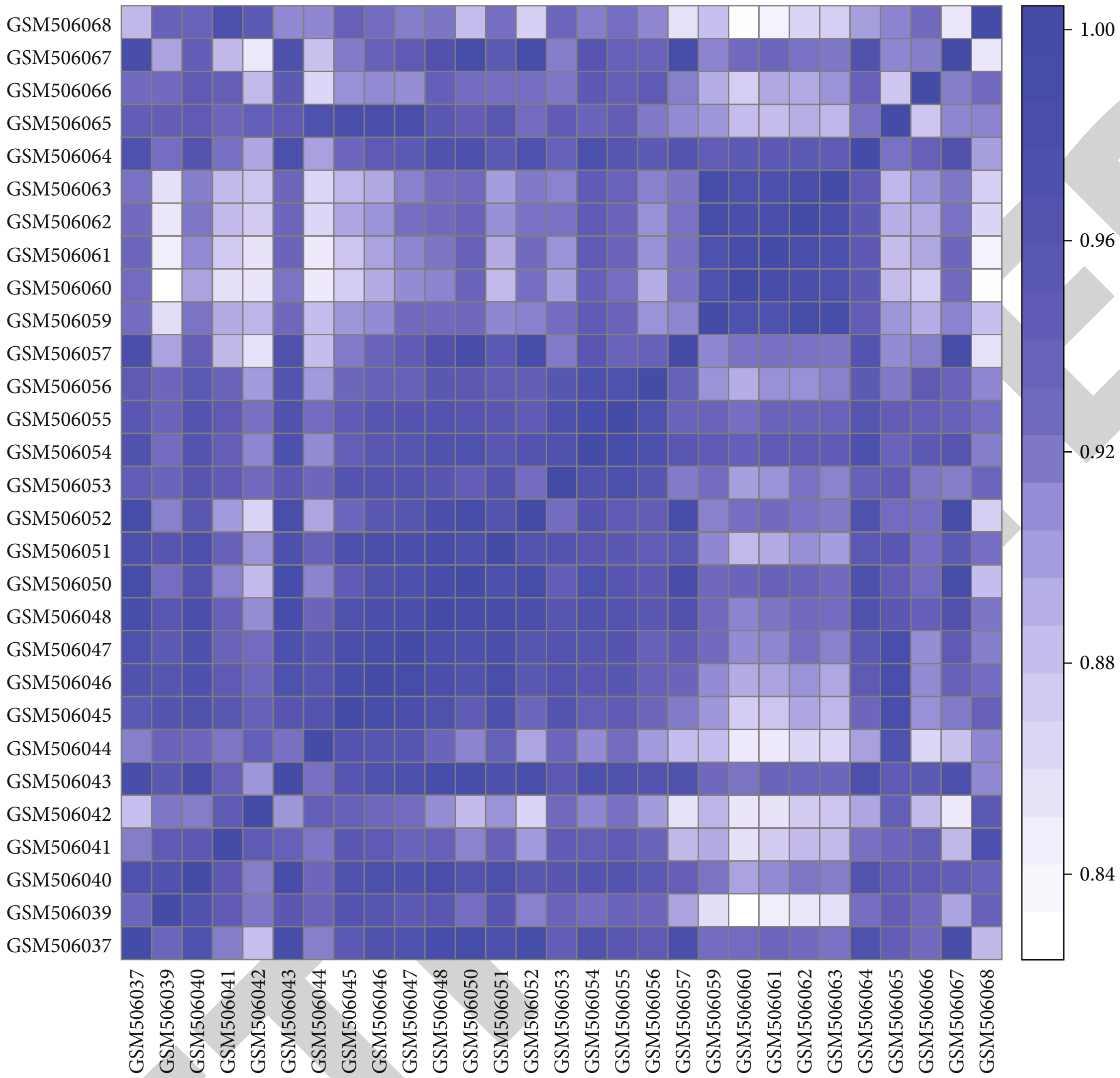


Figure 2: Continued.



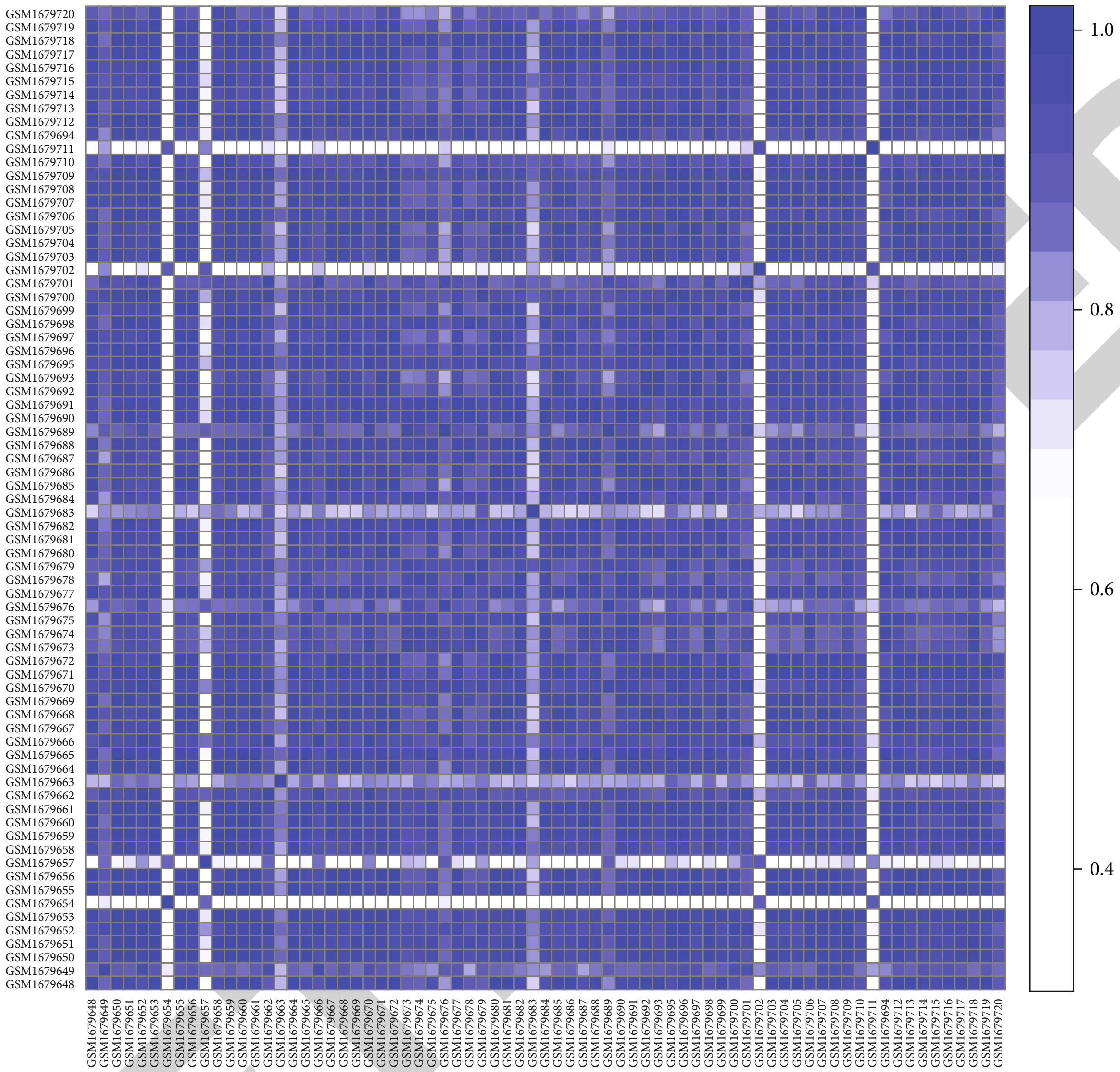
(c)

Figure 2: Continued.



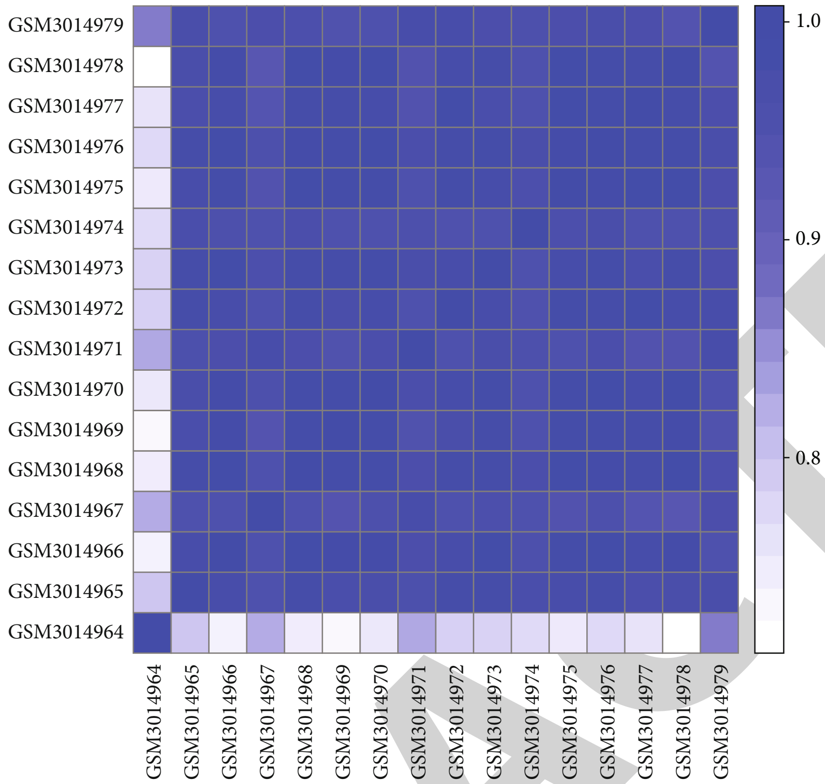
(d)

Figure 2: Continued.



(e)

Figure 2: Continued.



(f)

Figure 2: Heat maps depicting sample correlation analysis between PD and normal samples: (a, b) substantia nigra (GSE20292 and GSE7621); (c) putamen (GSE20291); (d, e) prefrontal cortex area (GSE20168 and GSE68719); (f) cingulate gyrus (GSE110716). The correlation coefficient indicates the similarity between samples. The closer the correlation coefficient is to 1, the higher the similarity between the two samples.

of the node was calculated and hub genes in the PPI network were determined [18].

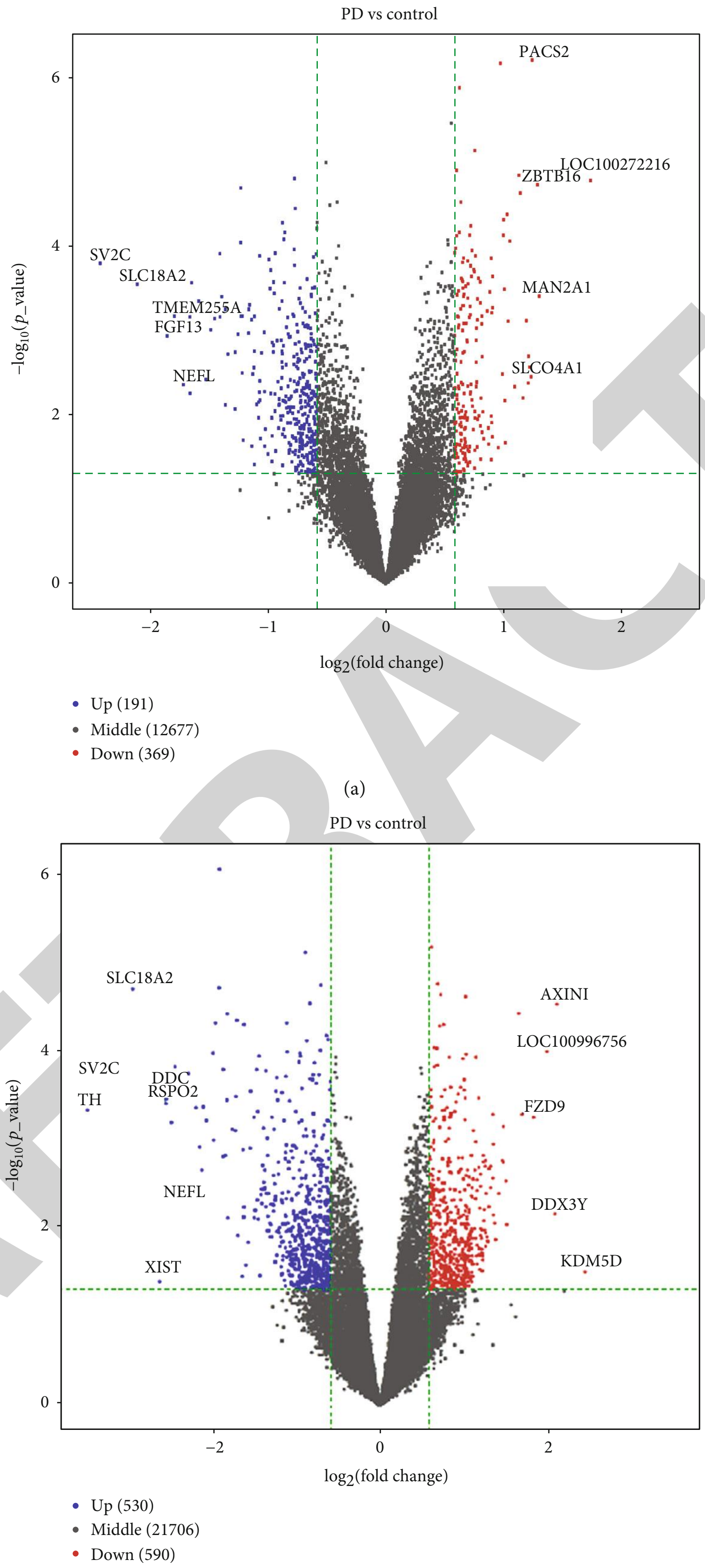
2.5. Functional Enrichment Analysis. Gene Ontology (GO) enrichment analysis including biological process, cellular component, and molecular function was carried out through the Gene Ontology database [19]. Moreover, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis was also presented by KEGG PATHWAY DATABASE [20]. Fisher's exact test was used to find out which items or pathways were significantly related to a set of genes. p value < 0.05 indicated significant enrichment.

2.6. Gene Set Enrichment Analyses (GSEA). The clusterProfiler package [21] was used to perform GSEA on the transcriptional data of multiple brain regions in PD from the GSE20295 dataset [17, 22]. Using the gene set file wikipathways-20180810-gmt-Homo_sapiens.gmt from the clusterProfiler package, GSEA was presented based on the default parameters.

3. Results

3.1. Principal Component Analysis for Multiple Brain Regions of PD Samples. In this study, we obtained expression profiles from multiple brain regions of PD, including the substantia nigra (GSE20292 and GSE7621), putamen (GSE20291), prefrontal cortex area (GSE20168 and GSE68719), and cingulate gyrus (GSE110716). Before downstream analysis, all samples were assessed by PCA. The results showed that PD substantia nigra samples (Figures 1(a) and 1(b)), putamen (Figure 1(c)), prefrontal cortex area (Figures 1(d) and 1(e)), and cingulate gyrus (Figure 1(f)) were distinctly distinguished from normal samples. Furthermore, based on these gene expression data, we calculated the correlation coefficients between the two samples. There was a significant high correlation between different samples for PD substantia nigra samples (Figures 2(a) and 2(b)), putamen (Figure 2(c)), prefrontal cortex area (Figures 2(d) and 2(e)), cingulate gyrus (Figure 2(f)), and corresponding normal samples.

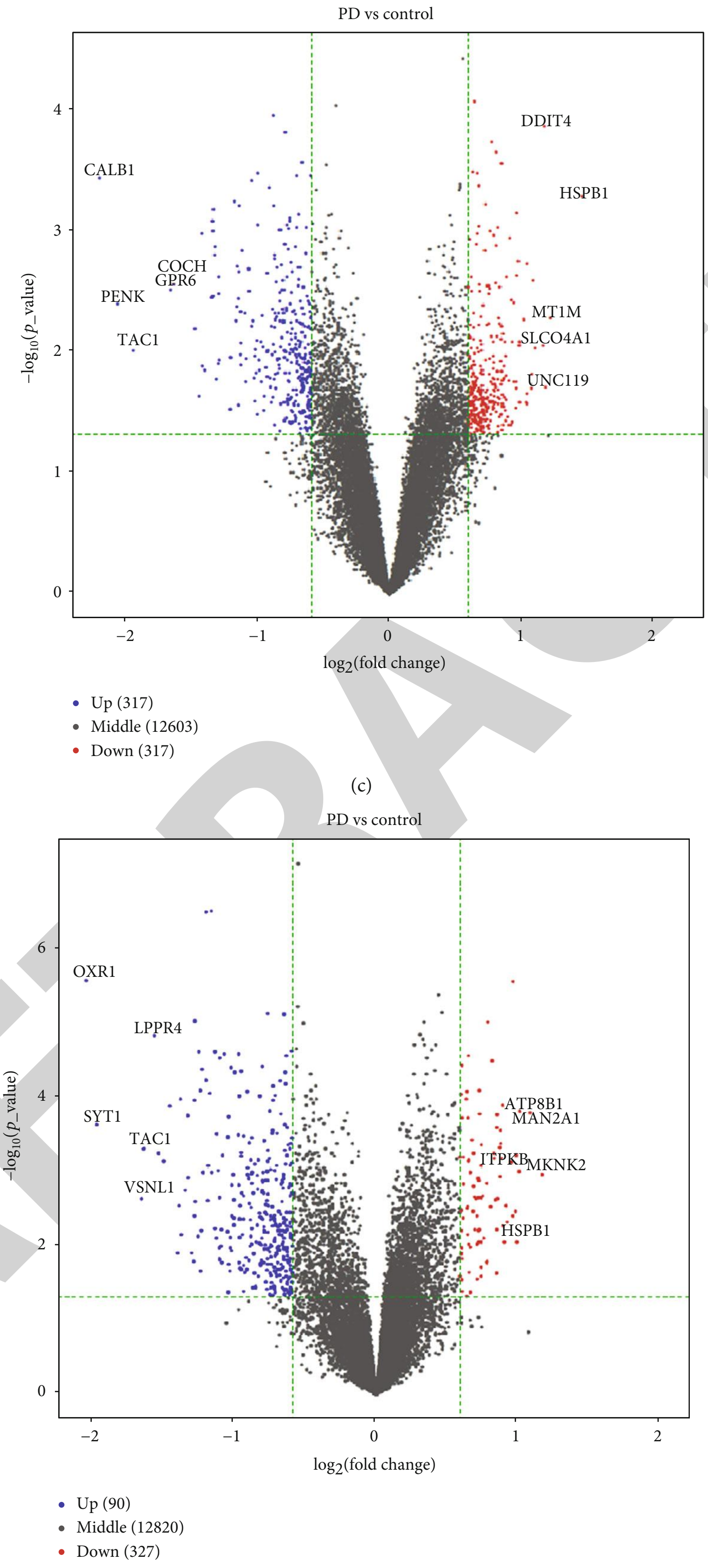
3.2. Differentially Expressed Genes for Multiple Brain Regions of PD. With the screening criteria of p value < 0.05 and $|\log_2 FC| > 0.585$, DEGs between multiple brain regions of PD samples and normal samples were identified. In the GSE20292 dataset, there were 191 upregulated and 369 downregulated genes between the substantia nigra of PD and normal samples (Figure 3(a)). 530 upregulated and 590 downregulated genes were screened for the substantia nigra of PD samples compared to normal samples in the



(a)

(b)

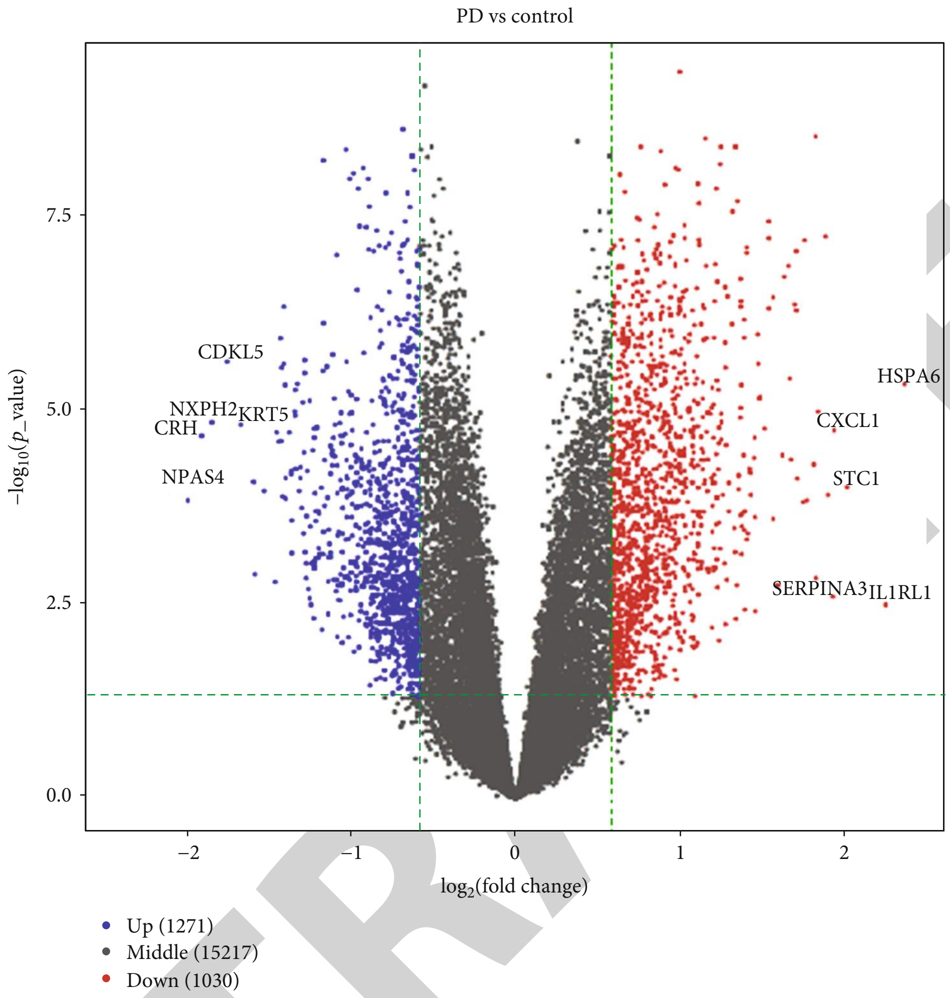
Figure 3: Continued.



(c)

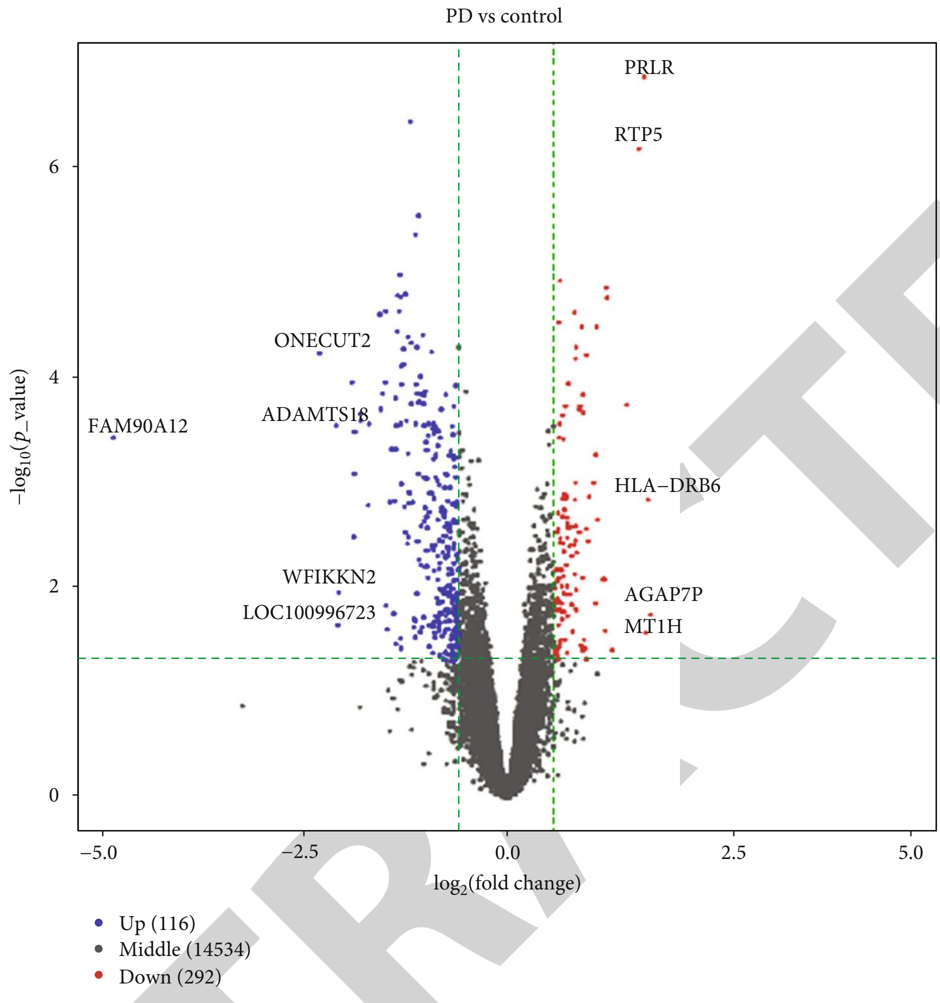
(d)

Figure 3: Continued.



(e)

Figure 3: Continued.



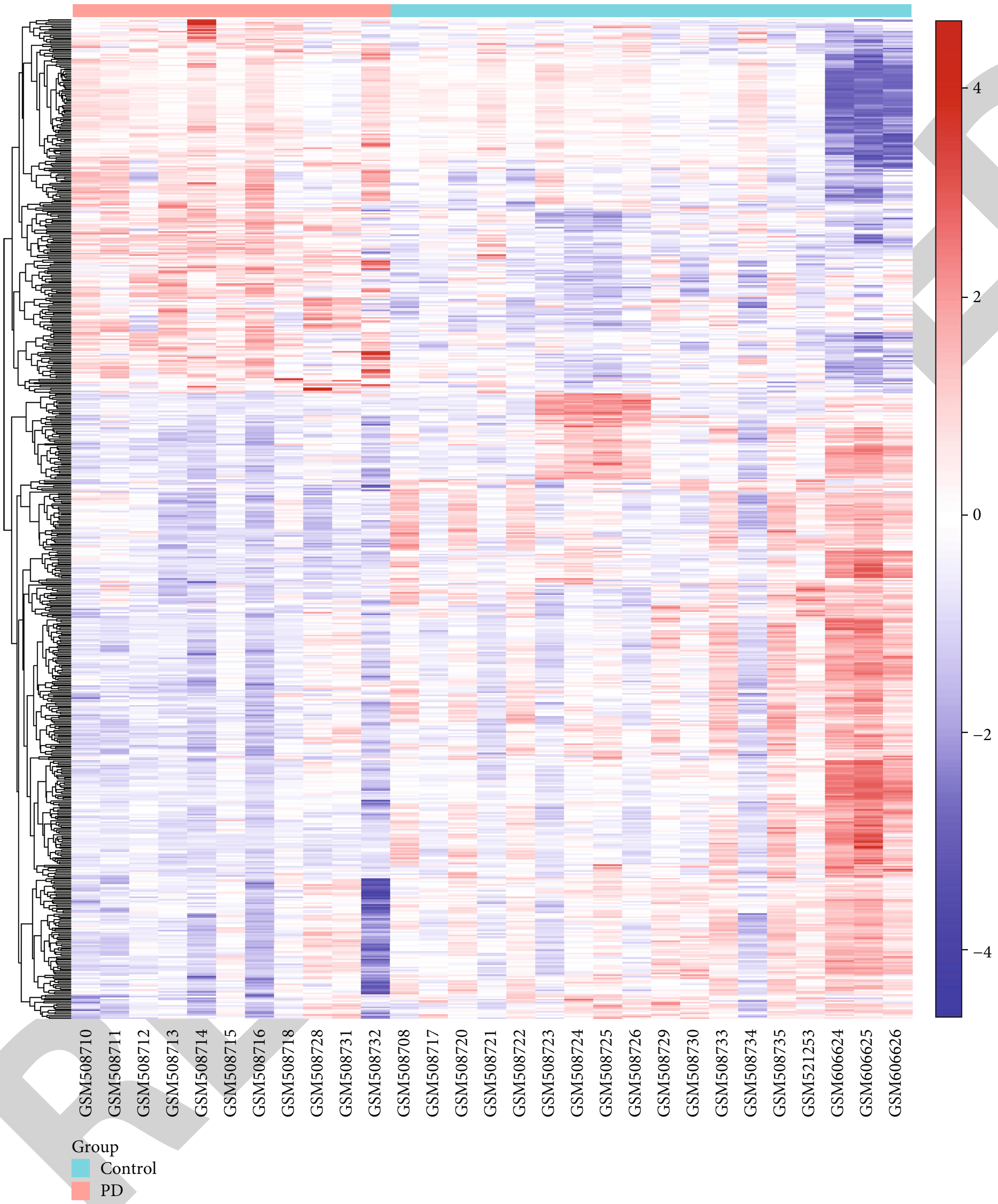
(f)

Figure 3: Volcano plots depicting differentially expressed genes between multiple brain regions of PD and corresponding normal tissues: (a, b) substantia nigra (GSE20292 and GSE7621); (c) putamen (GSE20291); (d, e) prefrontal cortex area (GSE20168 and GSE68719); (f) cingulate gyrus (GSE110716). Red suggests upregulation, and blue suggests downregulation. The top five most significant upregulated genes and downregulated genes were labeled separately.

GSE7621 dataset (Figure 3(b)). After overlapping the results from the two datasets, 17 upregulated and 52 downregulated genes were identified for the substantia nigra of PD. In the GSE20291 dataset, 317 upregulated and 317 downregulated genes were identified between PD putamen and normal tissues (Figure 3(c)). Following the intersections of DEGs from the GSE20168 dataset (90 upregulated and 327 downregulated genes; Figure 3(d)) and the GSE68719 dataset (1271 upregulated and 1030 downregulated genes; Figure 3(e)), 39 upregulated and 72 downregulated genes were identified for PD prefrontal cortex area tissues in comparison to normal tissues. In the GSE110716 dataset, there were 116 upregulated and 292 downregulated genes between PD cingulate gyrus and normal tissues (Figure 3(f)). Heat maps depicted that these DEGs could significantly distinguish PD substantia nigra samples (Figures 4(a) and 4(b)), putamen (Figure 4(c)), prefrontal cortex area (Figures 4(d) and 4(e)), and cingulate gyrus (Figure 4(f)) from the corresponding normal samples.

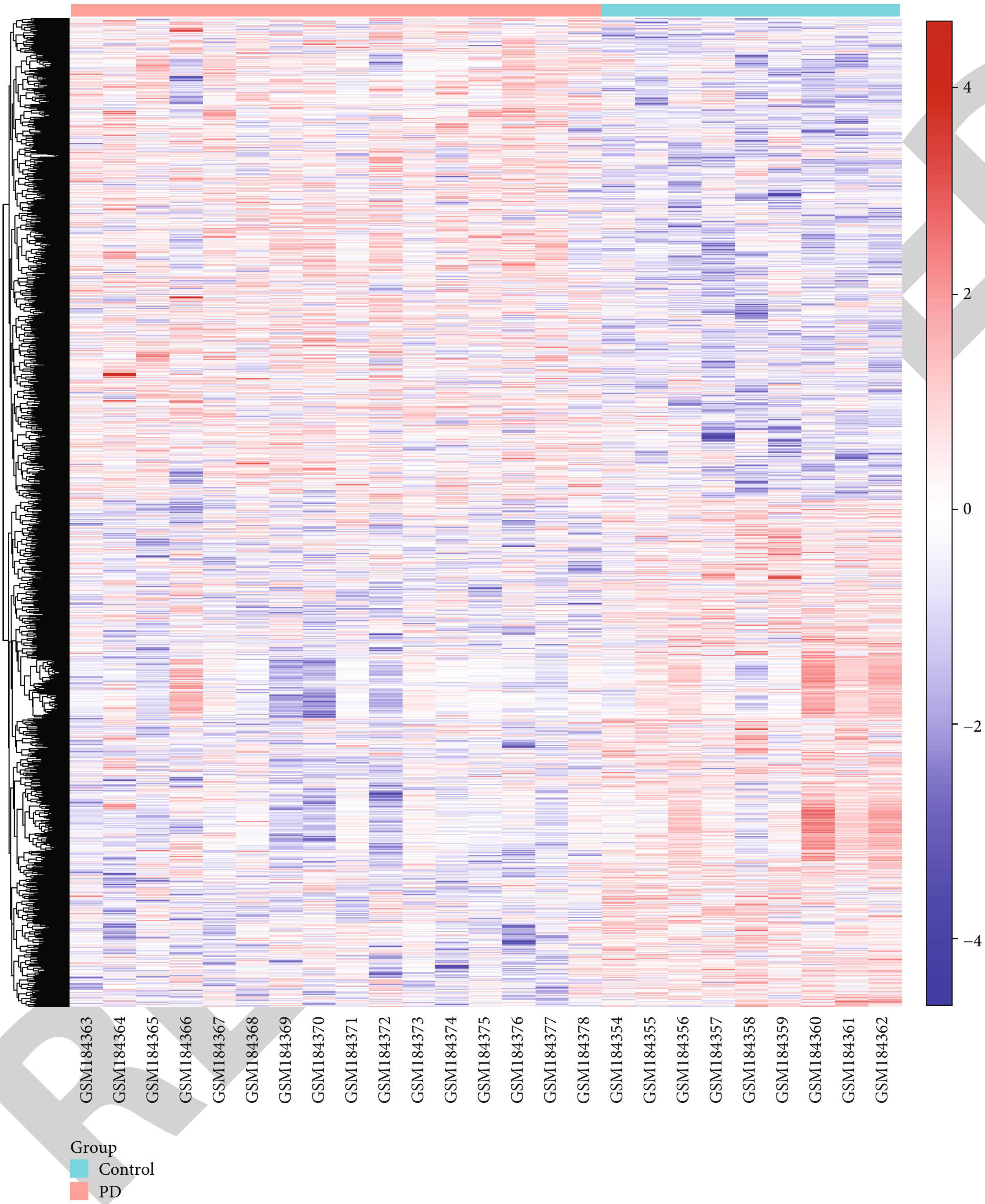
3.3. *Construction of WGCNA for the Substantia Nigra of PD.*

11 substantia nigra PD and 18 control samples were used for coexpression analysis in the GSE20292 dataset. Coexpression module analysis was easily affected by outlier samples, so removing outlier sample data before constructing the network was especially important for obtaining meaningful analysis results. Herein, there were no outliers among them (Figure 5(a)). Thus, no samples were removed. By dynamic cutting tree method, gene modules were divided, and highly similar modules were merged (Figure 5(b)). Finally, nine modules were constructed. Among them, the purple module was significantly related to PD ($r = -0.44$ and $p = 0.02$) (Figure 5(c)). Heat maps showed that there was a high correlation between different genes (Figure 5(d)). We further assessed the coexpression similarity of modules. These modules were divided into two main clusters, which were validated by adjacency heat maps (Figure 5(e)).



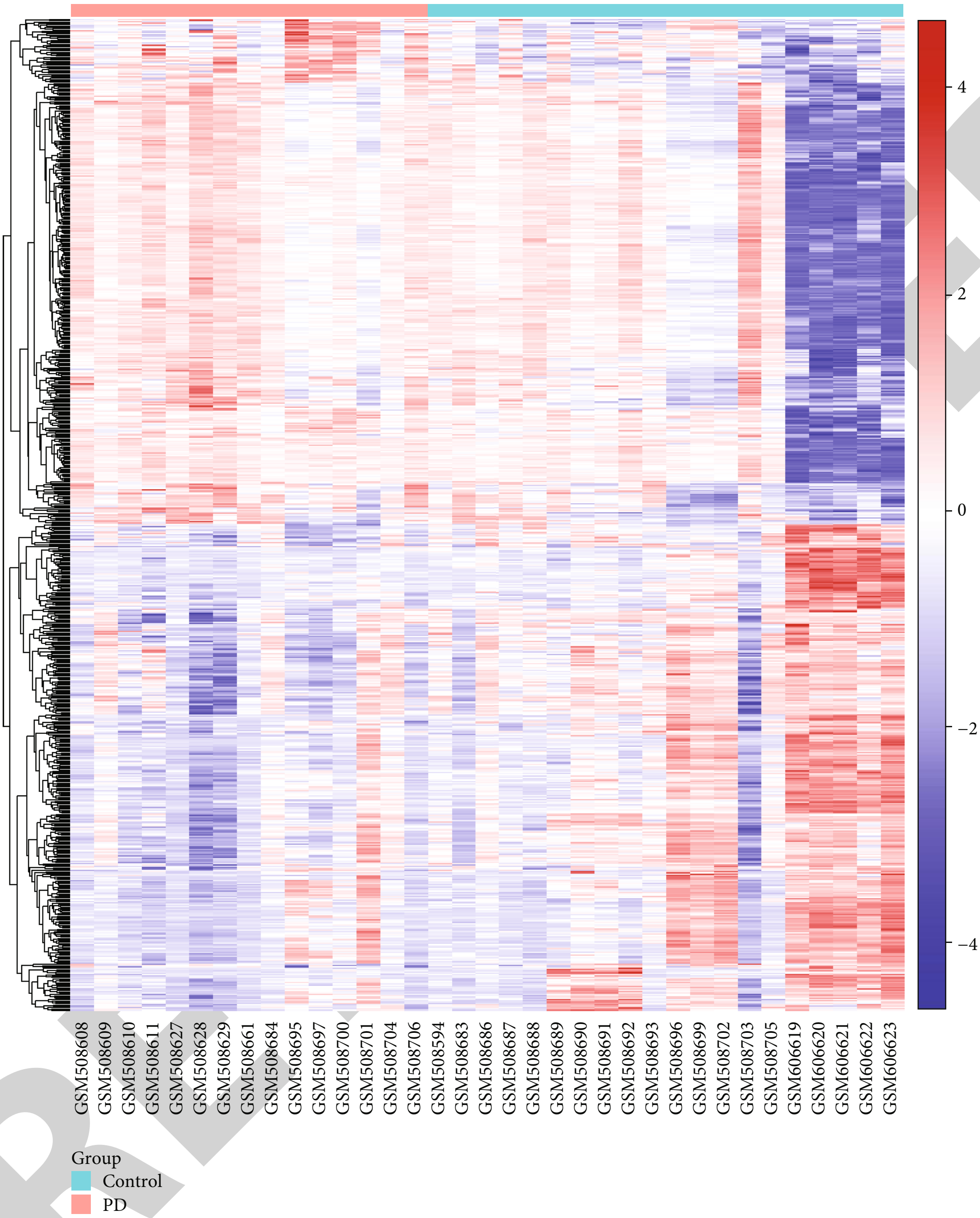
(a)

Figure 4: Continued.



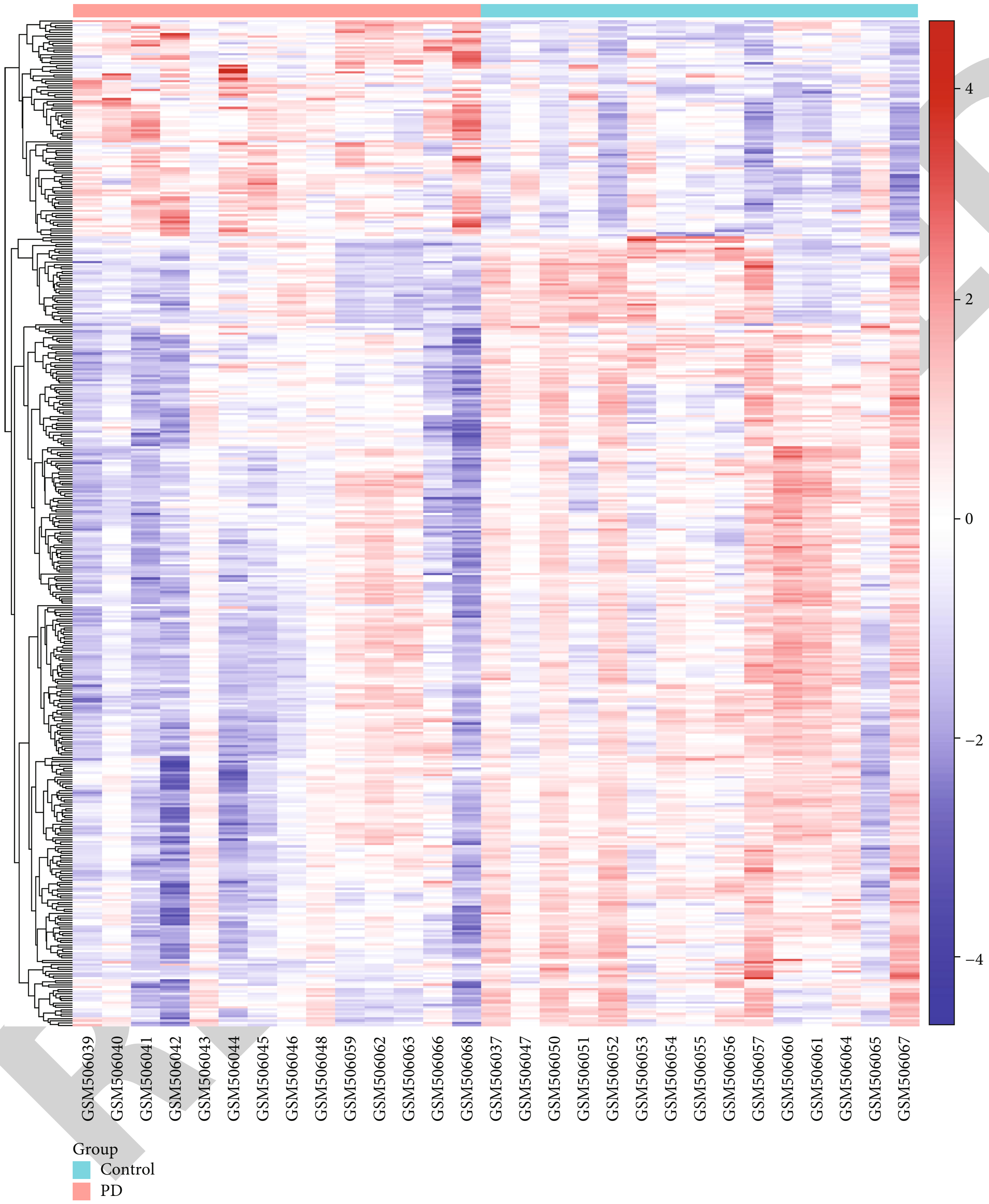
(b)

Figure 4: Continued.



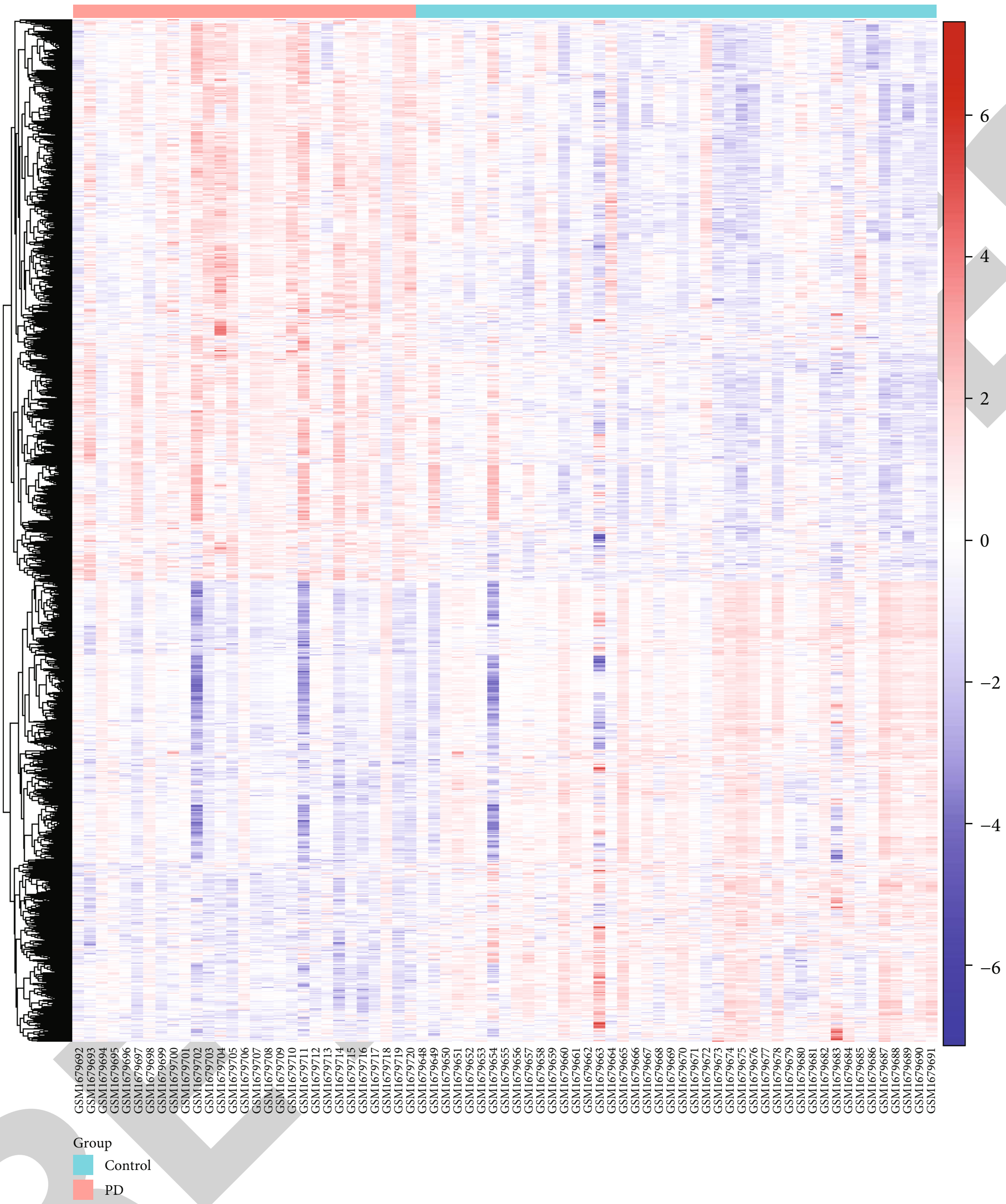
(c)

Figure 4: Continued.



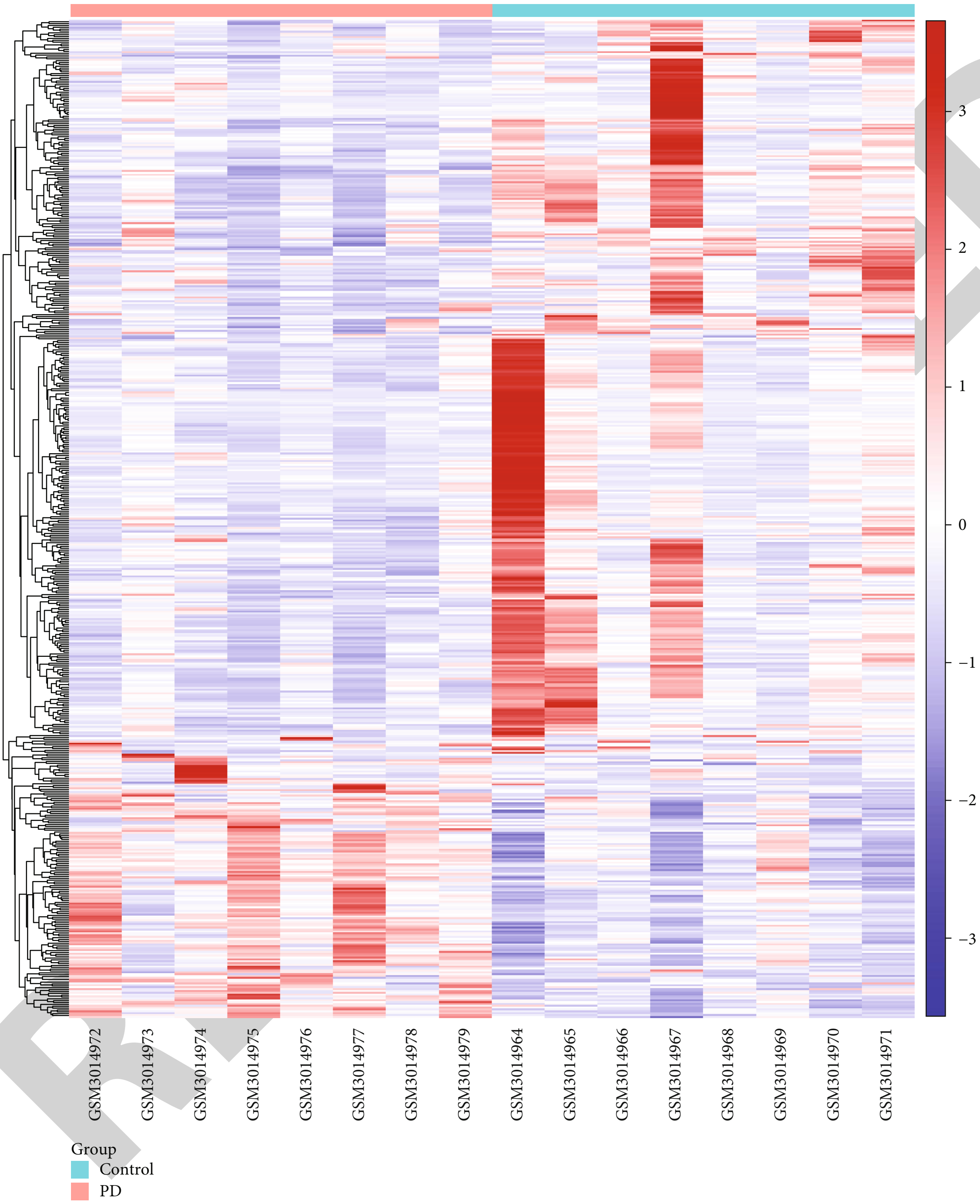
(d)

Figure 4: Continued.



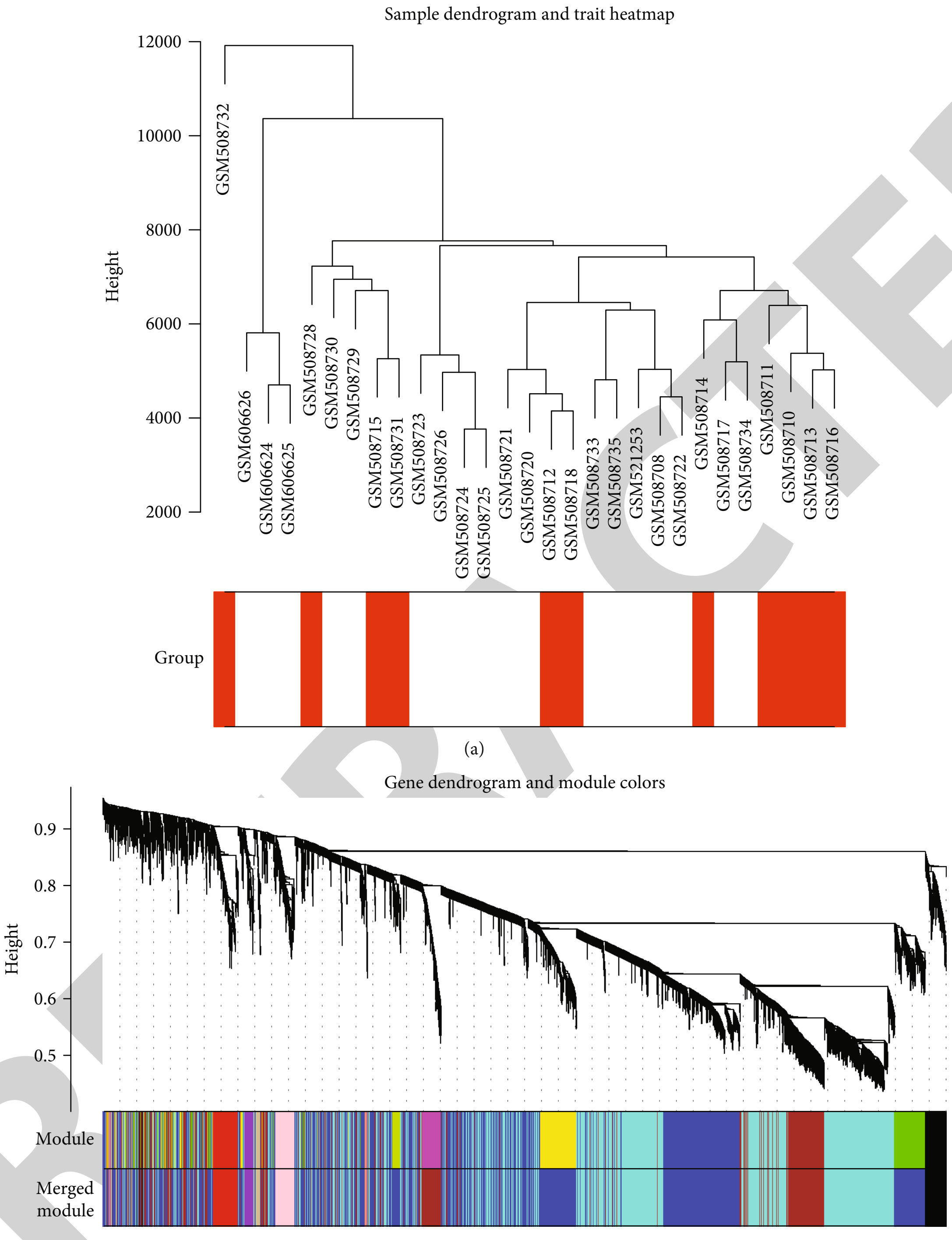
(e)

Figure 4: Continued.



(f)

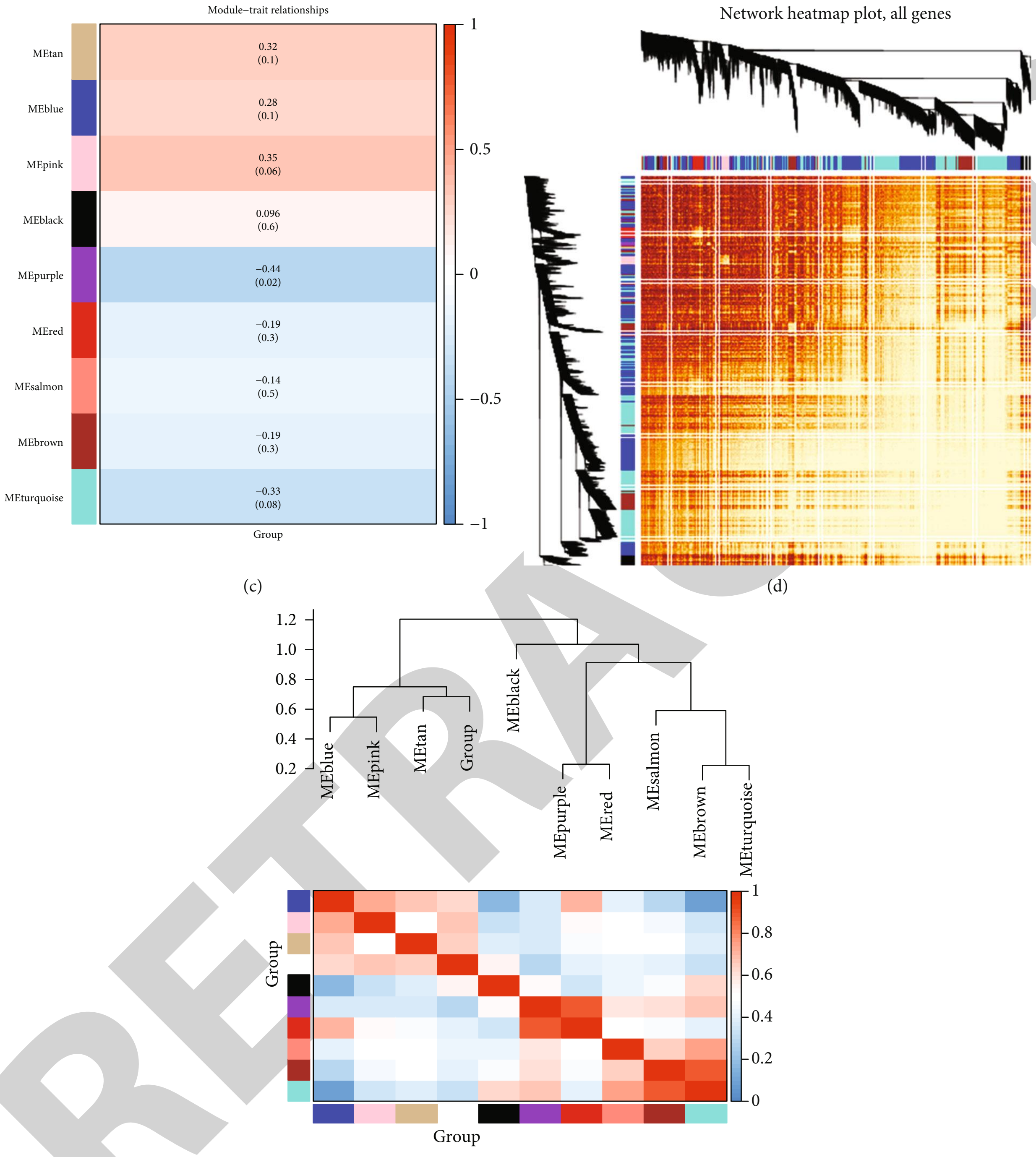
Figure 4: Heat maps showing expression patterns of differentially expressed genes between PD and corresponding normal tissues: (a, b) substantia nigra (GSE20292 and GSE7621); (c) putamen (GSE20291); (d, e) prefrontal cortex area (GSE20168 and GSE68719); (f) cingulate gyrus (GSE110716). Red suggests upregulation, and blue suggests downregulation.



(a)

(b)

Figure 5: Continued.



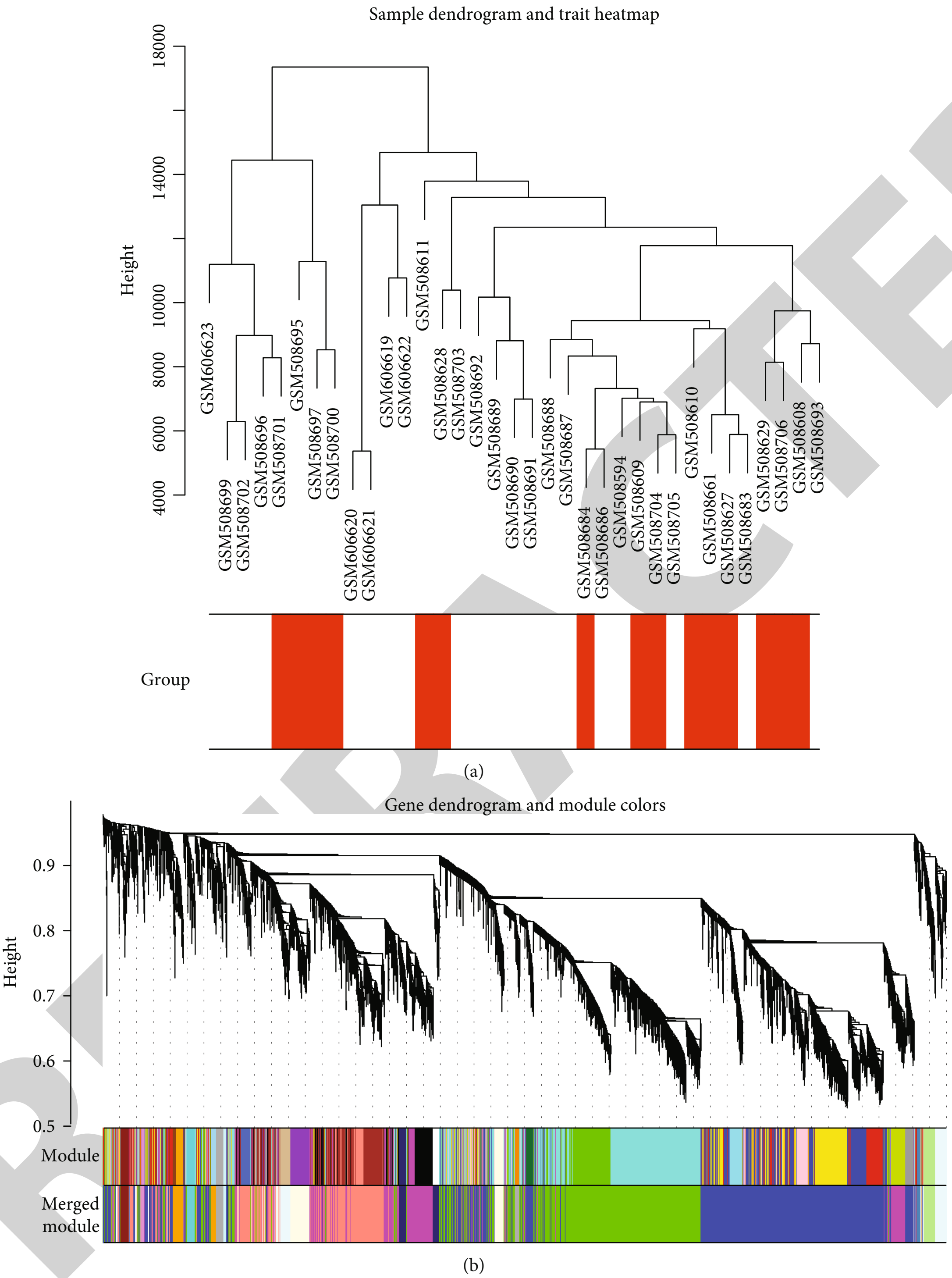
(c)

(d)

(e)

Figure 5: Construction of WGCNA for the substantia nigra of PD. (a) Sample hierarchical clustering tree to detect outliers. (b) Dynamic cutting tree method was utilized to determine gene modules. (c) Module-trait relationship network. The color of the square indicates the correlation between the module and the clinical traits. The p value is in brackets. (d) Hierarchical clustering dendrogram. The branches correspond to each module. The module memberships colored by different colors are shown in the color bar below and to the right of the tree diagram. Shades of color are proportional to coexpression interconnectedness. (e) Clustering of module eigengenes and eigengene adjacency heat map. Red represents high correlation and blue represents low correlation.

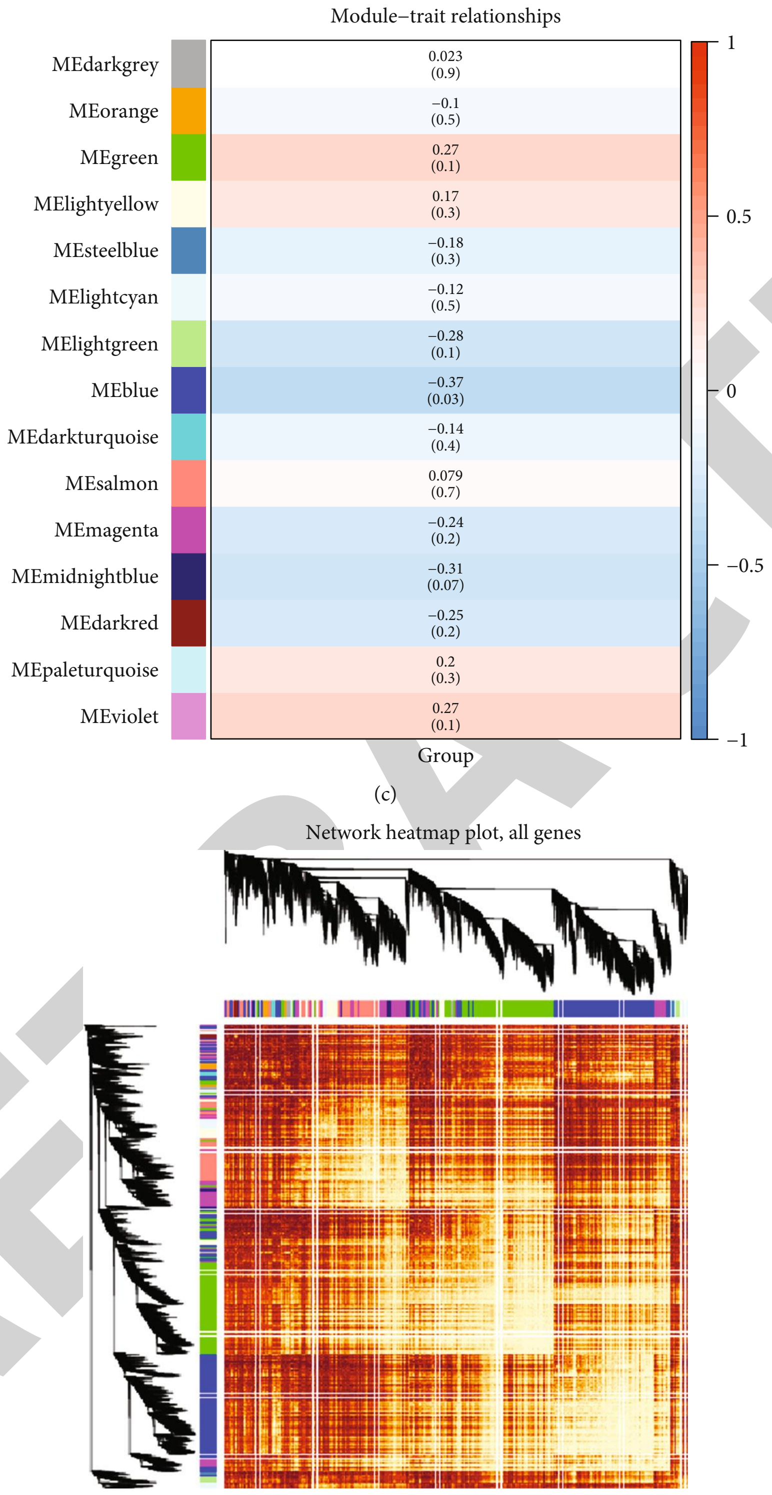
3.4. Construction of WGCNA for the Substantia Nigra of PD. In the GSE20291 dataset, 15 putamen PD and 20 normal samples were utilized for WGCNA. No outliers were detected and removed among them (Figure 6(a)). Using dynamic cutting tree method, gene modules were built (Figure 6(b)). Finally, fifteen coexpression modules with high similarity were merged. Among them, the blue module was distinctly correlated to PD ($r = -0.37$ and $p = 0.03$) (Figure 6(c)). Heat maps showed that there was a high correlation between different genes in different modules (Figure 6(d)). The



(a)

(b)

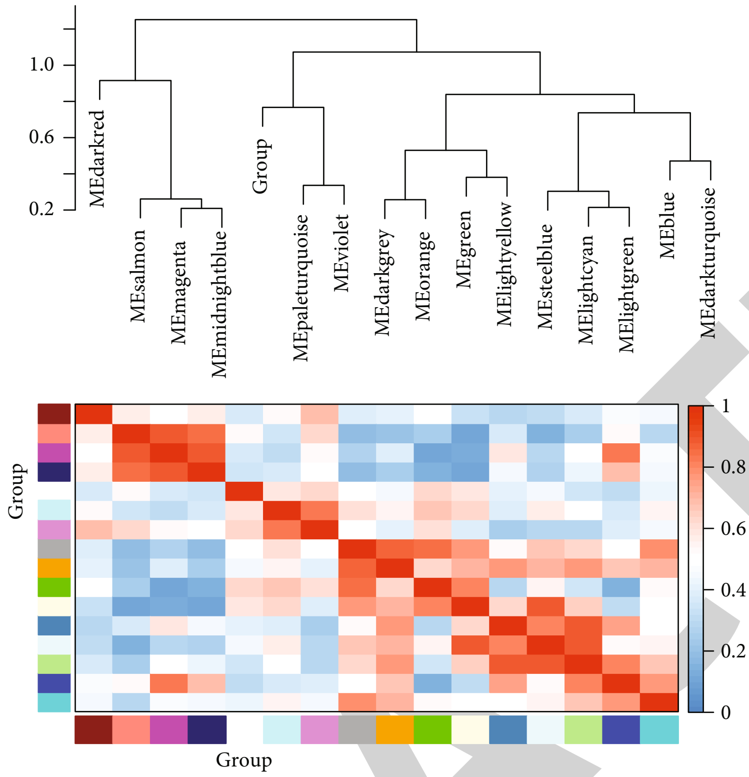
Figure 6: Continued.



(c)

(d)

Figure 6: Continued.



(e)

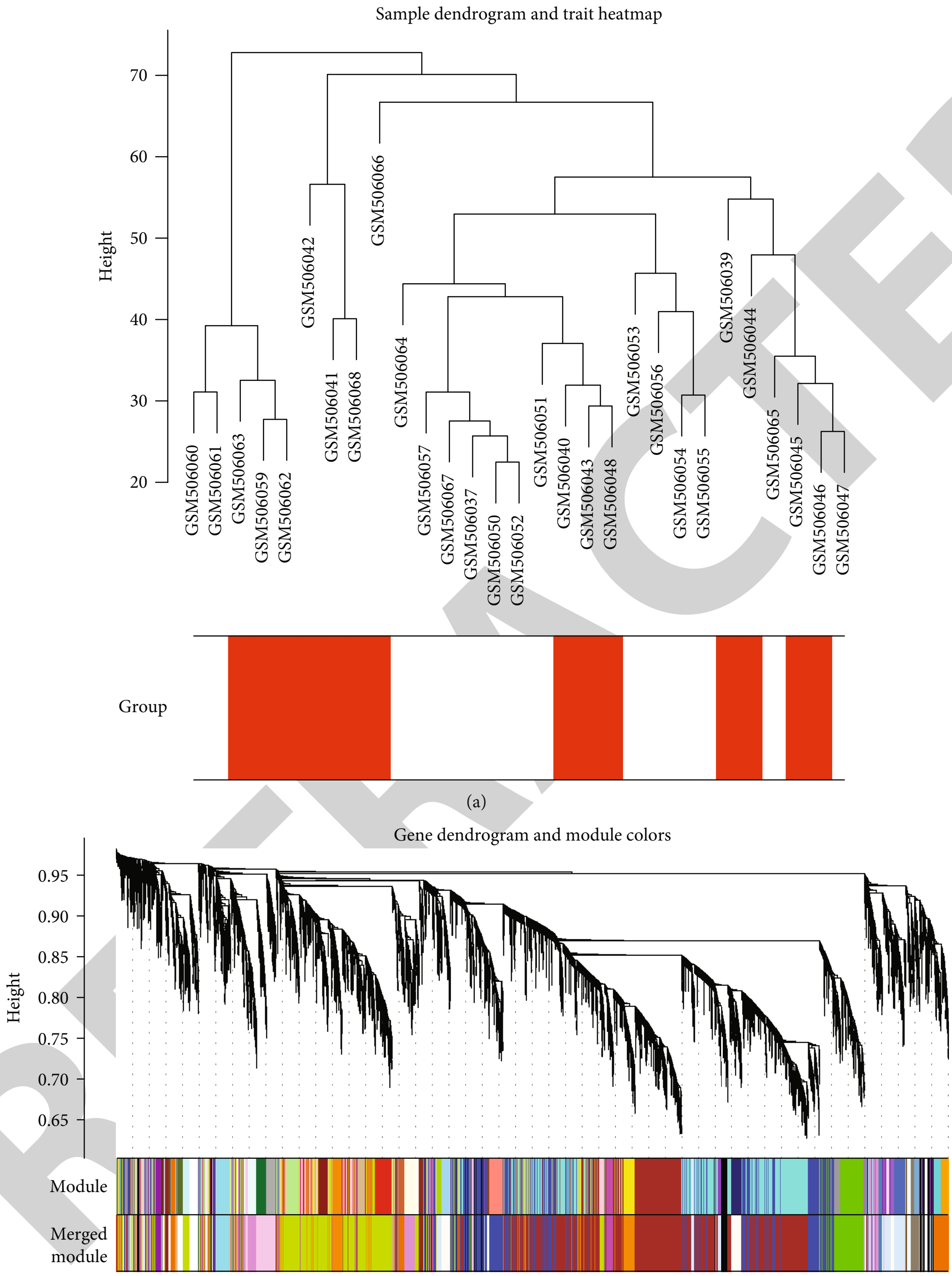
Figure 6: Construction of WGCNA for the putamen of PD. (a) Sample hierarchical clustering tree to detect outliers. (b) Gene modules were determined by dynamic cutting tree method. (c) Module-trait relationship network. (d) Hierarchical clustering dendrogram. (e) Clustering of module eigengenes and eigengene adjacency heat map.

coexpression similarity of modules was analyzed, as shown in Figure 6(e).

3.5. Construction of WGCNA for the Prefrontal Cortex of PD. 14 prefrontal cortex PD and 15 normal samples were analyzed by WGCNA in the GSE20168 dataset. There was no outlier sample among them (Figure 7(a)). Gene modules were divided using dynamic cutting tree method (Figure 7(b)). After merging, 25 modules were constructed. Among them, the green ($r = -0.43$ and $p = 0.02$), magenta ($r = -0.52$ and $p = 0.004$), and bisque ($r = -0.54$ and $p = 0.002$) modules were negatively correlated to PD (Figure 7(c)). Also, salmon ($r = 0.42$ and $p = 0.02$) and dark orange ($r = 0.49$ and $p = 0.006$) modules were positively related to PD. According to the network heat map plot, each module was independent of others (Figure 7(d)). Furthermore, their coexpression similarity was quantified by adjacency heat maps (Figure 7(e)).

3.6. Construction of WGCNA for the Cingulate Gyrus of PD. From the GSE110716 dataset, 8 cingulate gyrus PD and 8 control samples were obtained for WGCNA. No outlier samples were detected among them (Figure 8(a)). Gene modules were divided via dynamic cutting tree method (Figure 8(b)). Following merging, 40 coexpression modules were constructed. Among them, the orange red ($r = 0.65$ and $p = 0.006$) and thistle ($r = 0.51$ and $p = 0.04$) modules had positive correlations to PD. The medium purple ($r = -0.65$ and $p = 0.006$) and salmon ($r = -0.55$ and $p = 0.03$) modules had negative correlations to PD in Figure 8(c). In the network heat map plot, each module was independent of others (Figure 8(d)). Moreover, their coexpression similarity was evaluated by adjacency heat maps (Figure 8(e)).

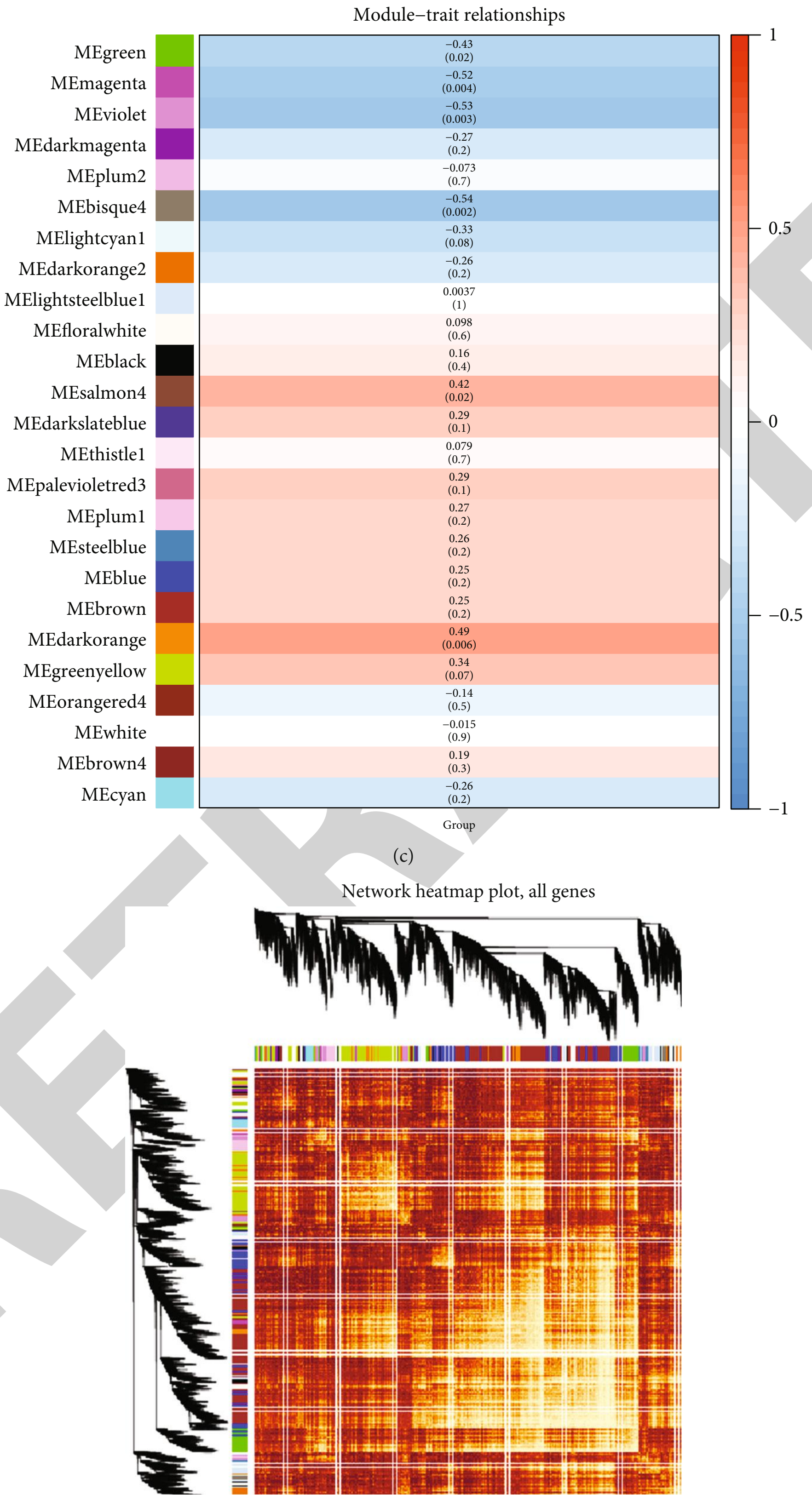
3.7. PPI Networks for DEGs in Multiple Brain Regions of PD. DEGs for the substantia nigra, putamen, prefrontal cortex area, and cingulate gyrus of PD were extracted for PPI networks by the STRING database. There were 69 nodes in the PPI network of substantia nigra PD, including 17 upregulated and 52 downregulated genes (Figure 9(a)). Among them, SLC6A3 (degree = 6), SLC18A2 (degree = 6), and TH (degree = 6) had the highest degree, which were considered as hub genes. In Figure 9(b), there were 317 upregulated genes and 317 downregulated genes in the PPI network of putamen. Among them, BMP4 (degree = 14) and SNAP25 (degree = 13) were two hub genes. As shown in Figure 9(c), there were 111 nodes in the PPI network of the prefrontal cortex area, including 39 upregulated and 72 downregulated genes. SNAP25 was identified as a hub gene (degree = 26). There were 408 nodes in the PPI network of the cingulate gyrus, composed of 116 upregulated and 292 downregulated genes in Figure 9(d). CTGF (degree = 3), CDH1 (degree = 3), and COL5A1 (degree = 3) were considered as hub genes.



(a)

(b)

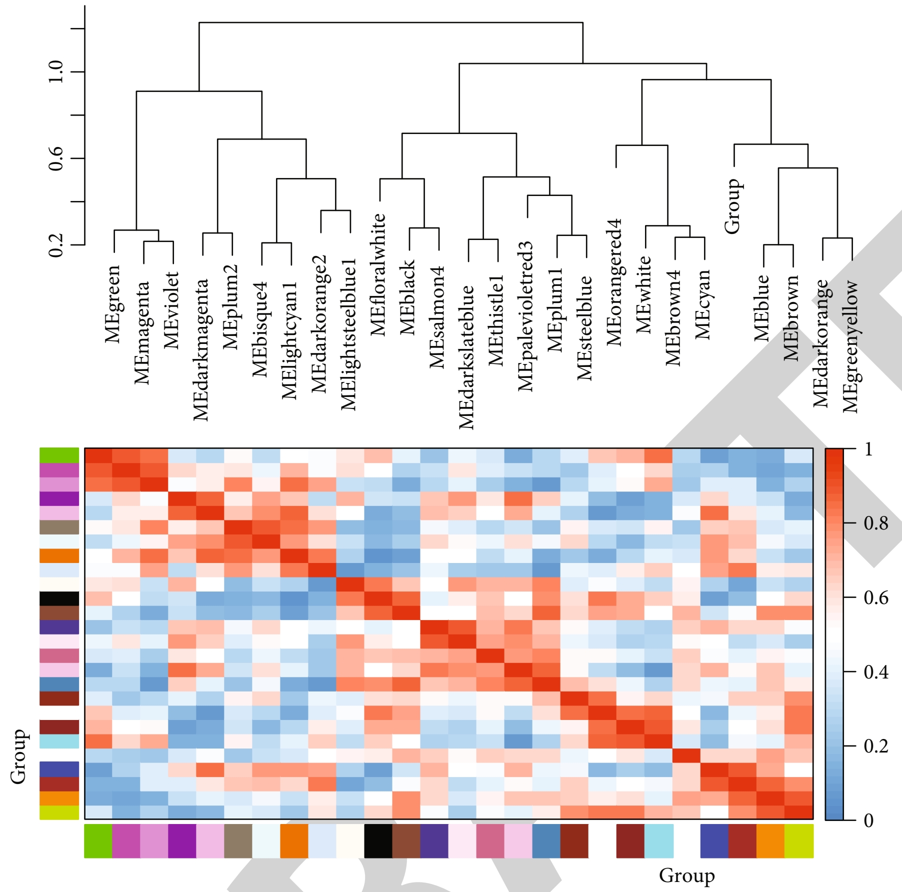
Figure 7: Continued.



(c)

(d)

Figure 7: Continued.



(e)

Figure 7: Construction of WGCNA for the prefrontal cortex of PD. (a) Sample hierarchical clustering tree to detect outliers. (b) Dynamic cutting tree method was utilized to determine gene modules. (c) Module-trait relationship network. (d) Hierarchical clustering dendrogram. The branches correspond to each module. (e) Clustering of module eigengenes and eigengene adjacency heat map.

3.8. Functional Enrichment Analysis of DEGs in the Substantia Nigra of PD. DEGs for the substantia nigra, putamen, prefrontal cortex area, and cingulate gyrus of PD were used for functional enrichment analysis. DEGs in the substantia nigra of PD were mainly enriched in PD-related biological processes such as aminergic neurotransmitter loading into synaptic vesicle, neurotransmitter transport, chemical synaptic transmission, amine transport, neurotransmitter loading into synaptic vesicle, dopamine biosynthetic process, phytoalexin metabolic process, cell-cell signaling, and isoquinoline alkaloid metabolic process (Figure 10(a)). These DEGs were involved in various key cellular components including neuron projection, cell projection, axon, plasma membrane bounded cell projection, postsynaptic membrane, synaptic membrane, presynapse, dendrite, dendritic tree, and neuronal cell body (Figure 10(b)). Also, they had several key molecular functions like monoamine transmembrane transporter activity, sodium : chloride symporter activity, dopamine binding, cytoskeletal adaptor activity, cation : chloride symporter activity, neurotransmitter : sodium symporter activity, spectrin binding, ammonium transmembrane transporter activity, protein serine/threonine kinase inhibitor activity, and chloride transmembrane transporter activity (Figure 10(c)). KEGG enrichment analysis results revealed that they were significantly related to a variety of PD-related pathways such as cocaine addiction, dopaminergic synapse, amphetamine addiction, serotonergic synapse, ECM-receptor interaction, alcoholism, tyrosine metabolism, Parkinson’s disease, PPAR signaling pathway, and synaptic vesicle cycle (Figure 10(d)).

3.9. Functional Enrichment Analysis of DEGs in the Putamen of PD. GO enrichment analysis results showed that DEGs in the putamen of PD were involved in the regulation of ossification, cell population proliferation, cartilage development, kidney morphogenesis, mesonephros development, chondrocyte differentiation, detection of abiotic stimulus, nephron morphogenesis, and cartilage development (Figure 10(e)). They were significantly involved in integral component of plasma membrane, intrinsic component of plasma membrane, amino acid transport complex, endoplasmic reticulum lumen, cell periphery, plasma membrane, extracellular space, cell surface, cell leading edge, and cytoplasmic side of plasma membrane (Figure 10(f)).

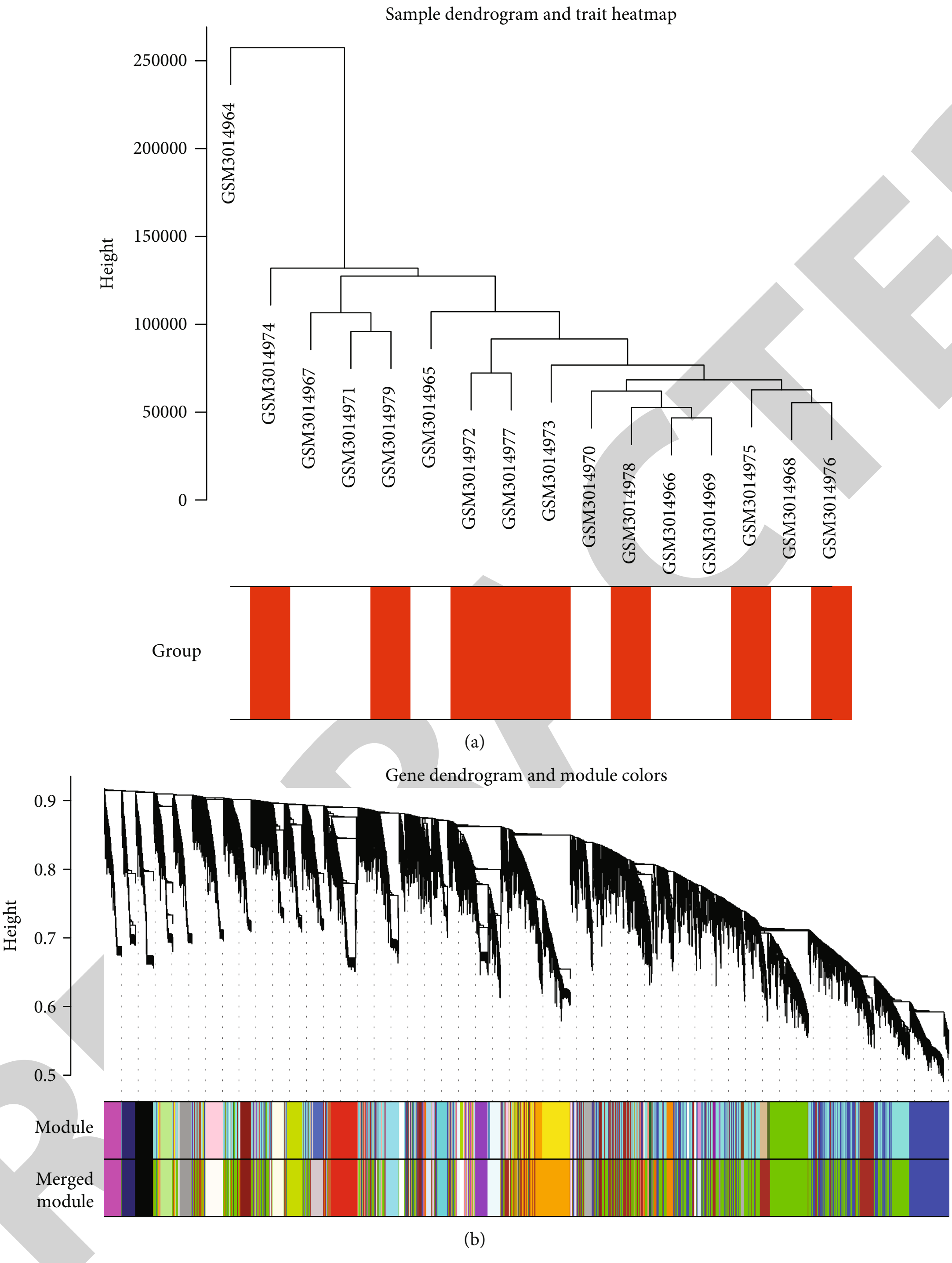
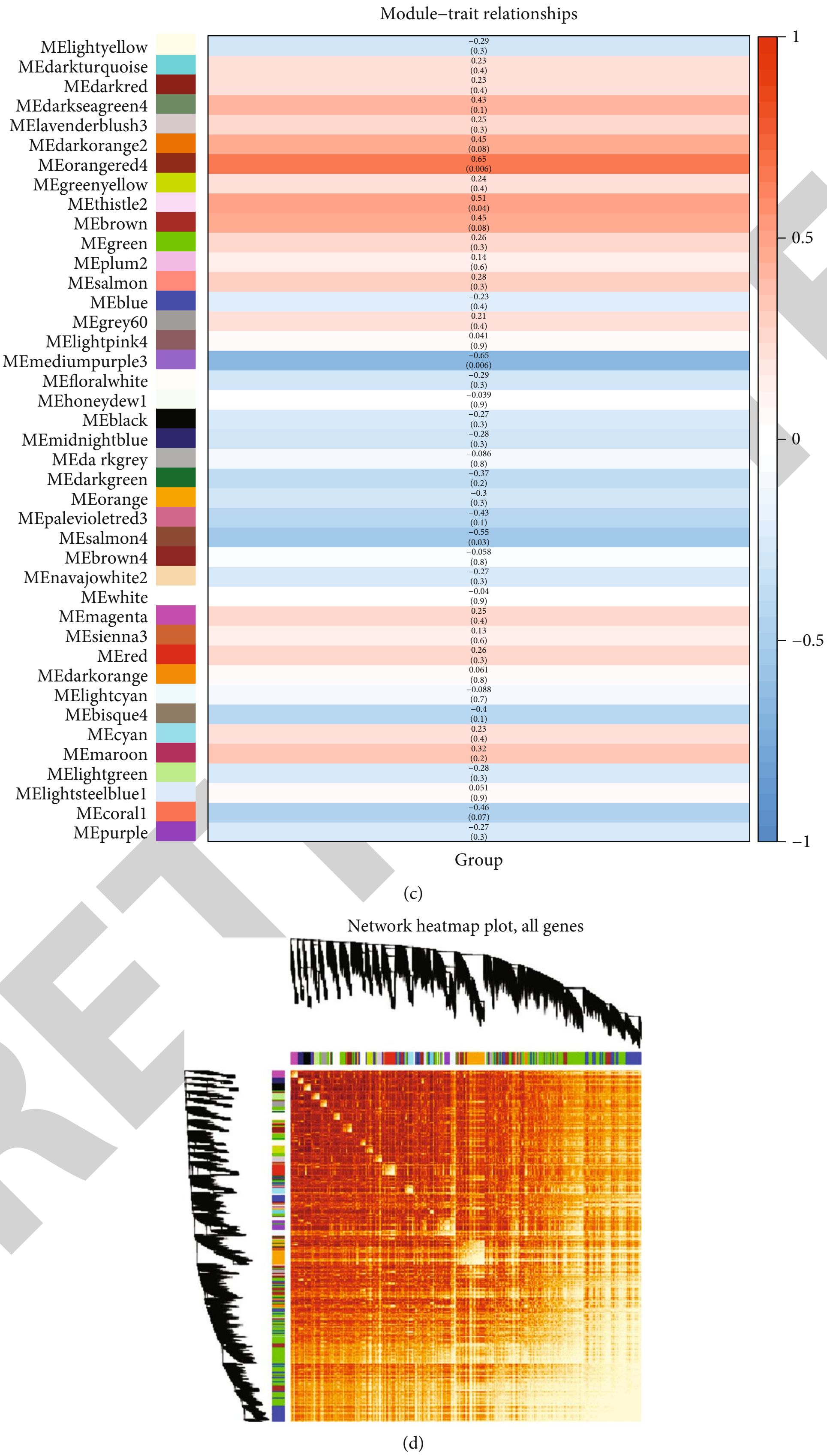


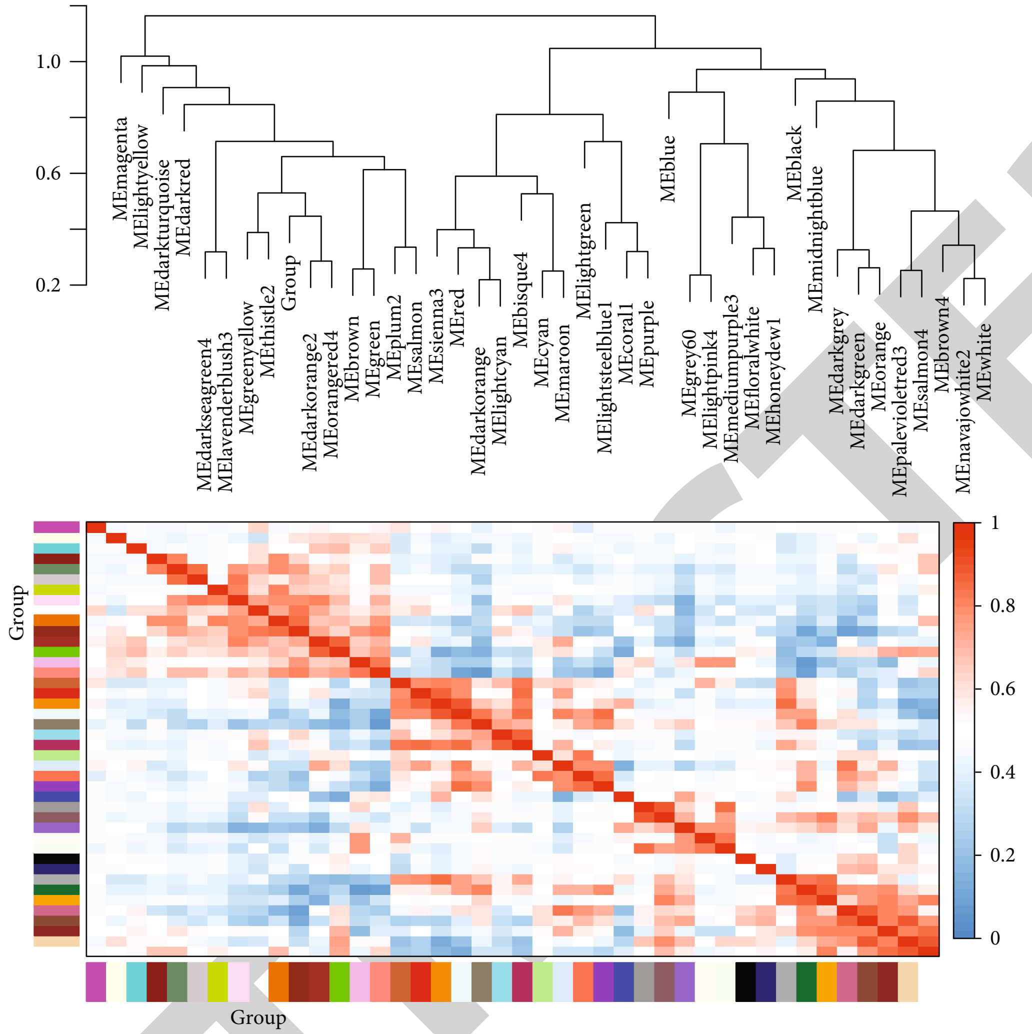
Figure 8: Continued.



(c)

(d)

Figure 8: Continued.

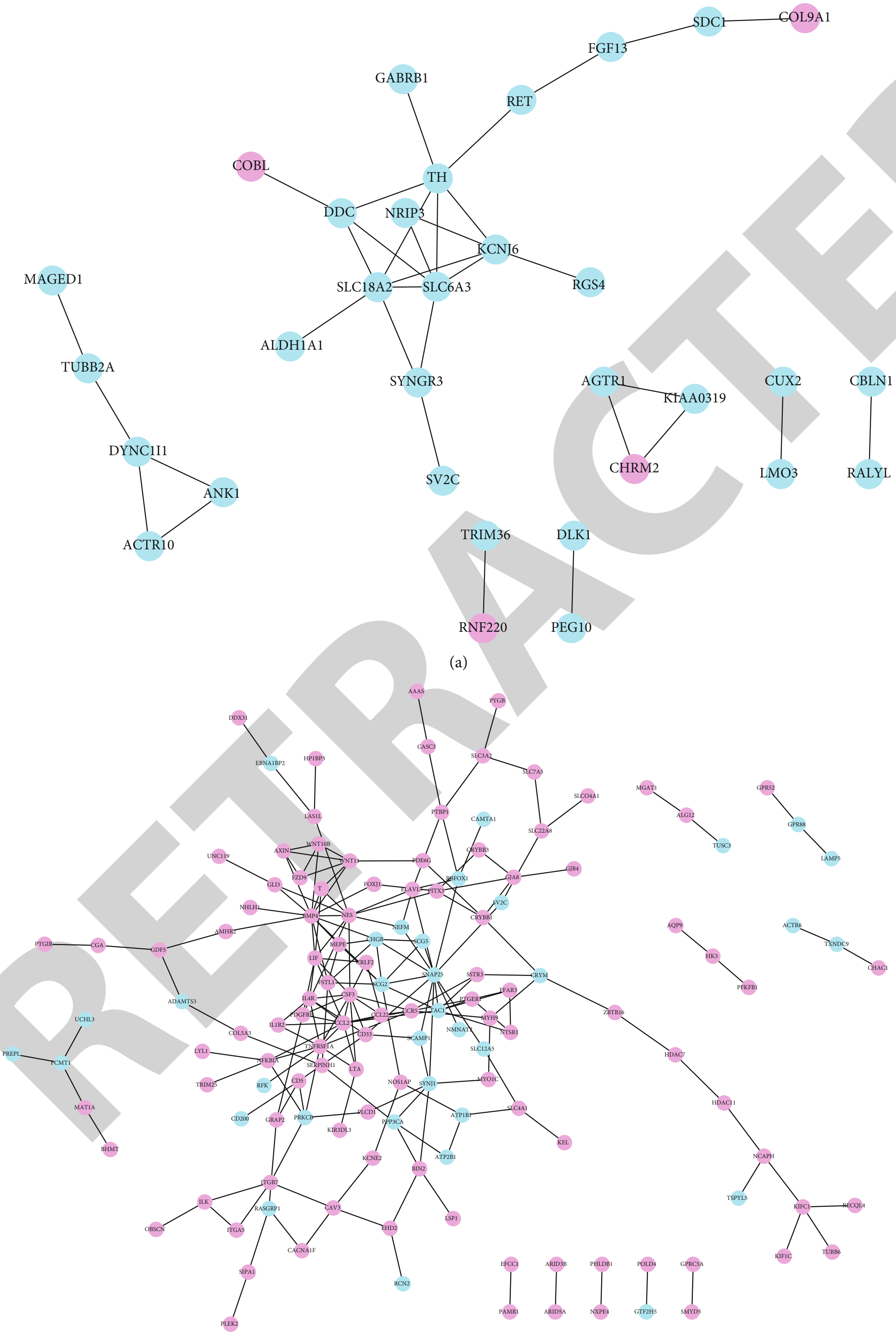


(e)

Figure 8: Construction of WGCNA for the cingulate gyrus of PD. (a) Sample hierarchical clustering tree to detect outliers. (b) Dynamic cutting tree method was utilized to determine gene modules. (c) Module-trait relationship network. The color of the square indicates the correlation between the module and the clinical traits. The p value is in brackets. (d) Hierarchical clustering dendrogram. The branches correspond to each module. The module memberships colored by different colors are shown in the color bar below and to the right of the tree diagram. Shades of color are proportional to coexpression interconnectedness. (e) Clustering of module eigengenes and eigengene adjacency heat map. Red represents high correlation and blue represents low correlation.

Also, they possessed a variety of molecular functions like heparin binding, cytokine binding, prostaglandin receptor activity, NAD-dependent histone deacetylase activity, cytokine activity, phosphoric diester hydrolase activity, cytokine receptor activity, actin-dependent ATPase activity, inorganic anion exchanger activity, and protein membrane anchor (Figure 10(g)). In Figure 10(h), these DEGs participated in key signaling pathways like cytokine-cytokine receptor interaction, basal cell carcinoma, hematopoietic cell lineage, calcium signaling pathway, mTOR signaling pathway, aldosterone synthesis and secretion, viral protein interaction with cytokine and cytokine receptor, signaling pathways regulating pluripotency of stem cells, NF-kappa B signaling pathway, and central carbon metabolism in cancer.

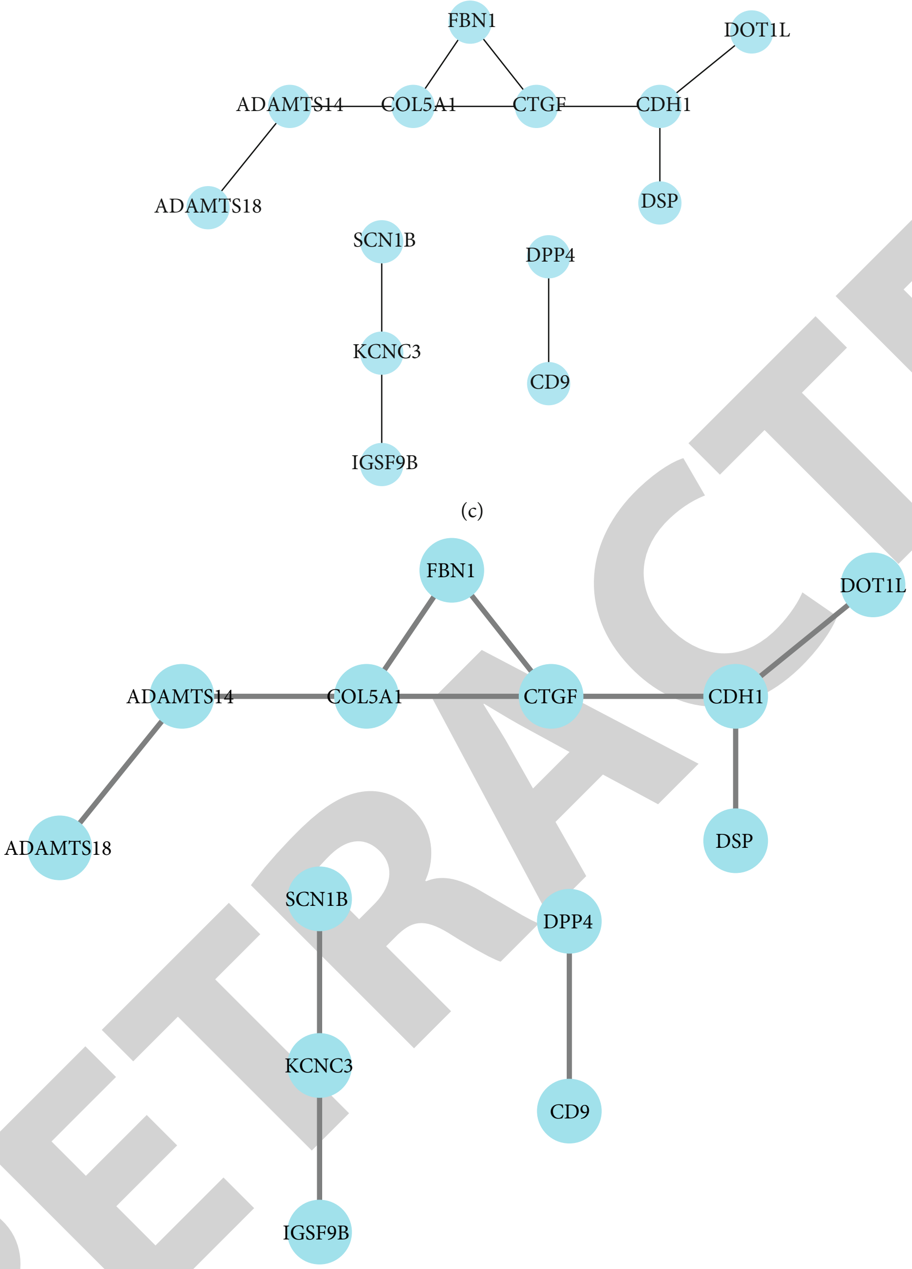
3.10. Functional Enrichment Analysis of DEGs in the Prefrontal Cortex of PD. GO enrichment analysis of DEGs in the prefrontal cortex of PD revealed that detoxification of copper ion, stress response to metal ion, chemical synaptic transmission, cellular response to zinc ion, cellular response to copper ion, nervous system development, cellular zinc ion homeostasis, cell communication, zinc ion homeostasis, and cellular response to cadmium ion were mainly enriched in Figure 10(i). They could regulate various cellular components like neuron projection, axon, synaptic membrane, plasma membrane bounded cell projection, cell projection, postsynapse, presynaptic membrane, presynapse, GABA-ergic synapse, and cell body (Figure 10(j)). In Figure 10(k), they had a variety of molecular functions like calcium ion binding, adiponectin binding, structural constituent of



(a)

(b)

Figure 9: Continued.



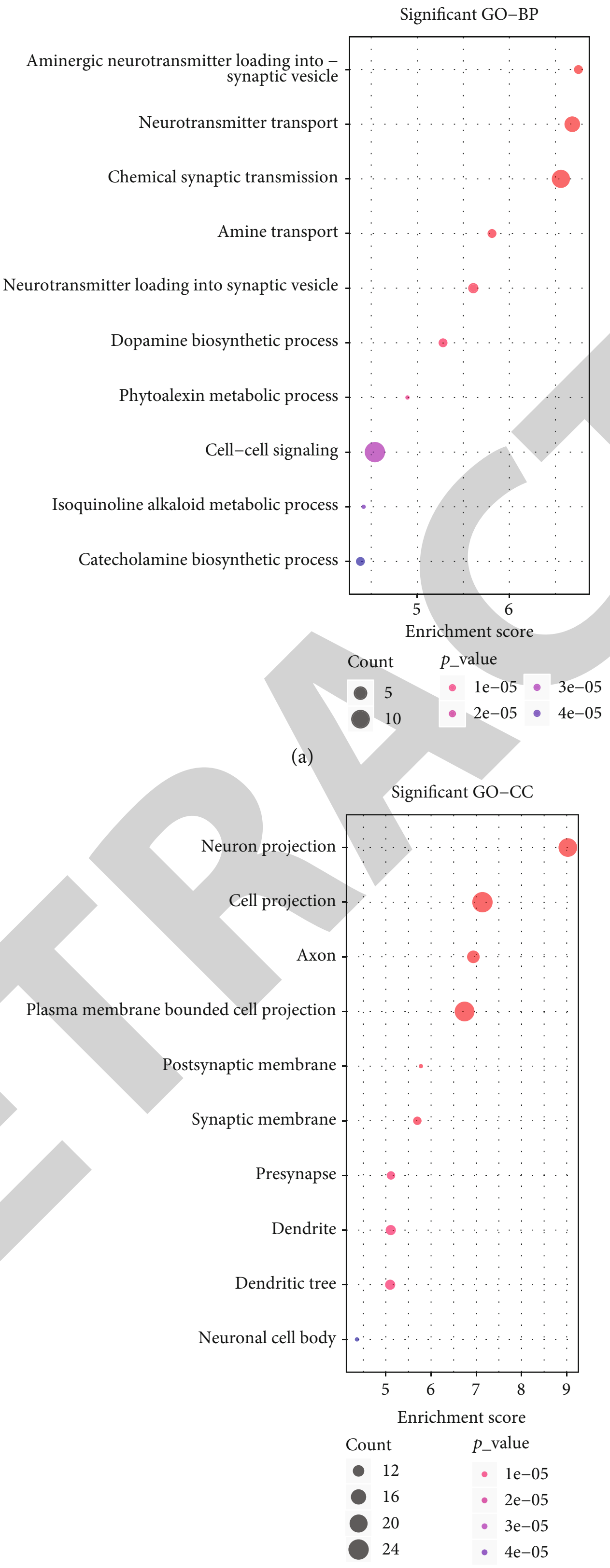
(c)

(d)

Figure 9: PPI networks for of DEGs in multiple brain regions of PD: (a) substantia nigra; (b) putamen; (c) prefrontal cortex area; (d) cingulate gyrus. Red expresses upregulation and blue expresses downregulation.

presynaptic active zone, G-protein alpha-subunit binding, calmodulin binding, benzodiazepine receptor activity, GABA-gated chloride ion channel activity, inositol 1,4,5-trisphosphate binding, inhibitory extracellular ligand-gated ion channel activity, and 1-phosphatidylinositol binding. KEGG enrichment analysis results demonstrated that mineral absorption, IL-17 signaling pathway, TNF signaling pathway, adipocytokine signaling pathway, synaptic vesicle cycle, mTOR signaling pathway, insulin secretion, gap junction, neuroactive ligand-receptor interaction, and phosphatidylinositol signaling system (Figure 10(l)).

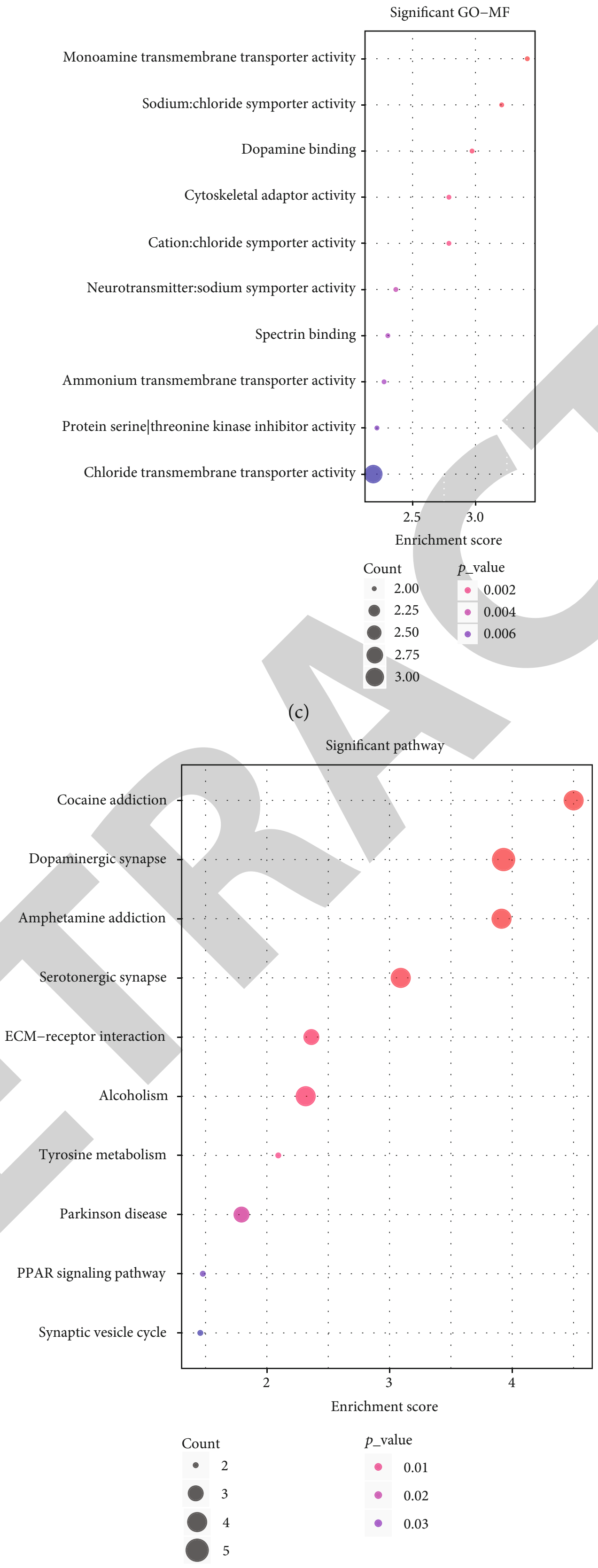
3.11. Functional Enrichment Analysis of DEGs in the Cingulate Gyrus of PD. GO enrichment analysis of DEGs in the cingulate gyrus of PD was performed. These genes could regulate a variety of biological processes like ion transmembrane transport, heart rate by cardiac conduction, ion transport, atrial cardiac muscle cell membrane depolarization, cell communication involved in cardiac conduction, cell-cell junction organization, cofactor transport, platelet aggregation, monovalent inorganic cation transport, and bundle of His cell to Purkinje myocyte communication (Figure 10(m)). As shown in Figure 10(n), they could



(a)

(b)

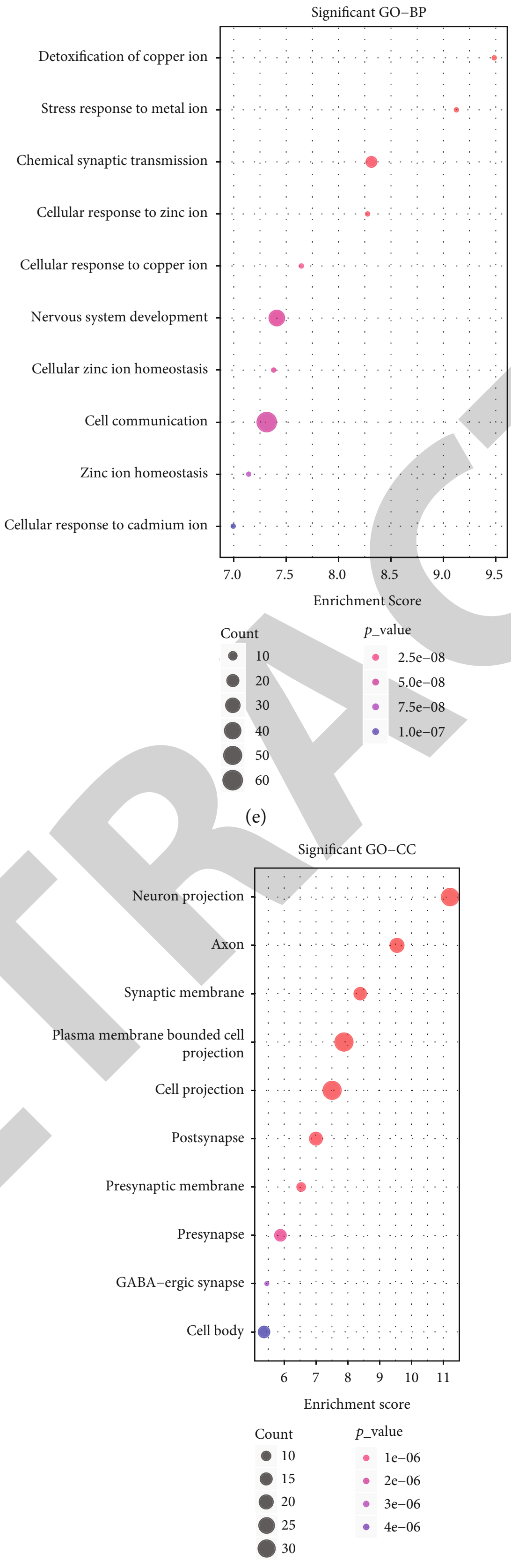
Figure 10: Continued.



(c)

(d)

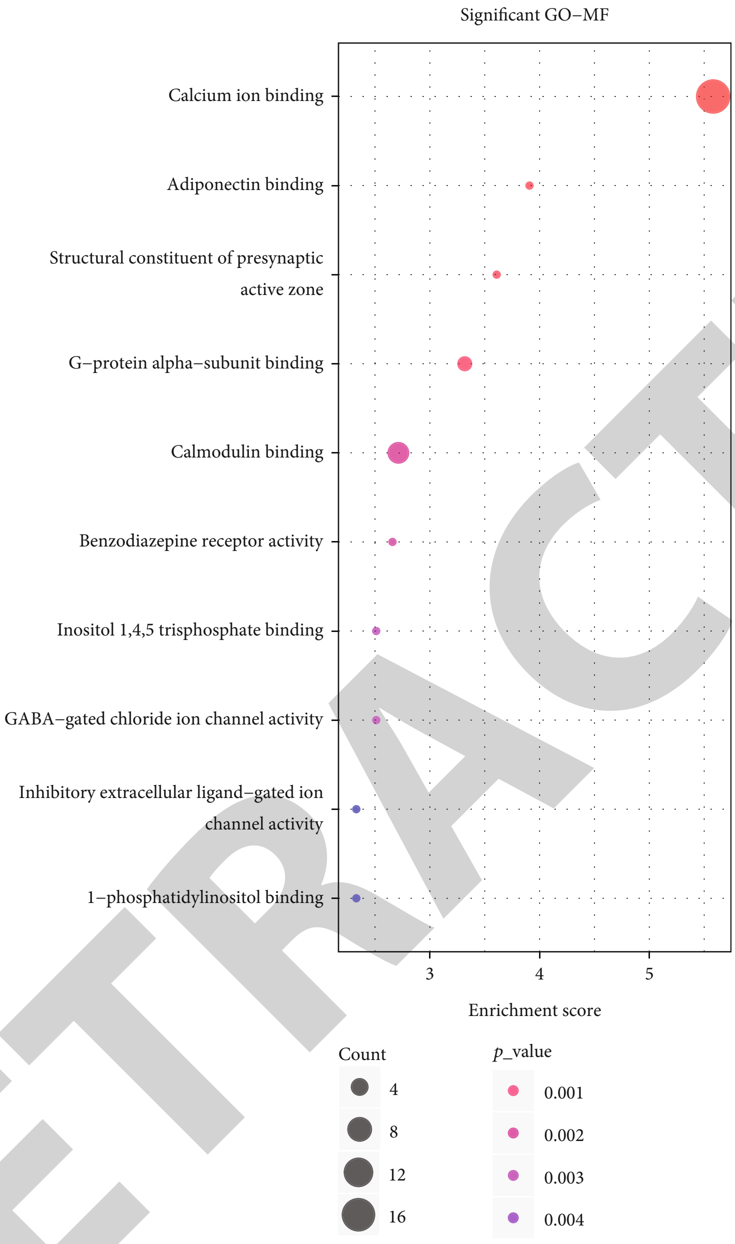
Figure 10: Continued.



(e)

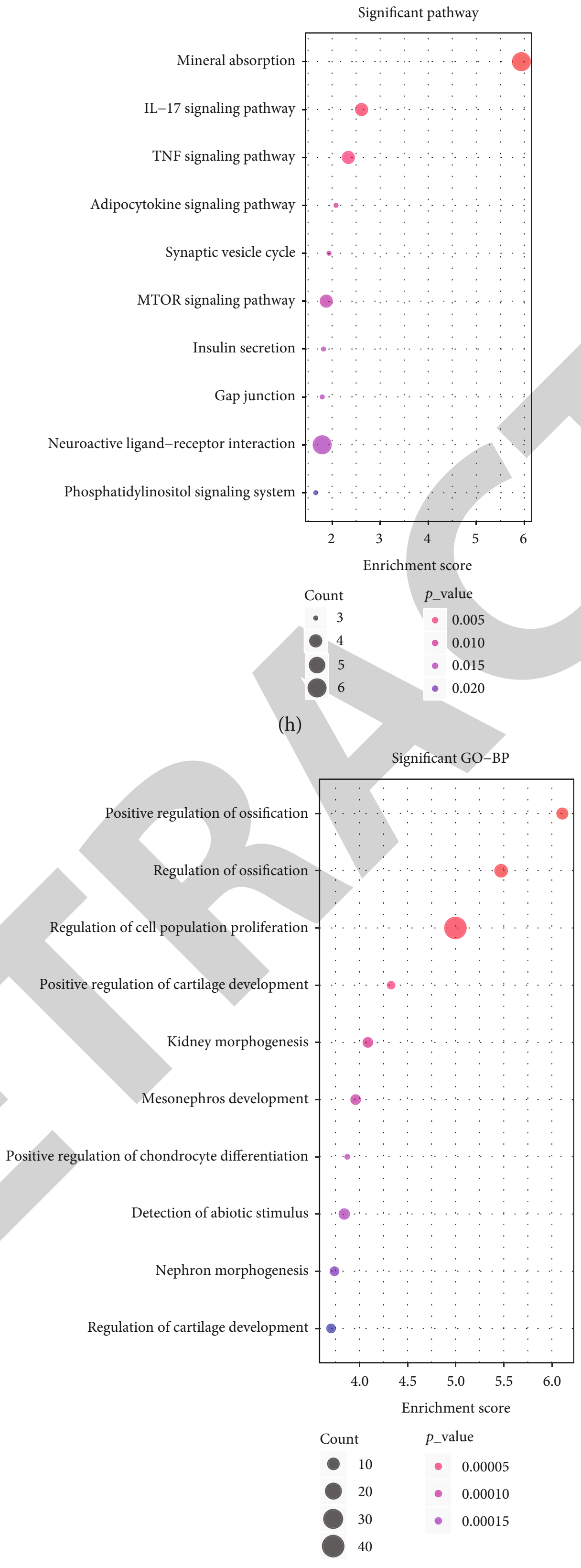
(f)

Figure 10: Continued.



(g)

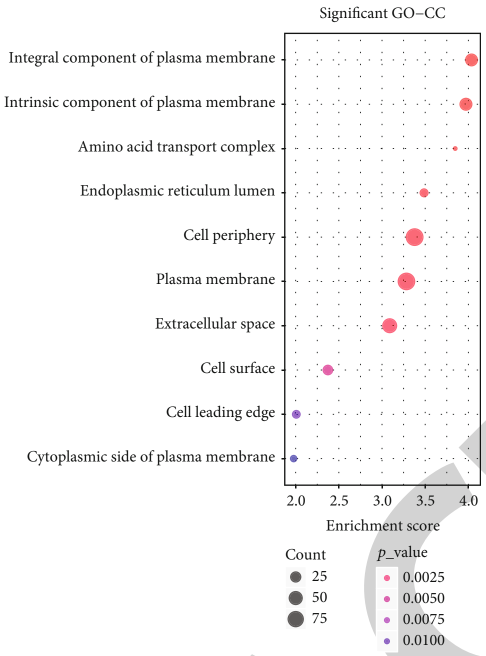
Figure 10: Continued.



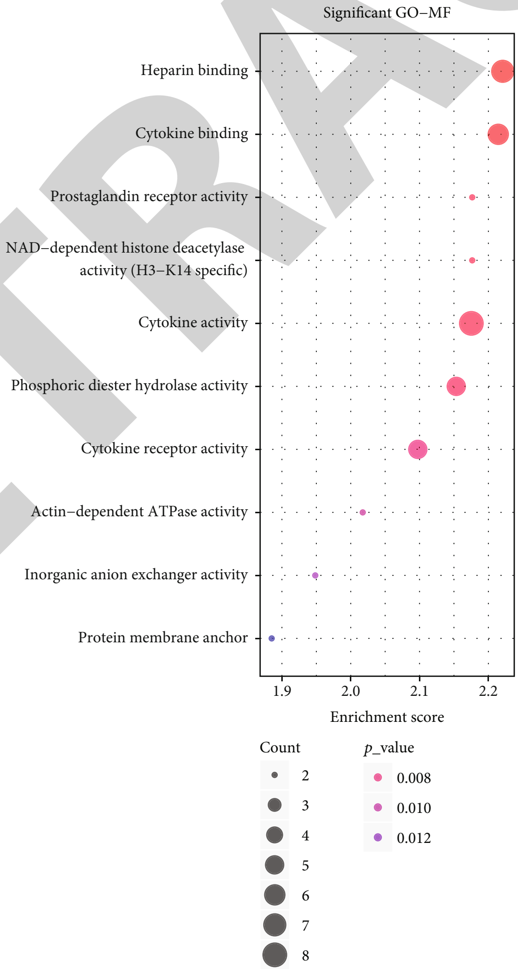
(h)

(i)

Figure 10: Continued.

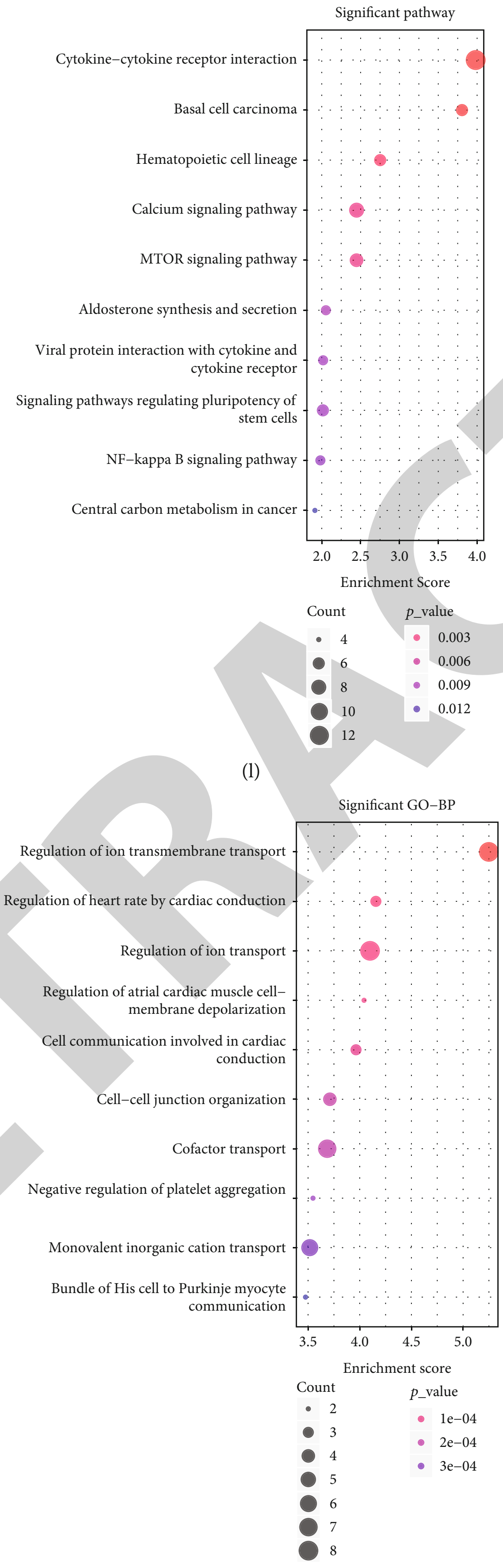


(j)



(k)

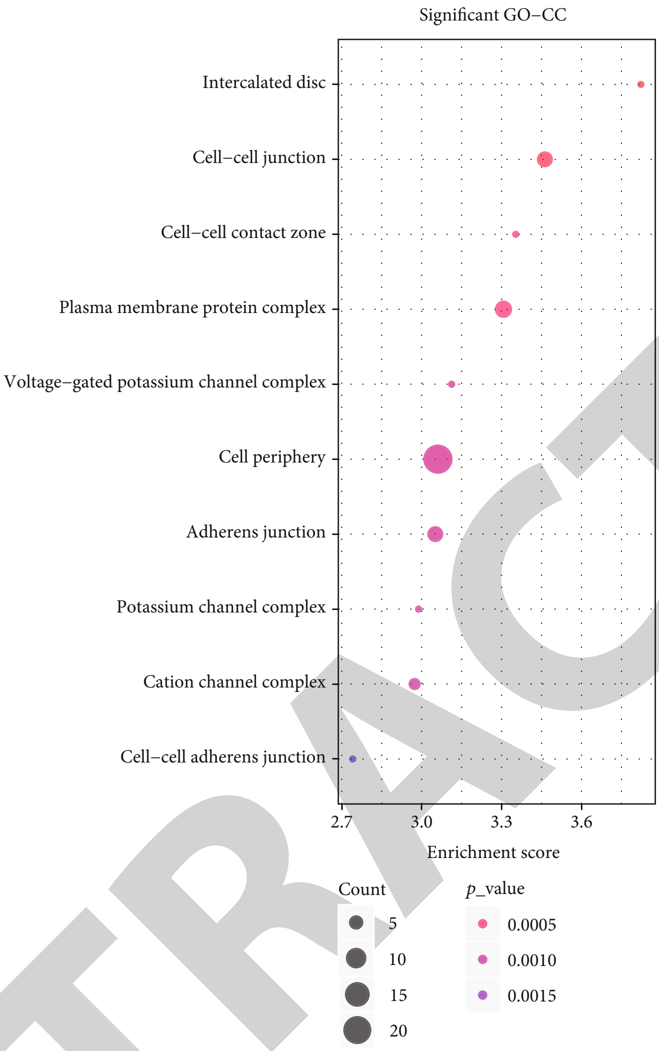
Figure 10: Continued.



(l)

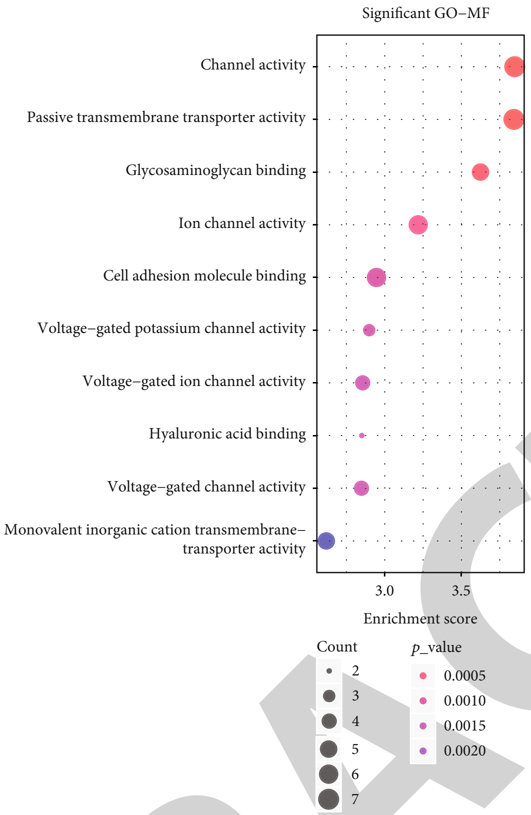
(m)

Figure 10: Continued.

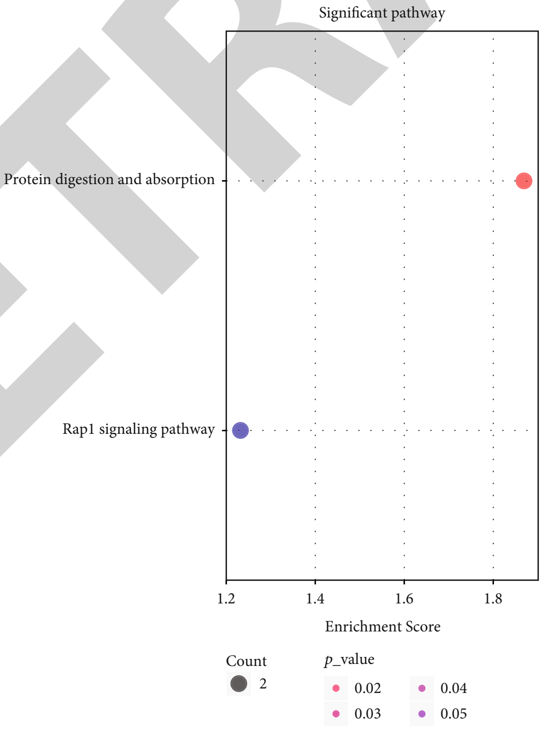


(n)

Figure 10: Continued.



(o)



(p)

Figure 10: GO and KEGG enrichment analysis results of DEGs for the substantia nigra, putamen, prefrontal cortex area, and cingulate gyrus of PD. GO terms included biological process (BP), cellular component (CC), and molecular function (MF). (a–d) DEGs in the substantia nigra of PD. (e–h) DEGs in the putamen of PD. (i–l) DEGs in the prefrontal cortex area of PD. (m–p) DEGs in the cingulate gyrus of PD.

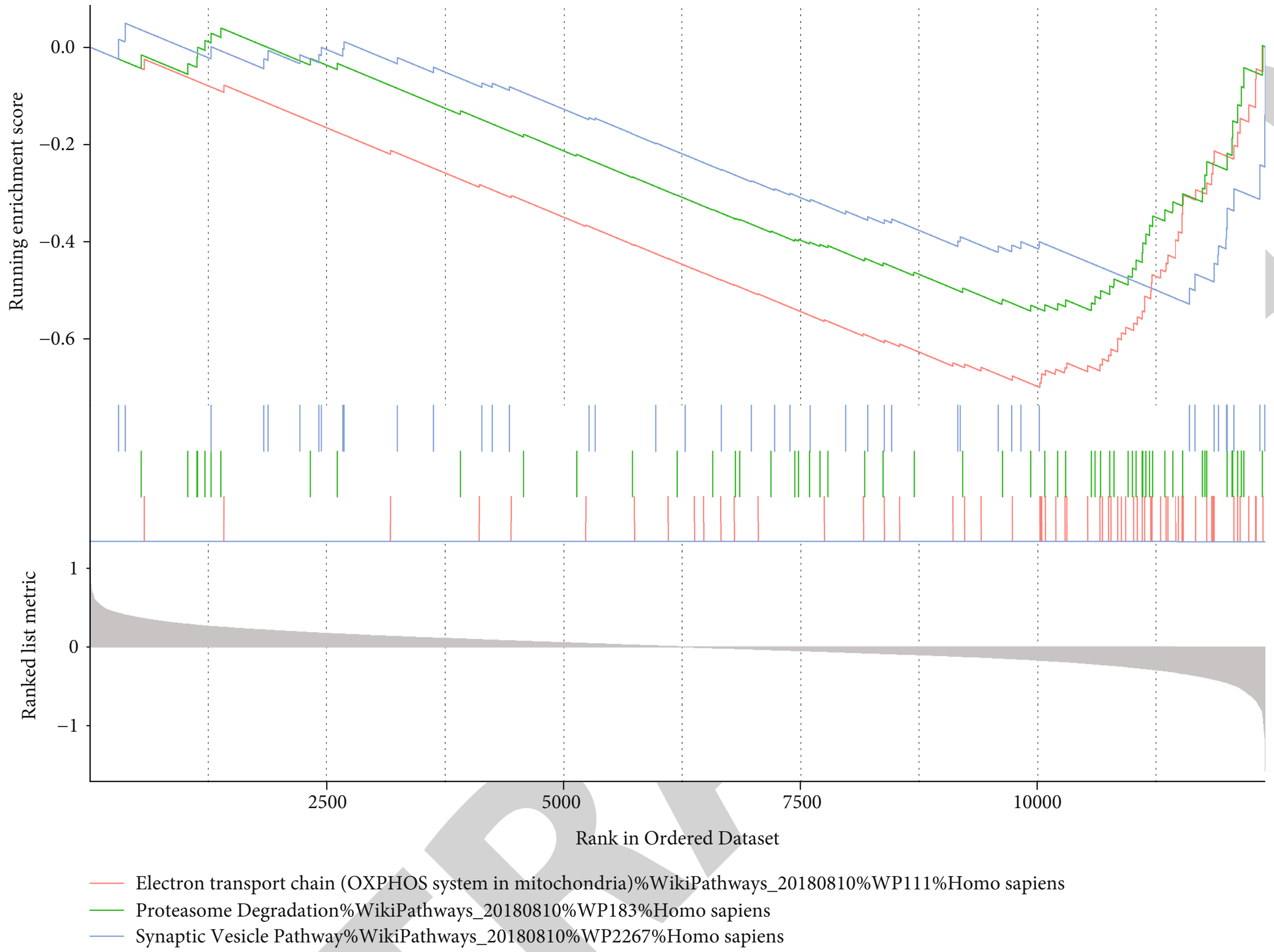


Figure 11: GSEA results of DEGs in multiple brain regions of PD.

distinctly participate in cellular components of intercalated disc, cell-cell junction, cell-cell contact zone, plasma membrane protein complex, voltage-gated potassium channel complex, cell periphery, adherens junction, potassium channel complex, cation channel complex, and cell-cell adherens junction. They could significantly have molecular functions of channel activity, passive transmembrane transporter activity, glycosaminoglycan binding, ion channel activity, cell adhesion molecule binding, voltage-gated potassium channel activity, voltage-gated ion channel activity, hyaluronic acid binding, voltage-gated channel activity, and monovalent inorganic cation transmembrane transporter activity (Figure 10(o)). Furthermore, our KEGG enrichment analysis results demonstrated that protein digestion and absorption and Rap1 signaling pathway were significantly enriched (Figure 10(p)).

3.12. GSEA of DEGs in Multiple Brain Regions of PD. GSEA was carried out based on DEGs in multiple brain regions of PD in the GSE20295 dataset. As depicted in Figure 11, these DEGs were most significantly enriched in electron transport chain, proteasome degradation, and synaptic vesicle pathway.

4. Discussion

Herein, we identified critical genes and pathways for multiple brain regions including the substantia nigra, putamen, prefrontal cortex area, and cingulate gyrus in PD by WGCNA, which deepened the understanding of PD-related molecular mechanisms.

In this study, we screened DEGs for the substantia nigra (17 upregulated and 52 downregulated genes), putamen (317 upregulated and 317 downregulated genes), prefrontal cortex area (39 upregulated and 72 downregulated genes), and cingulate gyrus (116 upregulated and 292 downregulated genes) of PD based on microarray and RNA-seq expression profiles. The regulatory relationship between genes is specific in time and space. In different organs and tissues, this regulatory relationship changes accordingly, which determines the occurrence and development of PD. To achieve specific biological functions of living organisms, the modularization of biological networks was conducted. WGCNA provides us with a simple and effective method to understand the regulatory relationship between genes, which is an indispensable method in systems biology research. Using WGCNA, gene modules were separately built for multiple brain regions of

PD. Based on PD-related DEGs, we visualized the PPI networks for the substantia nigra, putamen, prefrontal cortex area, and cingulate gyrus of PD by the Cytoscape software. The typical feature of the PPI network is that most of the nodes in the network are connected to only a few nodes, and there are very few nodes connected to a very large number of nodes. These nodes connected to many nodes are important nodes called as hub genes in the network. These hub genes could be involved in regulating many biological processes. In this study, hub genes in the PPI networks were identified for the substantia nigra (SLC6A3, SLC18A2, and TH), putamen (BMP4 and SNAP25), prefrontal cortex area (SNAP25), and cingulate gyrus (CTGF, CDH1, and COL5A1) of PD through the cytoHubba plug-in. Among them, SLC6A3 gene polymorphism has been found to be related to dopamine overdose in PD [23]. It has been identified as a hub gene for PD progression in a previous study, which is consistent with our study [24]. SLC6A3 genotype may affect cortical striatum activity in PD [25]. A meta-analysis reveals that SLC6A3 is a risk factor for PD [26]. SLC18A2 functions abnormally in the human PD brain. Improving SLC18A2 levels can increase the efficacy of levodopa [27]. SNAP25 gene polymorphism may prevent PD and mediate the severity of disease [28]. Furthermore, CDH1 expression is related to substantia nigra degeneration in a PD mouse model [29]. However, the functions of most of genes should be further explored in PD.

These DEGs in multiple brain regions were involved in distinct biological functions and pathways. GSEA showed that these DEGs were all significantly enriched in electron transport chain, proteasome degradation, and synaptic vesicle pathway, which have been widely accepted to be related to PD progression. For example, Coenzyme Q10 as a component of the electron transport chain may prevent neurodegeneration in response to mitochondrial deficiency and oxidative stress, which possesses potential as a target for treatment and intervention of PD [30]. Proteasome degradation induced by misfolding could contribute to the development of PD [31]. Also, abnormal accumulation of synaptic vesicle-associated protein is related to PD [32]. Thus, these critical pathways enriched by DEGs may be involved in the pathogenesis of PD.

Our results identified biologically significant gene modules by WGCNA and discovered clinical information-related hub genes, which were consistent with literature reports, thereby proving the accuracy and effectiveness of our WGCNA analysis results. Further excavation of gene module information may assist us to have an in-depth understanding on the role and significance of hub genes and signal pathways during PD progression. In our future studies, we will continue to validate the biological functions of these hub genes and key pathways in PD progression by a series of in vivo and in vitro experiments.

5. Conclusion

Taken together, this study identified hub genes for multiple brain regions including the substantia nigra (SLC6A3, SLC18A2, and TH), putamen (BMP4 and SNAP25), prefrontal cortex area (SNAP25), and cingulate gyrus (CTGF, CDH1, and COL5A1) in PD based on WGCNA. Furthermore, PD-related key pathways were identified including electron transport chain, proteasome degradation, and synaptic vesicle pathway. These findings could provide novel insights into the molecular mechanisms of PD.

Abbreviations

PD:	Parkinson’s disease
WGCNA:	Weighted gene coexpression network analysis
DEGs:	Differentially expressed genes
FC:	Fold change
GEO:	Gene Expression Omnibus
PPI:	Protein-protein interaction
GO:	Gene Ontology
KEGG:	Kyoto Encyclopedia of Genes and Genomes
GSEA:	Gene Set Enrichment Analyses.

Data Availability

The datasets analyzed during the current study are available from the corresponding author on reasonable request.

Conflicts of Interest

The authors declare no conflicts of interest.

Authors’ Contributions

Jianjun Huang and Li Liu contributed equally to this work.

Acknowledgments

This study was supported by the Middle-Aged and Young Teachers in Colleges and Universities in Guangxi Basic Ability Promotion Project (2017KY0512), the Starting Research Projects for Introducing Dr. of Youjiang Medical University for Nationalities (2015bsky001), and the First Batch of High-Level Talent Scientific Research Projects of the Affiliated Hospital of Youjiang Medical University for Nationalities in 2019 (R20196315).

References

[1] S. G. Reich and J. M. Savitt, “Parkinson’s disease,” *The Medical Clinics of North America*, vol. 103, no. 2, pp. 337–350, 2019.

[2] L. M. Chi, L. P. Wang, and D. Jiao, “Identification of differentially expressed genes and long noncoding RNAs associated with Parkinson’s disease,” *Parkinson’s Disease*, vol. 2019, Article ID 6078251, 7 pages, 2019.

[3] M. T. Hayes, “Parkinson’s disease and parkinsonism,” *The American Journal of Medicine*, vol. 132, no. 7, pp. 802–807, 2019.

[4] P. M. Antony, N. J. Diederich, R. Krüger, and R. Balling, “The hallmarks of Parkinson’s disease,” *The FEBS Journal*, vol. 280, no. 23, pp. 5981–5993, 2013.

[5] A. S. Chen-Plotkin, R. Albin, R. Alcalay et al., “Finding useful biomarkers for Parkinson’s disease,” *Science Translational Medicine*, vol. 10, no. 454, article eaam6003, 2018.

[6] T. Kakati, D. K. Bhattacharyya, P. Barah, and J. K. Kalita, “Comparison of methods for differential co-expression analysis for disease biomarker prediction,” *Computers in Biology and Medicine*, vol. 113, article 103380, 2019.

Hindawi
BioMed Research International
Volume 2024, Article ID 9858439, 1 page
https://doi.org/10.1155/2024/9858439

Retraction

Retracted: Optimum Feature Selection with Particle Swarm Optimization to Face Recognition System Using Gabor Wavelet Transform and Deep Learning

BioMed Research International

Received 12 March 2024; Accepted 12 March 2024; Published 20 March 2024

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope

(2) Discrepancies in the description of the research reported

(3) Discrepancies between the availability of data and the research described

(4) Inappropriate citations

(5) Incoherent, meaningless and/or irrelevant content included in the article

(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] S. Ahmed, M. Frikha, T. D. H. Hussein, and J. Rahebi, "Optimum Feature Selection with Particle Swarm Optimization to Face Recognition System Using Gabor Wavelet Transform and Deep Learning," *BioMed Research International*, vol. 2021, Article ID 6621540, 13 pages, 2021.

Hindawi
BioMed Research International
Volume 2021, Article ID 6621540, 13 pages
https://doi.org/10.1155/2021/6621540

Research Article

Optimum Feature Selection with Particle Swarm Optimization to Face Recognition System Using Gabor Wavelet Transform and Deep Learning

Sulayman Ahmed,[1] Mondher Frikha,[1] Taha Darwassh Hanawy Hussein,[2] and Javad Rahebi [3]

[1] *ENETCOM, Universite de Sfax, Tunisia*
[2] *Kirkuk University, Kirkuk, Iraq*
[3] *Department of Software Engineering, Istanbul Ayvansaray University, Istanbul, Turkey*

Correspondence should be addressed to Javad Rahebi; cevatrahebi@ayvansaray.edu.tr

Received 1 January 2021; Revised 6 February 2021; Accepted 24 February 2021; Published 10 March 2021

Academic Editor: B. D. Parameshachari

In this study, Gabor wavelet transform on the strength of deep learning which is a new approach for the symmetry face database is presented. A proposed face recognition system was developed to be used for different purposes. We used Gabor wavelet transform for feature extraction of symmetry face training data, and then, we used the deep learning method for recognition. We implemented and evaluated the proposed method on ORL and YALE databases with MATLAB 2020a. Moreover, the same experiments were conducted applying particle swarm optimization (PSO) for the feature selection approach. The implementation of Gabor wavelet feature extraction with a high number of training image samples has proved to be more effective than other methods in our study. The recognition rate when implementing the PSO methods on the ORL database is 85.42% while it is 92% with the three methods on the YALE database. However, the use of the PSO algorithm has increased the accuracy rate to 96.22% for the ORL database and 94.66% for the YALE database.

1. Introduction

Face recognition has attracted a lot of interest in recent years [1]. It has become one of the main areas of study in machine vision, pattern recognition, and machine learning. In face recognition, the system selects a face that is more like the desired face according to the trained faces and considers it as the final answer.

Facial recognition was proposed in the 1960s. The first semiautomatic facial recognition system was produced by Woody Bledsoe, Helen Chan Kurt, and Charles Bisson [2]. However, the human face includes a number of details that have been used in many systems, such as artificial age classification [3, 4], facial identification [5], forecasting images and restoration apps [6, 7], description of gender and gestures [8], human-computer interaction (HCI), electronic consumer experience management and audience recording, and tracking of security cameras. Applications for face recognition include monitoring, forensic and medical apps, security applications, in the banks, detection of the person in international centers of transition, access control, and several different fields. Recently, facial recognition technologies were widely used in particular in areas needing strict security measures (airports, police stations, banks, sports fields, and surveillance of entry and exit from business companies).

Computer security is considered to be important in the world today [9]. Face recognition remains an important subject in computer vision sciences. This is because the current systems perform well in relatively controlled conditions but appear to fail when there are issues with facial images, for example, presenting a particular face that differs by various factors, such as variations in posture, position, occlusion, lighting, make-up, and noise- and blur-induced image damage. Although researchers have developed many technologies, multiple different solutions have been attempted to address the problem of changing conditions of the environment. These conditions are the main challenges to facial recognition. Difficulties of the face recognition problem derive from the fact that the faces tend to be approximately similar in their most typical shape (i.e., the front view), and the variations between them are very slight. As a consequence, frontal images formalize a large concentration of the size of the image. This size makes it nearly difficult for typical pattern recognition methods to recognize correctly with a high degree of level of success [10]. Another concentration is the database images [11]. It must have sufficient information for effective face recognition, so the recognition must be possible when dealing with the test image. It is also difficult to determine if there is enough information in the stored images so that the relevant information can be extracted from the databases. Often, unnecessary information is also present in the images of the database, resulting in higher storage consumption and higher processing times. In addition, the optimal size of the images requires to be stored in the databases for effective results [12, 13]. The image size can be compressed to the required size and be stored in the databases. When the image size is compressed, there would be a loss of features, but large numbers of these images can be stored and transmitted through the network fast [14].

In this paper, we used Gabor wavelet transform for feature extraction and then for reducing the features. To find the best feature, the PSO method is used. For the recognition of a face, the deep learning method with 6 layers is used.

2. Literature Review

Facial recognition is currently divided into two general categories: appearance-based methods, which statistically process the face, and model-based methods that operate geometrically [15]. For face recognition [16], discriminative dictionary learning and sparse representation are used. In their method, the Gabor amplitude images are implemented by the bank of Gabor filter. Furthemore, the local binary pattern (LBP) is used for feature extraction [17]. Face recognition can be considered one of the most significant applications in the image processing domain [18]. However, illumination and pose invariant recognitions are still the most obvious problems. Viewpoint and illumination are vital to the efficiency of the recognition system because these two factors differ when face images are taken in an uncontrolled environment. Elastic bunch graph matching [19], one of the feature-based methods, has been known for a long time to be accomplished toward several factors such as illumination and viewpoint [18]. Their excessive susceptibility to feature extraction and measurement of the extracted features [20] are what make them unreliable. As a result, the dominant method in the literature is appearance-based methods.

Ahonen et al. proposed a face recognition model with native binary patterns (LBP) [5]. In their study, the point ensuring the robustness of their work is that the algorithm is not sensitive to light.

The fisherface [21] technique is one of the milestones for face recognition under variations. In linear discriminant analysis (LDA), interperson alteration is used optimally with large and intraperson alteration efficaciously small to construct a subspace [21]. Like the PCA [22], the main disadvantage of this technique is that the data space is a consideration of Euclidean. The method does not succeed as multimodal distributed face images when data points are located in a nonlinear subspace.

The sparse representation algorithm based on the Gabor feature is proposed by Yang and Zhang [23]. In their method, the SRC and Gabor features are combined. Using this technique, they improved the human face recognition rate and reduced the complexity of computation.

The deep learning approaches are investigated [24]. The cross-resolution face recognition scenario based on the deep learning method is performed [25]. They robustly extracted the features by deep properties with a cross-resolution scenario. In [26], the angularly discriminative features based on deep learning for face recognition are utilized.

Xu et al. presented the new artificial neural network to face recognition called coupled autoencoder networks (CAN). This helps to overcome age-invariant face recognitions and redemption troubles [27].

The effect of variations in condition on face recognition has been investigated by authors [28]. Consequently, the dominant method has been the appearance-based method. Nikan and Ahmadi [29] introduced a new procedure that propped up fusion of global and local structures.

In [30], local linear transformations were used on behalf of one global transformation, which is a good improvement. The technique suggests different pose classes to different mapping functions. When a probe image is examined, its pose is determined by soft clustering. Deciding the number of pose clusters is a difficult task as in all clustering algorithms. Moreover, novel poses cannot be treated in case of critical variations. In [31], the authors used the neighborhood structure of the input space to determine the underlying nonlinear manifold of multimodal face images. What is used to calculate the basic set is called Laplacian Faces Locality Preserving Projections (LPP). When examining face images with other poses, facial expressions, and illumination conditions, their recognition performance was higher than that of fisherfaces or eigenfaces. In [32], pose variation using view-based eigenfaces was studied. For every view, eigenfaces were numbered to apply a standard dimensional subspace as separate transformations. In addition, a feature-based scheme is included within the eigenfeatures introduced by the authors. As in [33], their performance depends highly on decoupling. Here, the eigenilluminant field technique was used to identify the subspace of poses. Zhao et al. [34] prepared the blurry invariant binary identifier to face recognition. They enhanced the corral among the binary codes of sharp face images and blurred face images of positive image pairs about to learn matrix of projection. Then, they used the learned projection matrix to procure blur-robust binary codes by quantizing projected pixel difference vectors (PDVs) in the trial phase. The discriminative DL method by training a classifier of the coding coefficients is proposed

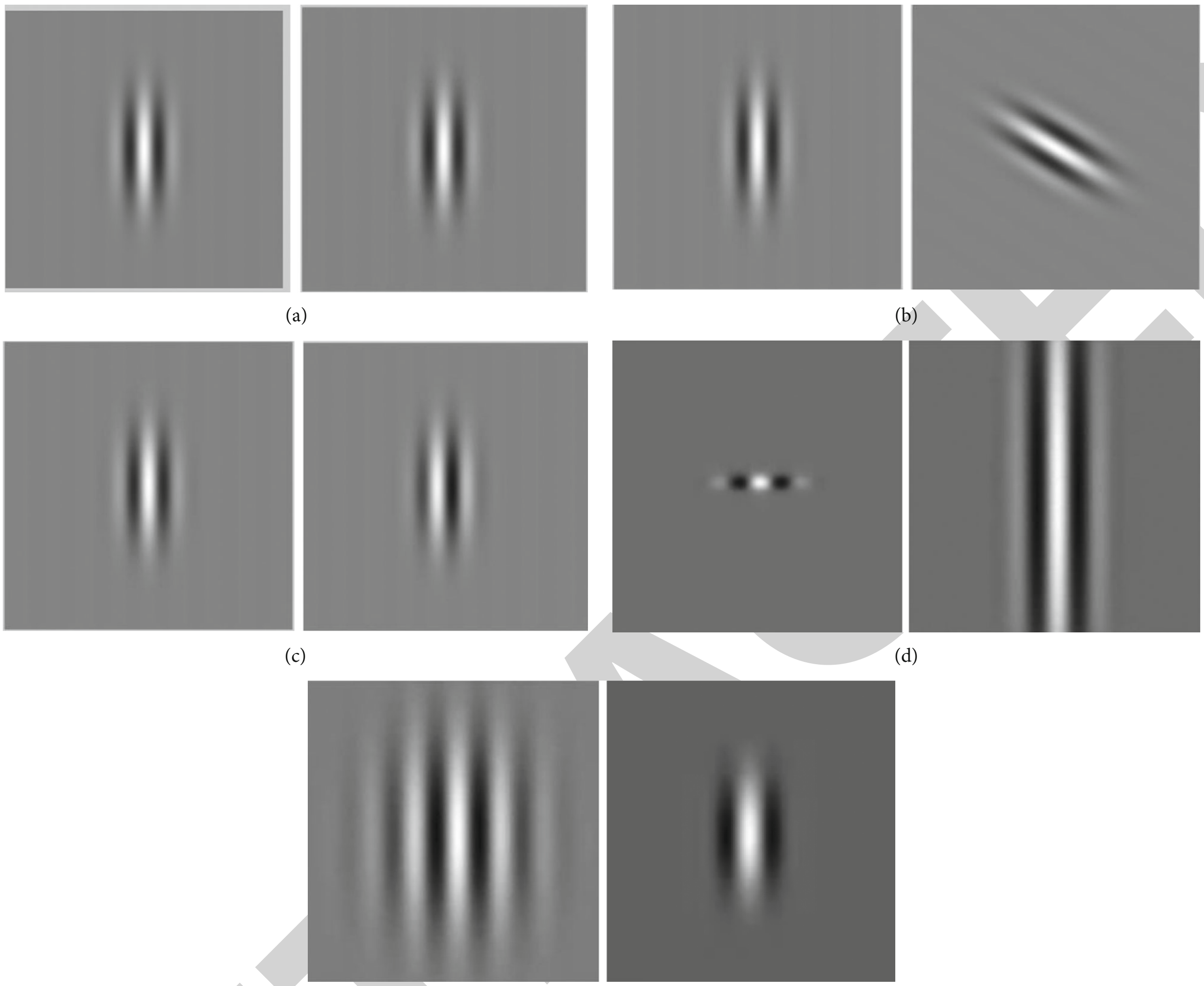


Figure 1: (a) Different values of wavelength (left: 25, right: 50), (b) different orientation (left: 0, right: 45), (c) changing the values of phase shift (left: 180, right: 90), (d) aspect ratio very large (left) and very small (right), and (e) different bandwidth values (left: large, right: small) [57].

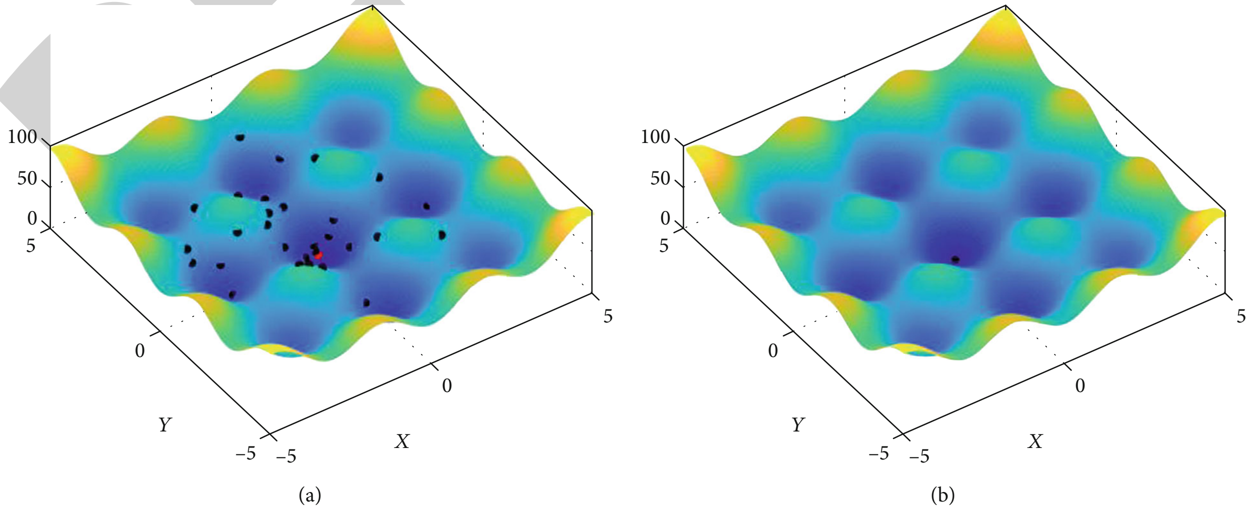


Figure 2: Particles searching for the best and optimum solution: (a) first iteration and (b) last iteration.

by Mairal et al. [35]. For texture classification and digit recognition, they verified their method. In [36], the sex and country population density interact to predict face recognition talent. The face plus word recognition based on the euro magnetic correlates of hemispheric specialization is presented in [37].

A method that is insensitive to illumination changes was produced by the authors in [38] through combining the generalized concept of photometric stereo and eigenlight field. 3D morphable face models were used in [20, 39, 40, 41] to defined novel poses, which have performances higher than that of the previous research works. Rendering ability for new poses and illumination conditions is exceptional with 3D morphable models [41]. However, the computational cost of generating 3D models from 2D images or using laser scanners to access 3D models decreases the efficiency of the recognition system.

Royer et al. [42] used the eye region to identify a face accurately. The mixed neighborhood topology with the cross decoded patterns is done by [43].

Illumination variance was studied in [44]. The quotient image was suggested by the authors as an identity signature that is insensitive to illumination. While the approximation does not work well, then the probe image has an unexpected shade. Its probe images could be identified with particular illumination then the gallery images. The technique requires only one gallery image for a thing. The technique in [45] introduced additional constraints on the albedo and the surface normal to solve the shadow problem. An illumination cone model was proposed in [20]. The authors discussed a series of images of an object in a fixed pose only describing a convex cone in all lighting conditions. The method needs some images to test their identity and then to guess its surface geometry and albedo map. They defined different illumination cones for each sampled viewpoint to deal with pose variations. The authors discussed in [46, 47] the use of all Lambert reflecting functions to create all kinds of illumination conditions for Lambertian objects. The researchers presented the approximation of plenty of variation of illumination achieved using only nine spherical harmonics. The multiple virtual views and alignment errors are presented in [48]. They manipulated the cross-pose face recognition method.

A methodology for recognition was also used in [46]. In [40], a spherical harmonics approach was exploited, and good recognition results were presented. They designed a 3D morphable model to achieve pose invariance, and this needs to generate 3D face models from 2D images.

Original and symmetrical examples of face training were used [49] to perform collaborative representation for face recognition.

A nonlinear subspace approach was introduced using the tensor representation of faces, such as facial expressions, illumination, and poses [50]. The *n* mode tensor Singular Value Decomposition (SVD) could form the basis of an image. In this technique, various images are required under different variations for each training identity. In [51], there was another nonlinear assumption for each identity in the database, and a gallery manifold is stored. When a test identity with several new poses needs to be defined, first, its probe manifold is constructed, then using manifold to manifold distance can help to define its identity.

Table 1: Parameter for particle swarm optimization to select the best features.

Parameter	Description	Value
N	Number of particles (population size)	40
T	Maximum number of iterations	15
c_1	Cognitive factor	3
c_2	Social factor	2.5
$w_{\max}$	Maximum bound on inertia weight	0.8
$w_{\min}$	Minimum bound on inertia weight	0.5
$V_{\max}$	Maximum velocity	5

The main drawback is the requirement of multiple images of the test person. The authors in [52] introduced a considerable idea by bilinear generative models to decompose orthogonal factors. They showed a separable bilinear mapping between the input space and the lower dimensional subspace. After determining all the parameters of mappings, identity and pose information can be separated explicitly. The recognition and synthesizing capabilities of the technique were analyzed, and the results were encouraging. In [53], illumination invariance was examined using a similar framework. In addition, a ridge regression technique was designed to come through the matrix inversion needed in the symmetric bilinear model. A modified asymmetric model in [54] is aimed at overcoming pose variations. One of the most important factors affecting performance is the solvation of the pose space. The authors in [55] incorporated the nonlinearity of the generative models. They recommended a nonlinear scheme combined with the bilinear model and tried to remove the linearity constraint of the classical generative models. Wright et al. [56] presented a robust method for face recognition. They used sparse representation for feature extraction.

3. Proposed Method

In this paper, the face recognition system undergoes stages. These three stages are feature extraction using Gabor wavelet transform, selecting the best features with the PSO method, and face recognition with the deep learning method.

3.1. Feature Extraction Using Gabor Wavelet Transform. A much useful instrument in image processing, especially in image identification, is the Gabor filter. The Gabor filter over the spatial field, which has two dimensions, is a Gaussian kernel function as explained below by a complex sinusoidal plane wave:

$$G(x, y) = \frac{f^2}{\pi\gamma\eta} \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) \exp\left(j2\pi f x' + \phi\right). \tag{1}$$

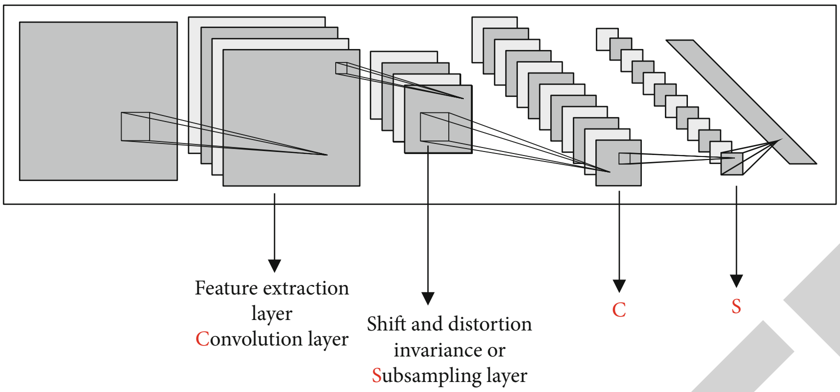


Figure 3: CNN's topology.

Here, f represents the sinusoid's frequency, θ is the orientation of the normal to the parallel Gabor function's stripes, ϕ is the phase offset, σ is the Gaussian envelope's standard deviation, and γ is the spatial aspect ratio that determines the elliptic support for the function of Gabor.

x' and y' can be calculated as the following equations:

$$\begin{aligned} x' &= x\cos\theta + y\sin(\theta), \\ y' &= -x\sin\theta + y\cos(\theta). \end{aligned} \tag{2}$$

Figure 1 shows the influence of changing some parameters for Gabor's function.

Some of the various benefits of Gabor filters are invariance rotation, scaling, translation, and resistance to distortion of images such as illumination change [58, 59]. They are specially proper for fabric representation plus discrimination.

A range of Gabor filters with other frequencies and directions can be used to extract many features such as texture analysis and segmentation from an image [60]. By varying the orientation, we can look for fabric orientation in a specific direction. By varying the standard deviation of the Gaussian envelope, we change the basis' support or the image's size region being analyzed.

When the features are extracted, the best relative set of features are selected using the PSO method for a flexible face recognition system.

3.2. Feature Selection with Particle Swarm Optimization. Particle swarm optimization (PSO) or known as the bird swarm algorithm was initially created in 1995 by Kenny and Eberhart [61]. PSO is a mathematical method that tries to solve optimization problems. For each problem, there are particles (solutions) flying over the problem area based on some mathematical calculations for the velocity and position of the particle. Each particle has fitness values that are measured by the fitness function to be optimized and has velocity that guides the flying of the particles [62].

In computational techniques, PSO is used as a random optimization algorithm for feature selection and classification. This is done by iteratively selecting the most relative and useful set of features to improve or maintain the classification performance for a robust facial recognition system [63].

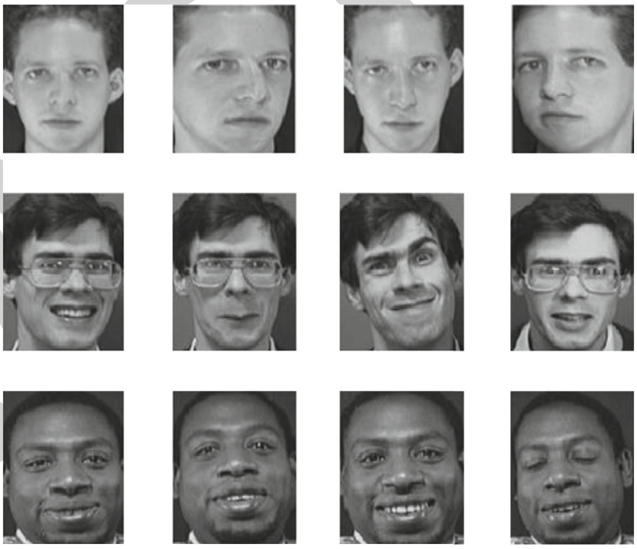


Figure 4: ORL database.

The basic idea behind this algorithm is the coevolvement of different classes of birds rather than focusing on a certain class of birds. This algorithm contributes to effective search abilities [64]. The PSO algorithm is illustrated in Figure 2.

First, all the particles are assigned primary values; after that, fit values for each particles are estimated. Then, the current fit value is determined; if it is better than the previous one, then we upgrade it to the current value, but if the old fit value is better; we keep it [65]. The algorithm ends, and this process is repeated until the best solution is obtained.

The equation of the PSO algorithm is demonstrated below:

$$\begin{aligned} v_i^d(t+1) = {} & w v_i^d(t) + c_1 r_1 \left(pbest_i^d(t) - x_i^d(t) \right) \\ & + c_2 r_2 \left(gbest^d(t) - x_i^d(t) \right). \end{aligned} \tag{3}$$

Each particle is upgraded with two "best" values in each iteration. Here, v denotes velocity which is bounded between $w_{\max}$ and $w_{\min}$, w is inertia weight, and x is solution [66, 67].

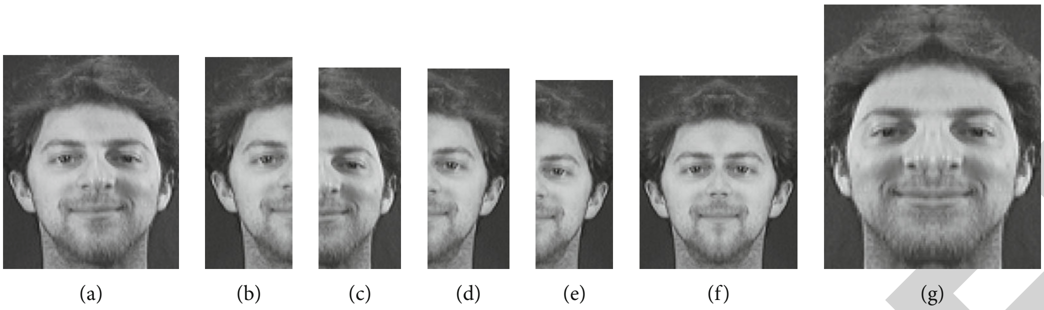


Figure 5: (a) Real image, (b) left side, (c) right side, (d) left side's mirror, (e) right side's mirror, (f) integration of left side with mirrors, and (g) integration of right side with mirrors.

Continuing, t refers to the number of irritation, i to the order of practicality in population, and d to the dimension of search space. c_1 and c_2 indicate acceleration factor; r_1 and r_2 are two independent random numbers in $[0, 1]$. $p$best implies the personal best solution (the best solution that has been found yet), while $g$best implies the global solution which is recorded by the particle swarm optimizer. This optimizer is the best worth yet achieved by any particle for the entire population.

Afterwards, velocity is updated to a probability value as demonstrated in the following equation:

$$s\left(v_i^d(t+1)\right) = \frac{1}{1+e^{\left(-v_i^d(t+1)\right)}}. \tag{4}$$

Practical position and $p$best with $g$best are converted to the following equations:

$$x_i^d(t+1) = \begin{cases} 1, & \text{if rand} < s\left(v_i^d(t+1)\right), \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

where rand is a random number between 0 and 1.

$$\begin{gathered} p\text{best}_i(t+1) = \begin{cases} x_i(t+1), & \text{if } F(x_i(t+1)) < F\,(p\text{best}_i(t)), \\ p\text{best}_i(t), & \text{otherwise,} \end{cases} \\ g\text{best}_i(t+1) = \begin{cases} p\text{best}_i(t+1), & \text{if } F(p\text{best}_i(t+1)) < F\,(g\text{best}_i(t)), \\ g\text{best}_i(t), & \text{otherwise,} \end{cases} \end{gathered} \tag{6}$$

where F is the fitness function:

$$w = w_{\max} - (w_{\max} - w_{\min})\left(\frac{t}{T_{\max}}\right). \tag{7}$$

The parameter used for particle swarm optimization is shown in Table 1.

We obtained these parameters experimentally.

Table 2: Specifications of the deep learning.

	Type	Activation	Learnable
1	Sequence input	5141	—
2	LSTM	200	Input weights (800 ∗ 5141) Recurrent weights (800 ∗ 200) Bias (800 ∗ 1)
3	Fully connected	50	Weights 50 ∗ 200 Bias 50 ∗ 1
4	Dropout	50	—
5	Fully connected	1	Weights 1 ∗ 50 Bias 1 ∗ 1
6	Regression output	—	—

3.3. Convolutional Neural Network. The main component of a convolution neural network (CNN) is the convolution layer. The approach behind a convolution layer is a feature which has been learned locally for any given input (for example, any 2D images). It should be helpful in other regions of that same input source. For example, a feature for edge detection, which was proved useful in one part of the image, might be helpful in the other regions of the image at a possible general feature extraction stage. The learning of other features in an image such as edges oriented at an angle or curves is obtained by sliding the filters across the image with a step or stride size which is constant for a given convolution layer.

Layers of more than one subsampling and convolutional layer, preferably fully added layers, are called CNN. M is the height and width of the image, and r is the number of channels, while the input of an accessible layer is the image $m \times m \times r$, e.g., an RGB image has $r = 3$. The convolutional layer can differ in every core it has; this is because it will have k kernels or filters of size $n \times n \times q$, where n is much smaller than the size of the image and q could be smaller than the number of channels. Figure 3 shows the general topology of a CNN.

4. Experimental Result and Discussion

This chapter shows that the outcomes are derived from the simulation using MATLAB 2020a. The recognition system consists of three stages. The first is *the feature extraction*; in this stage, we used Gabor wavelet transform. The second is *feature selection*. In this stage, we used particle swarm

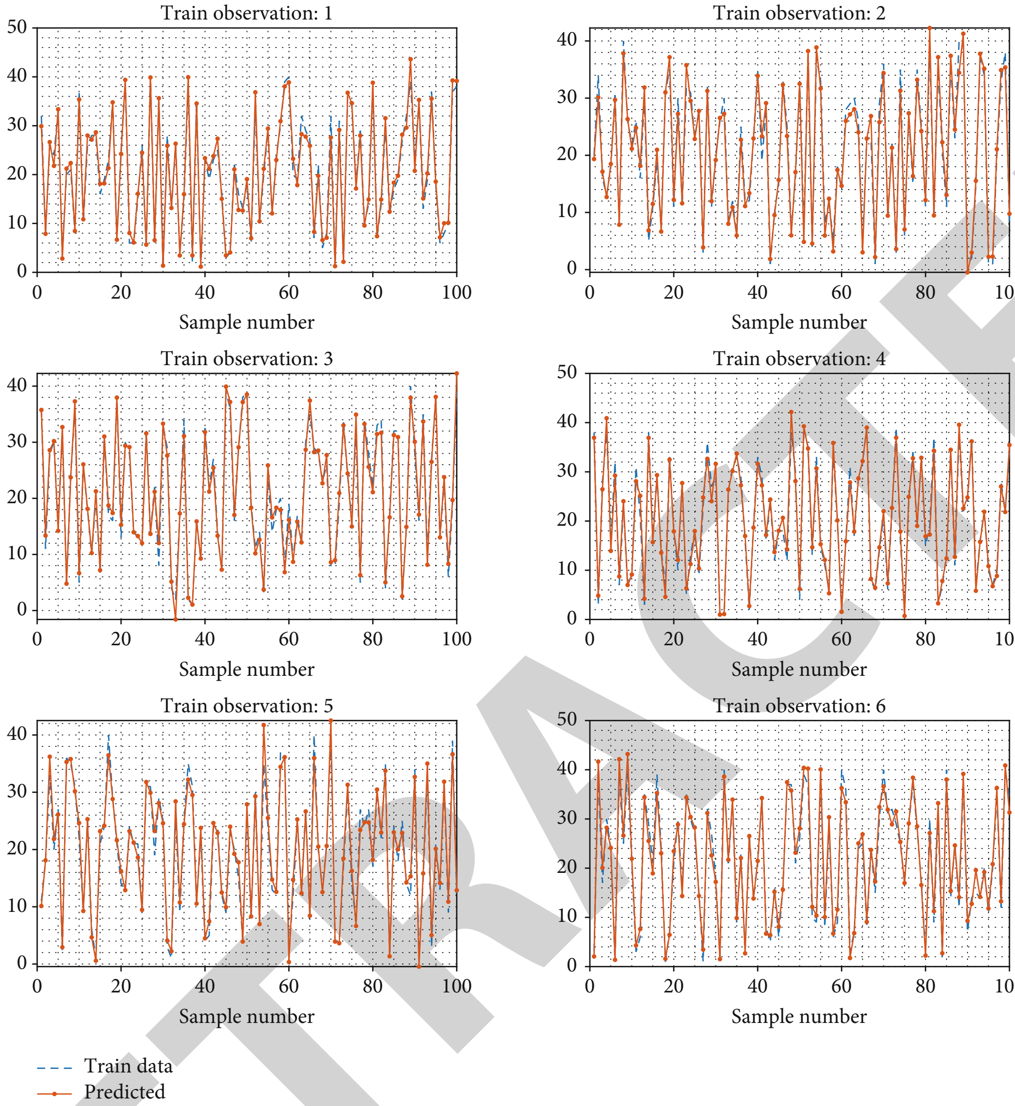


Figure 6: Train result.

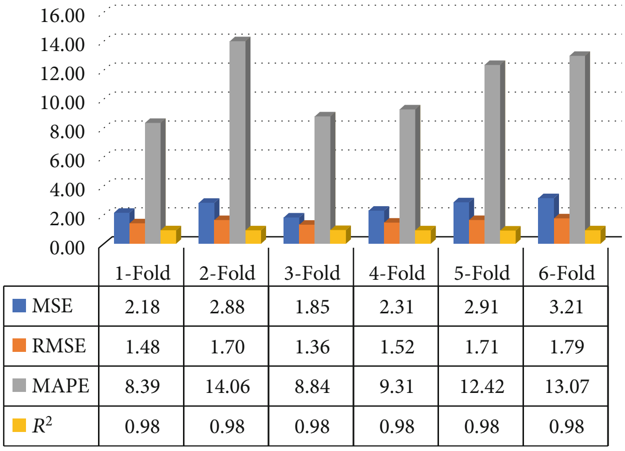
optimization (PSO), on features that are obtained from Gabor wavelet transform. In the final stage, *the classification*, we used deep learning with 6 layers.

The database in this study is used from ORL databases. The ORL (Olivetti Research Laboratory) face database contains 400 images of 40 different people. There are ten different grayscale images of each of 40 distinct persons. Images were captured at various times, and they have various variations including various expressions (closed/open eyes, not smiling/smiling). The details of the face (with/-without glasses) are included. Images were taken with a tolerance for some tilting and rotation of the face up to 20 degrees [49].

Some face images from the ORL database are shown in Figure 4.

Some simulation of the first face image is implemented on MATLAB 2020a, and the results are shown in Figure 5.

For evaluating the proposed method, we used the mean squared error (MSE), mean absolute percentage error



	1-Fold	2-Fold	3-Fold	4-Fold	5-Fold	6-Fold
🟦 MSE	2.18	2.88	1.85	2.31	2.91	3.21
🟧 RMSE	1.48	1.70	1.36	1.52	1.71	1.79
⬜ MAPE	8.39	14.06	8.84	9.31	12.42	13.07
🟨 R^2	0.98	0.98	0.98	0.98	0.98	0.98

Figure 7: The comparison of the MSE, RMSE, MAPE, and R for train data.

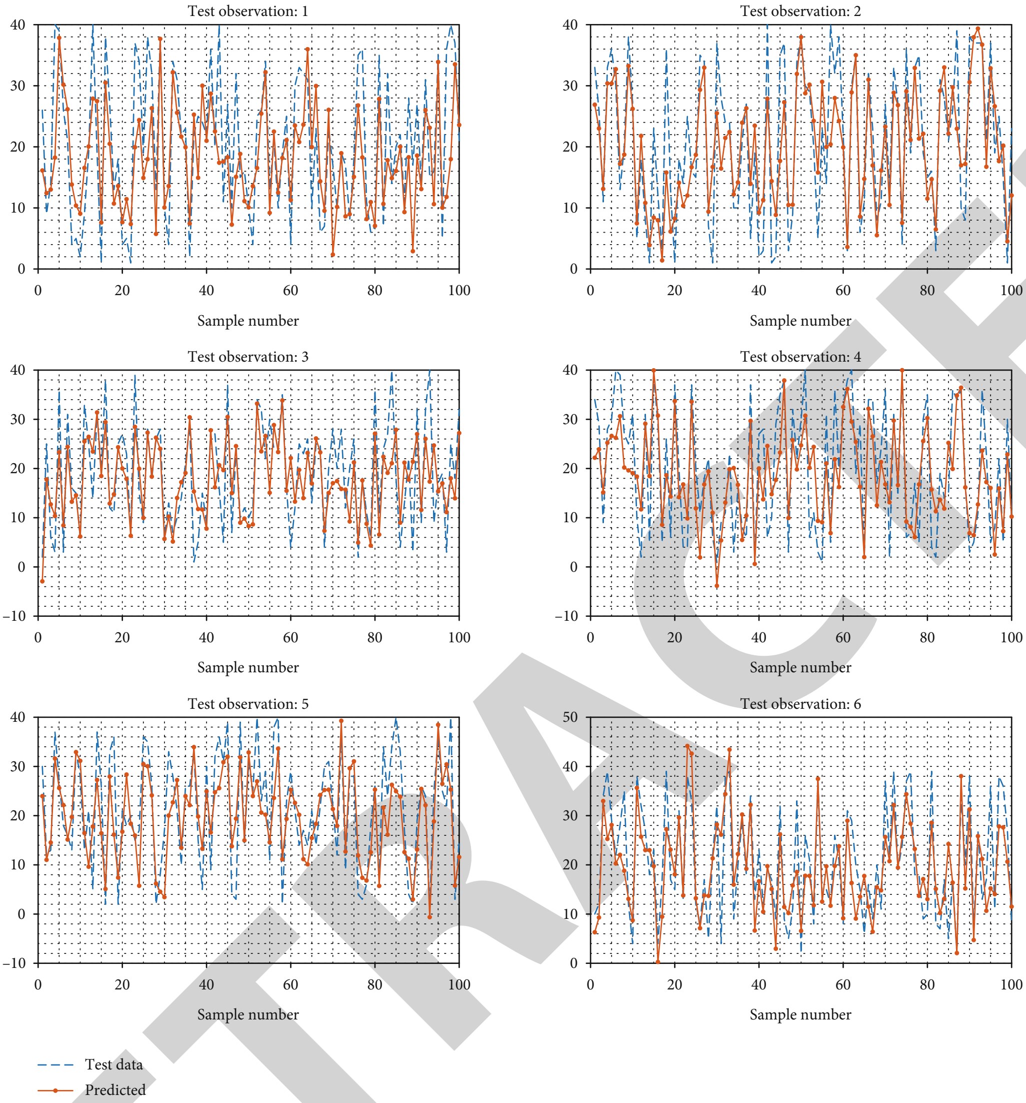


Figure 8: Test result.

(MAPE), and R-squared method. The mean squared error (MSE) is shown by

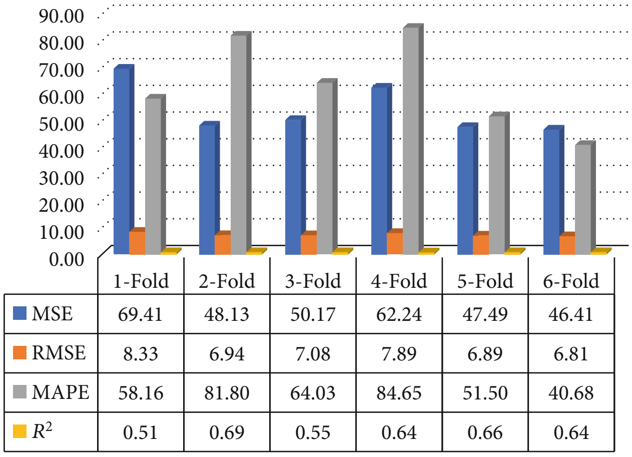
$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2. \tag{8}$$

The mean absolute percentage error (MAPE) is shown in the following equation:

$$\mathrm{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|. \tag{9}$$

For R square, we have the estimated value as

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i. \tag{10}$$



	1-Fold	2-Fold	3-Fold	4-Fold	5-Fold	6-Fold
MSE	69.41	48.13	50.17	62.24	47.49	46.41
RMSE	8.33	6.94	7.08	7.89	6.89	6.81
MAPE	58.16	81.80	64.03	84.65	51.50	40.68
R^2	0.51	0.69	0.55	0.64	0.66	0.64

Figure 9: The comparison of the MSE, RMSE, MAPE, and R for test data.

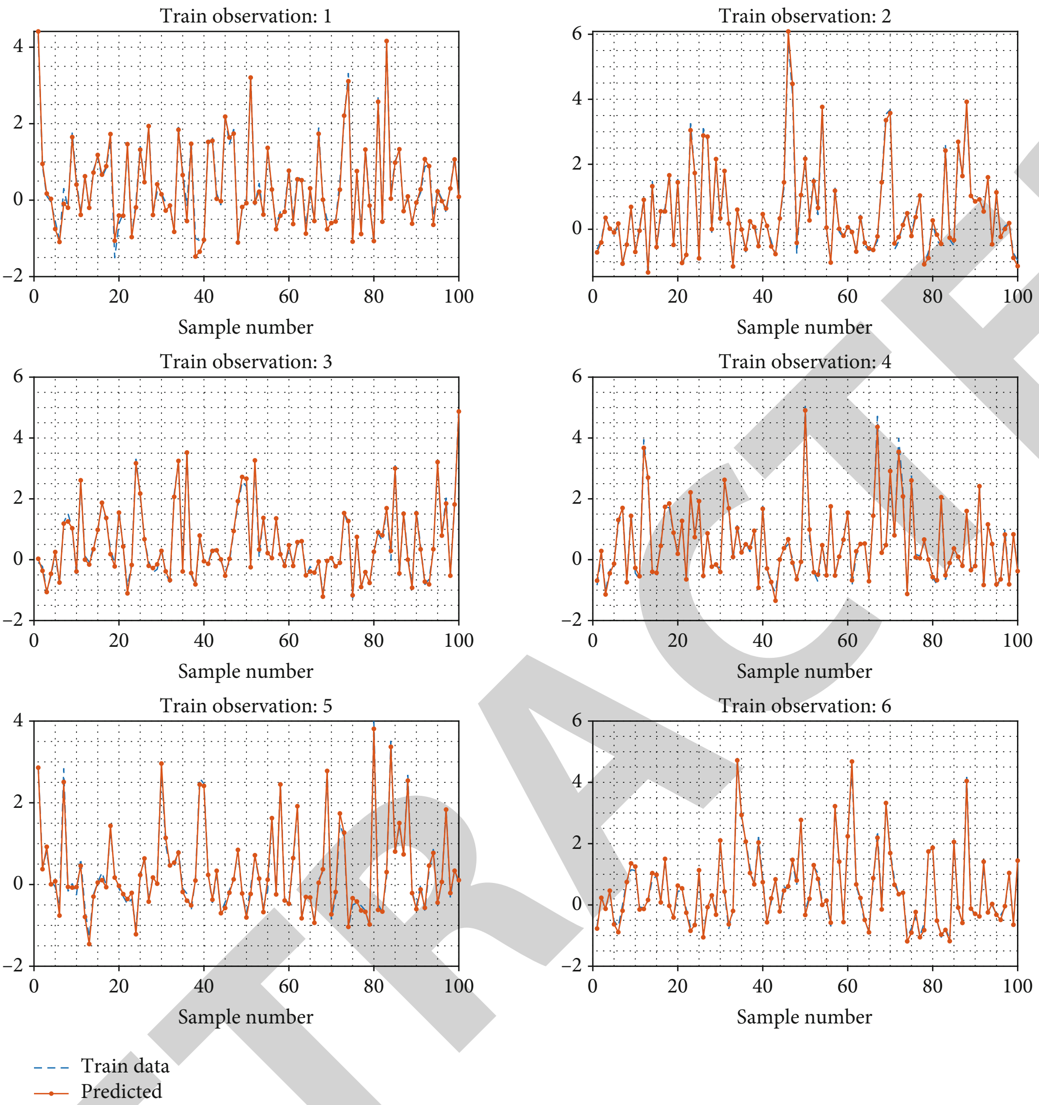


Figure 10: Train result with PSO.

Then, the variability of the data set can be measured using three sums of squares formulas. The total sum of squares is proportional to the variance of the data:

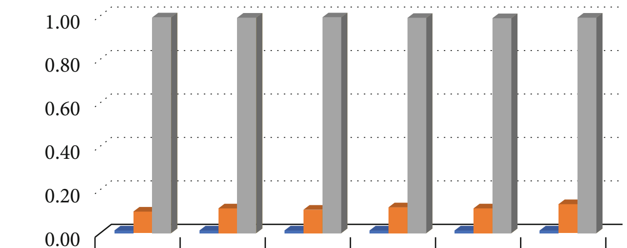
$$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2. \tag{11}$$

The regression sum of squares is also called the explained sum of squares:

$$SS_{\text{reg}} = \sum_i (f_i - \bar{y})^2. \tag{12}$$

The sum of squares of residuals is the residual sum of squares:

$$SS_{\text{res}} = \sum_i (y_i - f_i)^2 = \sum_i e_i^2. \tag{13}$$



	1-Fold	2-Fold	3-Fold	4-Fold	5-Fold	6-Fold
🟦 MSE	0.01	0.01	0.01	0.01	0.01	0.02
🟧 RMSE	0.10	0.11	0.10	0.12	0.11	0.13
⬜ R^2	0.99	0.99	0.99	0.99	0.99	0.99

Figure 11: The comparison of the MSE, RMSE, and R for train data.

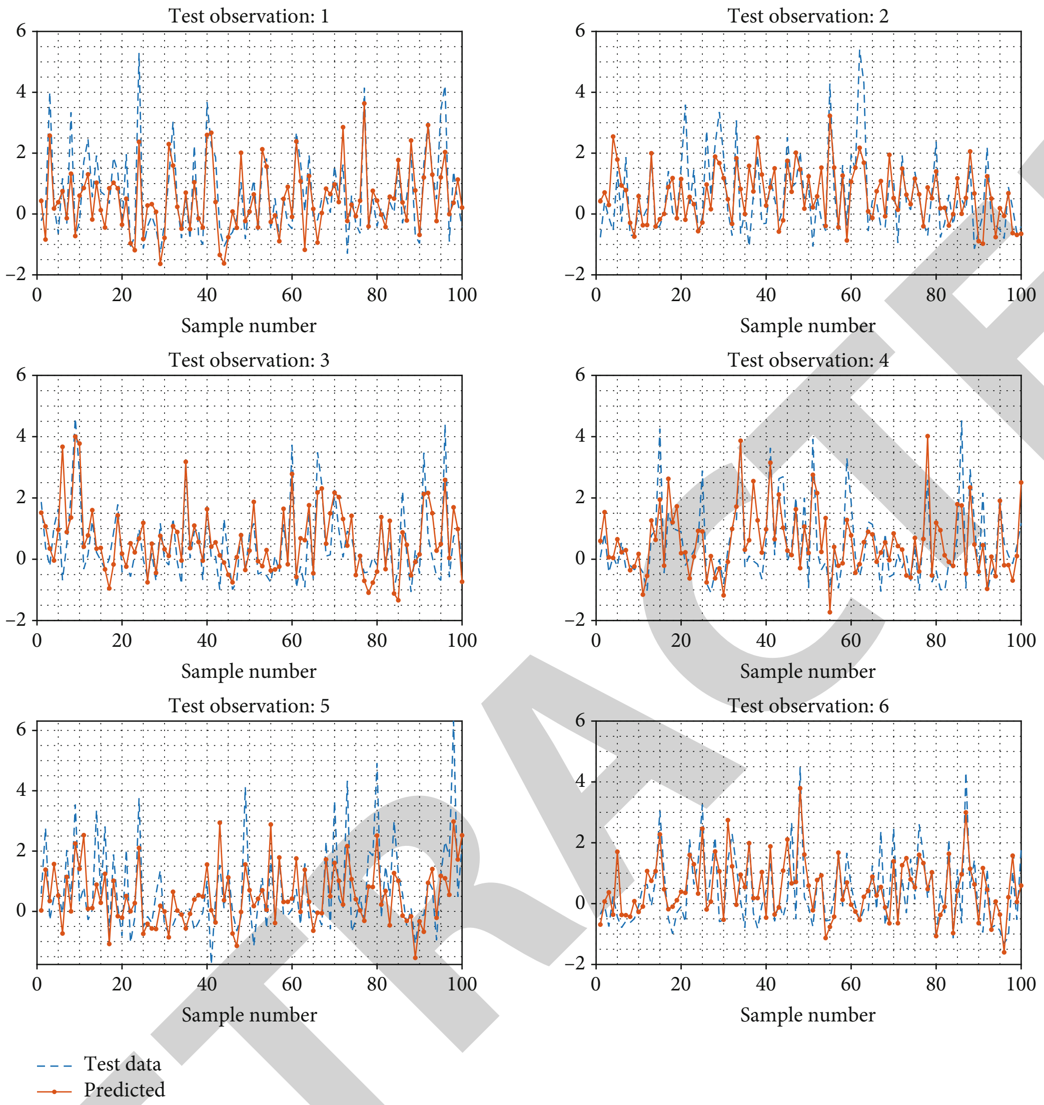
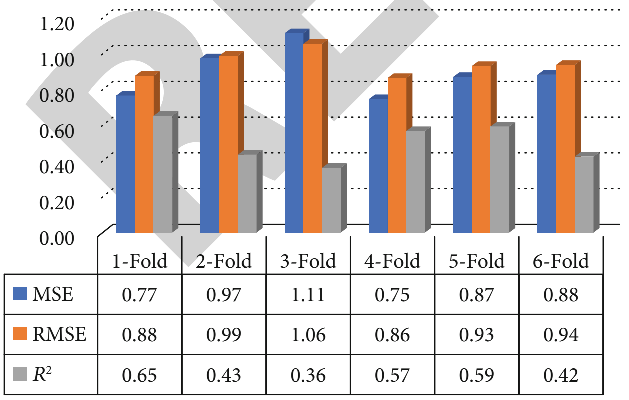


Figure 12: Test result with PSO.



	1-Fold	2-Fold	3-Fold	4-Fold	5-Fold	6-Fold
🟦 MSE	0.77	0.97	1.11	0.75	0.87	0.88
🟧 RMSE	0.88	0.99	1.06	0.86	0.93	0.94
⬜ R^2	0.65	0.43	0.36	0.57	0.59	0.42

Figure 13: The comparison of the MSE, RMSE, and R for train data.

Table 3: Comparing other methods with the proposed methods.

Method	Recognition rate
PCA [22]	53.2%
SRC [56]	75.12%
CRC [69]	79.4%
Gabor wavelet with Euclidian method [57]	83.44%
Symmetrical face sample method [49]	81.43%
Proposed method	85.25%

The most general definition of the coefficient of determination is

$$R^2 \equiv 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}. \qquad (14)$$

Table 2 shows the specifications of the layer that is used in deep learning.

The procedure and test were performed using actual with symmetrical species from ORL and YALE datasets. The results are shown in Figures 6–9. The results show that the system using the ORL dataset revealed how the preprocessing stage improves the accuracy. They also indicate how we can merge or fuse two methods of feature extraction to produce a powerful third method that can accomplish the job.

The comparison of the MSE, RMSE, MAPE, and R for train data is shown in Figure 7.

The result using the PSO is shown in Figures 10–13.

We have observed that the recognition rate and accuracy results from the experiments cannot be met when utilizing the Gabor wavelet and deep learning due to some variation of the values of features which corrupts the classification step. So, when compared with Gabor wavelet features, the variety will be large. Hence, the features are between -14 and 254. Therefore, optimum features are chosen.

PSO methods try to address this problem by selecting only the optimum features from Gabor wavelet. The performance of the classifier is based on the number of features. Too less or too redundant features can reduce the accuracy rates. Therefore, the number of features must be chosen carefully. In PSO, the basic process is that there are a number of particles; each one of them is flying through the problem area arbitrarily searching for the previous best solution and the global best solution of the whole swarm. Then, velocity is modified at each iteration which will define the movement of the particles to be more or less random. Therefore, the algorithms are converged. This method was used in literature [68], using the PSO method for selecting the best features.

In our experiments, we have used Gabor wavelet for feature extraction obtaining 10304 features. When the features were extracted, the implementation of PSO reduces the features to 5142. The best and most optimum features are selected by eliminating the highest and lowest values of features using the fitness function which determines the features that are the closest to each other in the amount. The experimental results obtained a 96% recognition rate on the ORL database when implementing the proposed method.

For the cause of completeness, we compare the performance of PCA [22], SRC [56], CRC [69], Gabor wavelet with Euclidian method [57], symmetrical face sample method [49], and the proposed method.

The comparison of other methods with the proposed methods is shown in Table 3.

5. Conclusion

The use of the symmetry property of the face is an efficient way to increase the performances of the face recognition systems. In this study, a new method is provided for the face recognition system. The new method is upgraded to use the benefits of symmetry property in the face data. The feature space is another way to implement the use of symmetry property in the face. There are many methods for feature extraction; however, none of them can handle the symmetry procedure in the feature space. The suggested methods can perform the symmetry procedure either in the image space or in the feature space. The introduced method is examined and tested for face recognition using data from ORL and YALE datasets.

Data Availability

All data available for readers are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] J. Bouguila and H. Khochtali, "Facial plastic surgery and face recognition algorithms: interaction and challenges. A scoping review and future directions," *Journal of Stomatology, Oral and Maxillofacial Surgery*, vol. 121, no. 6, pp. 696–703, 2020.

[2] W. W. Bledsoe, "Semiautomatic facial recognition," Technical Report SRI Project 6693, 1968.

[3] M. Yazdi, S. Mardani-Samani, M. Bordbar, and R. Mobaraki, "Age classification based on RBF neural network," *Canadian Journal on Image Processing and Computer Vision*, vol. 3, no. 2, pp. 38–42, 2012.

[4] W.-B. Horng, C.-P. Lee, and C.-W. Chen, "Classification of age groups based on facial features," *Journal of Applied Science and Engineering*, vol. 4, no. 3, pp. 183–192, 2001.

[5] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face recognition with local binary patterns," *European conference on computer vision*, pp. 469–481, 2004.

[6] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 915–928, 2007.

[7] M. Chandra Mohan, V. Vijaya Kumar, and A. Damodaram, "Novel Method of Adulthood Classification Based on Geometrical Features of Face," *GVIP Journal of Graphics, Vision and Image Processing*, 2010.

[8] V. V. Kumar, G. S. Murty, and P. S. Kumar, "Classification of facial expressions based on transitions derived from third order neighborhood LBP," *Global Journal of Computer Science and Technology*, vol. 14, no. 1-F, 2014.

[9] K. L. Kroeker, "Face recognition breakthrough," *Communications of the ACM*, vol. 52, no. 8, pp. 18-19, 2009.

[10] Ming Zhang and J. Fulcher, "Face recognition using artificial neural network group-based adaptive tolerance (GAT) trees," *IEEE Transactions on Neural Networks*, vol. 7, no. 3, pp. 555–567, 1996.

[11] X. Feng, M. Pietikainen, and A. Hadid, "Facial expression recognition with local binary patterns and linear programming," *Pattern Recognition And Image Analysis C/C of Raspoznavaniye Obrazov I Analiz Izobrazhenii*, vol. 15, no. 2, p. 546, 2005.

[12] M. Elad, R. Goldenberg, and R. Kimmel, "Low bit-rate compression of facial images," *IEEE Transactions on Image Processing*, vol. 16, no. 9, pp. 2379–2383, 2007.

[13] A. Skodras, C. Christopoulos, and T. Ebrahimi, "The jpeg 2000 still image compression standard," *IEEE Signal Processing Magazine*, vol. 18, no. 5, pp. 36–58, 2001.

[14] S. Rakshit and D. M. Monro, "An evaluation of image sampling and compression for human iris recognition," *IEEE Transactions on Information Forensics and Security*, vol. 2, no. 3, pp. 605–612, 2007.

[15] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, “Face recognition using LDA-based algorithms,” *IEEE Transactions on Neural Networks*, vol. 14, no. 1, pp. 195–200, 2003.

[16] Z. Lu and Linghua Zhang, “Face recognition algorithm based on discriminative dictionary learning and sparse representation,” *Neurocomputing*, vol. 174, pp. 749–755, 2016.

[17] J. Chen, V. M. Patel, L. Liu et al., “Robust local features for remote face recognition,” *Image and Vision Computing*, vol. 64, pp. 34–46, 2017.

[18] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, “Face recognition: a literature survey,” *ACM Computing Surveys (CSUR)*, vol. 35, no. 4, pp. 399–458, 2003.

[19] L. Wiskott, J.-M. Fellous, N. Kuiger, and C. Von Der Malsburg, “Face recognition by elastic bunch graph matching,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 775–779, 1997.

[20] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, “From few to many: illumination cone models for face recognition under variable lighting and pose,” *Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 643–660, 2001.

[21] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, “Eigenfaces vs. fisherfaces: recognition using class specific linear projection,” *Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[22] M. A. Turk and A. P. Pentland, “Face recognition using eigenfaces,” in *Proceedings 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 586–591, Maui, HI, USA, 1991.

[23] M. Yang and L. Zhang, “Gabor feature based sparse representation for face recognition with Gabor occlusion dictionary,” in *Computer Vision–ECCV 2010*, Lecture Notes in Computer Science, pp. 448–461, 2010.

[24] G. Guo and N. Zhang, “A survey on deep learning based face recognition,” *Computer Vision and Image Understanding*, vol. 189, article 102805, 2019.

[25] F. V. Massoli, G. Amato, and F. Falchi, “Cross-resolution learning for face recognition,” *Image and Vision Computing*, vol. 99, article 103927, 2020.

[26] M. Iqbal, M. S. I. Sameem, N. Naqvi, S. Kanwal, and Z. Ye, “A deep learning approach for face recognition based on angularly discriminative features,” *Pattern Recognition Letters*, vol. 128, pp. 414–419, 2019.

[27] C. Xu, Q. Liu, and M. Ye, “Age invariant face recognition and retrieval by coupled auto-encoder networks,” *Neurocomputing*, vol. 222, pp. 62–71, 2017.

[28] C.-K. Tran, C.-D. Tseng, and T.-F. Lee, “Improving the face recognition accuracy under varying illumination conditions for local binary patterns and local ternary patterns based on weber-face and singular value decomposition,” in *2016 3rd International Conference on Green Technology and Sustainable Development (GTSD)*, pp. 5–9, Kaohsiung, Taiwan, November 2016.

[29] S. Nikan and M. Ahmadi, “A modified technique for face recognition under degraded conditions,” *Journal of Visual Communication and Image Representation*, vol. 55, pp. 742–755, 2018.

[30] T.-K. Kim and J. Kittler, “Locally linear discriminant analysis for multimodally distributed classes for face recognition with a single model image,” *Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 318–327, 2005.

[31] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, “Face recognition using Laplacianfaces,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 328–340, 2005, http://ieeexplore.ieee.org/ielx5/34/30209/01388260.pdf?tp=&arnumber=1388260&isnumber=30209.

[32] A. Pentland, B. Moghaddam, and T. Starner, “View-based and modular eigenspaces for face recognition,” in *1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 84–91, Seattle, WA, USA, June 1994.

[33] R. Gross, I. Matthews, and S. Baker, “Eigen light-fields and face recognition across pose,” in *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition*, pp. 1–7, Washington, DC, USA, May 2002.

[34] C. Zhao, X. Li, and Y. Dong, “Learning blur invariant binary descriptor for face recognition,” *Neurocomputing*, vol. 404, pp. 34–40, 2020.

[35] J. Mairal, J. Ponce, G. Sapiro, A. Zisserman, and F. R. Bach, “Supervised dictionary learning,” *Advances in Neural Information Processing Systems*, pp. 1033–1040, 2009.

[36] M. A. Sunday, P. A. Patel, M. D. Dodd, and I. Gauthier, “Gender and hometown population density interact to predict face recognition ability,” *Vision Research*, vol. 163, pp. 14–23, 2019.

[37] S. Inamizu, E. Yamada, K. Ogata, T. Uehara, J.-i. Kira, and S. Tobimatsu, “Neuromagnetic correlates of hemispheric specialization for face and word recognition,” *Neuroscience Research*, vol. 156, pp. 108–116, 2019.

[38] S. K. Zhou and R. Chellappa, “Illuminating light field: image-based face recognition across illuminations and poses,” in *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings*, pp. 229–234, Seoul, Korea (South), May 2004.

[39] V. Blanz and T. Vetter, “Face recognition based on fitting a 3D morphable model,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1063–1074, 2003.

[40] L. Zhang and D. Samaras, “Face recognition from a single training image under arbitrary unknown lighting using spherical harmonics,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 3, pp. 351–363, 2006.

[41] V. Blanz, K. Scherbaum, T. Vetter, and H. P. Seidel, “Exchanging faces in images,” in *Computer Graphics Forum*, vol. 23, no. 3pp. 669–676, Wiley Online Library, 2004.

[42] J. Royer, C. Blais, I. Charbonneau et al., “Greater reliance on the eye region predicts better face recognition ability,” *Cognition*, vol. 181, pp. 12–20, 2018.

[43] M. Kas, Y. el merabet, Y. Ruichek, and R. Messoussi, “Mixed neighborhood topology cross decoded patterns for image-based face recognition,” *Expert Systems with Applications*, vol. 114, pp. 119–142, 2018.

[44] A. Shashua and T. Riklin-Raviv, “The quotient image: class-based re-rendering and recognition with varying illuminations,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 129–139, 2001.

[45] S. Zhou and R. Chellappa, “Rank constrained recognition under unknown illuminations,” in *IEEE International Workshop on Analysis and Modeling of Faces and Gestures, 2003. AMFG 2003*, pp. 11–18, Nice, France, October 2003.

[46] R. Basri and D. W. Jacobs, “Lambertian reflectance and linear subspaces,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 2, pp. 218–233, 2003.

[47] R. Ramamoorthi, “Analytic PCA construction for theoretical analysis of lighting variability in images of a Lambertian

Hindawi
BioMed Research International
Volume 2024, Article ID 9837415, 1 page
https://doi.org/10.1155/2024/9837415

Retraction

Retracted: Transcriptional Profiling Uncovers Biologically Significant RNAs and Regulatory Networks in Nucleus Pulposus from Intervertebral Disc Degeneration Patients

BioMed Research International

Received 12 March 2024; Accepted 12 March 2024; Published 20 March 2024

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope

(2) Discrepancies in the description of the research reported

(3) Discrepancies between the availability of data and the research described

(4) Inappropriate citations

(5) Incoherent, meaningless and/or irrelevant content included in the article

(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] Y. Chen, B. Cai, X. Lian, J. Xu, and T. Zhang, "Transcriptional Profiling Uncovers Biologically Significant RNAs and Regulatory Networks in Nucleus Pulposus from Intervertebral Disc Degeneration Patients," *BioMed Research International*, vol. 2021, Article ID 6696335, 33 pages, 2021.

Hindawi
BioMed Research International
Volume 2021, Article ID 6696335, 33 pages
https://doi.org/10.1155/2021/6696335

Research Article

Transcriptional Profiling Uncovers Biologically Significant RNAs and Regulatory Networks in Nucleus Pulposus from Intervertebral Disc Degeneration Patients

Yuanyuan Chen, Bin Cai, Xiaofeng Lian, Jianguang Xu, and Tao Zhang

Department of Orthopaedics, Shanghai Jiao Tong University Affiliated Sixth People's Hospital, Shanghai 200233, China

Correspondence should be addressed to Tao Zhang; zhangtao012345678@126.com

Received 20 October 2020; Revised 24 November 2020; Accepted 6 February 2021; Published 22 February 2021

Academic Editor: Parameshachari B. D.

Objective. This study aimed to uncover biologically significant RNAs in nucleus pulposus tissues of human intervertebral disc degeneration (IVDD) by integrated transcriptional profiling. *Methods*. From the Gene Expression Omnibus (GEO) database, three IVDD-related microarray profiling datasets were retrieved and assessed by intragroup data repeatability test. Then, differentially expressed circRNAs, lncRNAs, mRNAs, and miRNAs were screened in nucleus pulposus tissues between IVDD and control samples via the limma package. Coexpression networks were separately conducted via weighted gene correlation network analysis (WGCNA). Based on the feature RNAs in the IVDD-related modules, IVDD-related circRNA-miRNA-mRNA and lncRNA-miRNA-mRNA networks were conducted. The differentially expressed mRNAs in the two networks were analyzed by protein-protein interaction (PPI) and functional enrichment analyses. *Results*. By the intragroup data repeatability test, outlier samples were removed. Abnormally expressed RNAs were separately identified in nucleus pulposus between IVDD and controls. Via WGCNA, IVDD-related coexpression modules were constructed and the feature circRNAs, lncRNAs, mRNAs, and miRNAs were identified. Then, the circRNA- and lncRNA-miRNA-mRNA networks were built for IVDD. These mRNAs in the network exhibited complex interactions. Moreover, they were involved in distinct IVDD-related biological processes and pathways such as transcription, cell proliferation, TNF, TGF-β, and HIF signaling pathways. *Conclusion*. This study revealed biologically significant noncoding RNAs and their complex regulatory networks for IVDD.

1. Introduction

Intervertebral disc degeneration (IVDD) is a well-known cause of low back pain that is the main cause of disability [1]. IVDD is a multifactorial process, featured by changes in phenotype and genotype [2]. The nucleus pulposus is one of the main components of the intervertebral disc. Excessive apoptosis and extracellular matrix (ECM) degradation of nucleus pulposus cells (NPC) are the main pathological changes of IVDD [3]. The dysfunction of NPC facilitates the overproduction of proinflammatory factors in IVDD [4]. The inhibition of the abnormal activities of NPC can ameliorate the progression of IVDD [5]. Current treatments such as conservative treatment and surgery are mainly focused on pain relief, rather than treatment of the changes in pathology of degeneration [6]. Hence, a clear treatment strategy is urgently required. It is of importance to comprehensively comprehend the pathogenesis of IVDD, thereby accelerating the development of the therapeutic strategy concerning repairing the intervertebral disc injury.

Gene therapy has received extensive attention on account of its unique capacity to mediate cellular biological processes in IVDD, which provides an excellent synergistic treatment consequence by combining distinct genes [7]. Thus, it may be durable and beneficial to the therapy of IVDD. Targeting the specific noncoding RNAs (such as microRNAs (miRNAs), long noncoding RNA (lncRNA), and circular RNAs (circRNAs)) is a promising therapeutic strategy to lower toxicity and other harmful consequences [8]. miRNAs participate in various biological functions of IVDD cells by targeting different genes related to the IVDD process [9]. As a regulator of gene expression, miRNA has attracted widespread attention in the prevention and treatment of IVDD [10]. Different from linear RNAs, circRNAs with tissue specific and conservative features exhibit a closed circular structure, which are not affected by RNA exonuclease [11]. Thus, circRNAs are more stable and difficult to degrade. They are rich in binding sites of miRNA, as miRNA sponges to abrogate the inhibitory effect of miRNAs concerning their target genes during the progression of IVDD [12]. For instance, circERCC2 improves IVDD through mediating mitochondrial phagocytosis as well as apoptosis in the NPC via the miR-182-5p/SIRT1 axis [13]. lncRNA is abundantly expressed in human genomes, which may regulate key cell biological processes including IVDD cells [14]. lncRNAs can be used as a miRNA sponge to mediate the expression of their target genes in IVDD [15]. However, there is still a lack of systematic analysis of noncoding RNAs and their regulatory mechanisms in IVDD. Hence, this study probed into biologically significant RNAs and conducted their regulatory networks for IVDD, offering underlying therapeutic targets as well as molecular mechanisms concerning IVDD.

2. Materials and Methods

2.1. Microarray Datasets. IVDD-related microarray data were retrieved from the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/gds/), including GSE67566, GSE56081, and GSE63492 datasets. The GSE67566 dataset contained circRNA expression profiles in human nucleus pulposus tissues from 5 IVDD patients and 5 normal controls on the GPL19978 platform [16]. The GSE56081 dataset was composed of lncRNA and mRNA profiling of nucleus pulposus tissues from 5 IVDD patients and 5 controls on the GPL15314 platform. The GSE63492 dataset included miRNA expression profiling in nucleus pulposus specimens derived from 5 IVDD patients and 5 controls on the GPL19449 platform.

2.2. Data Preprocessing and the Intragroup Data Repeatability Test. The raw data were preprocessed via the Linear Models for Microarray Data (limma) package v3.34.0 in R [17], which were normalized by quantile normalization as well as log^2 conversion. When a gene corresponded to multiple probes, the average value was taken as the expression value of the gene. Intragroup data repeatability test was presented via principal component analysis (PCA) and Pearson's correlation analysis.

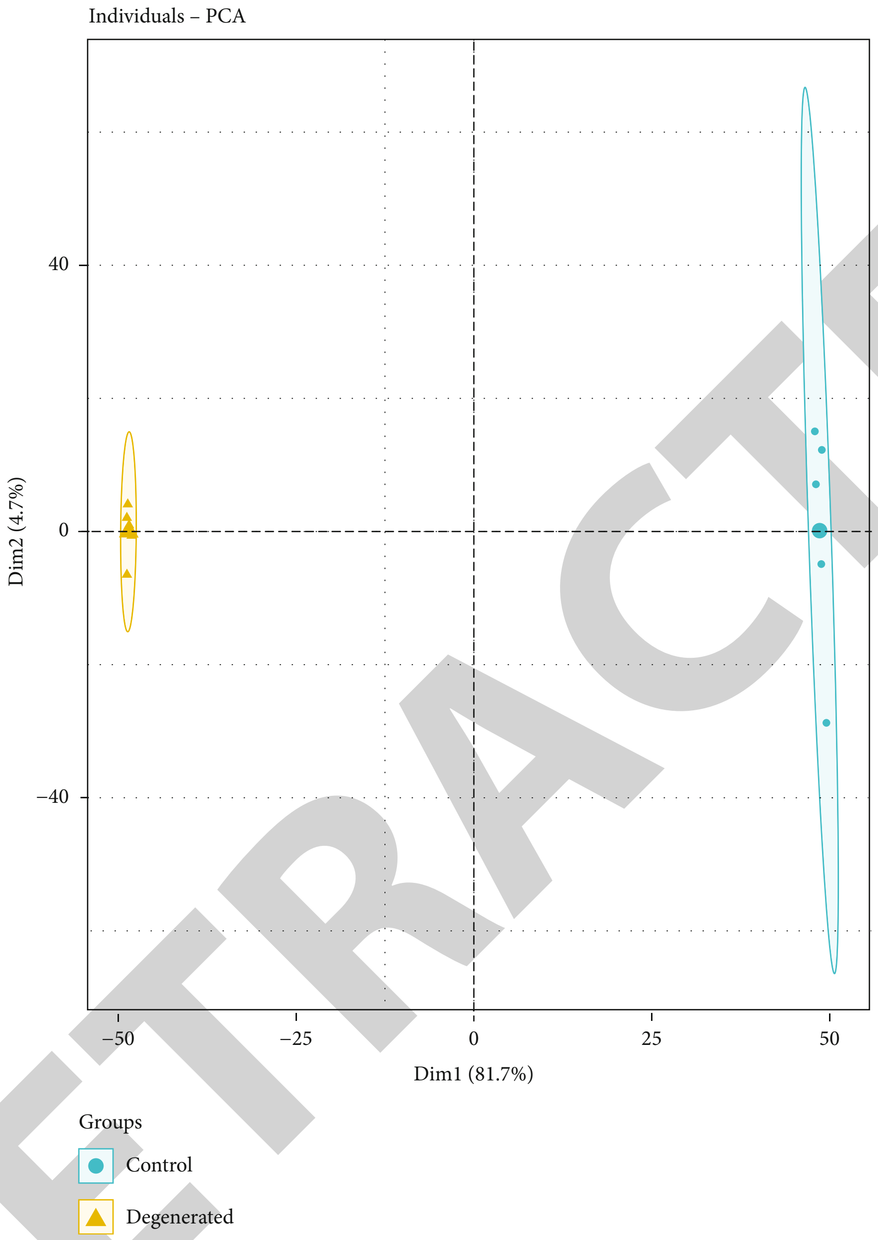
2.3. Differential Expression Analysis. Differentially expressed circRNAs (DEcircRNAs), lncRNAs (DElncRNAs), mRNAs (DEmRNAs), and miRNAs (DEmiRNAs) were screened by the limma package in R [17]. The criteria were set as adjusted $p < 0.05$ and $|\log^2 \text{fold change (FC)}| > 2$. The results were visualized into volcano plots and hierarchical clustering heat maps utilizing the limma or pheatmap packages in R. Furthermore, hierarchical clustering analysis was conducted on the basis of Euclidean distance.

2.4. Weighted Gene Correlation Network Analysis (WGCNA). Coexpression analysis was presented via the WGCNA package in R [18]. Although the data were subjected to batch effect removal and normalization conversion, the possibility of outliers cannot be ruled out. Outlier samples were prone to adversely affect the analysis results of the network module, so it was necessary to identify and remove outliers before constructing the network. Since the outliers could introduce excessive deviations to WGCNA, they were removed before WGCNA. The pickSoftThreshold function in the WGCNA package was used to establish the soft threshold β. The selected soft threshold β met the following conditions: (1) the degree of fitting between the constructed network and the scale-free network reached a high level; (2) a smaller β value was chosen when the degree of fit was up to a satisfactory level. The adjacency coefficient between gene i and gene j was calculated by exponential adjacency function a_{ij}, as follows: $a_{ij} = |\text{cor}\,(x_i, x_j)|^{\beta}$ (where x_i was the expression vector of gene x and a_{ij} was obtained by the β power of the Pearson correlation coefficient between x_i and x_j). The topological overlap matrix was then calculated. dissTOM was calculated as follows: dissTOM = 1 – TOM. The hClust function of the WGCNA package was applied to cluster the samples. The dynamic tree cutting method was utilized to identify and merge coexpressed gene modules with similar expression patterns. According to the genes in each module, module eigengene of each module was calculated, which was defined as the first principal component of the expression levels of all genes in the module, representing the overall level of gene expression in the module. Then, we calculated Pearson's correlation coefficient between the module and the traits of samples.

2.5. The Competing Endogenous RNA (ceRNA) Network. DEmiRNA-DEmRNA and DElncRNA-DEmiRNA relationships were predicted via the StarBase v2.0 database (http://starbase.sysu.edu.cn/) [19]. Based on circBase (http://www.circbase.org/cgi-bin/downloads.cgi) [20], DEcircRNA-DEmiRNA pairs were analyzed. After integration of DEcircRNA-DEmiRNA, DElncRNA-DEmiRNA, and DEmiRNA-DEmRNA pairs, IVDD-related circRNA-miRNA-mRNA and lncRNA-miRNA-mRNA networks were separately conducted via the Cytoscape v3.7.2 software (http://www.cytoscape.org/) [21].

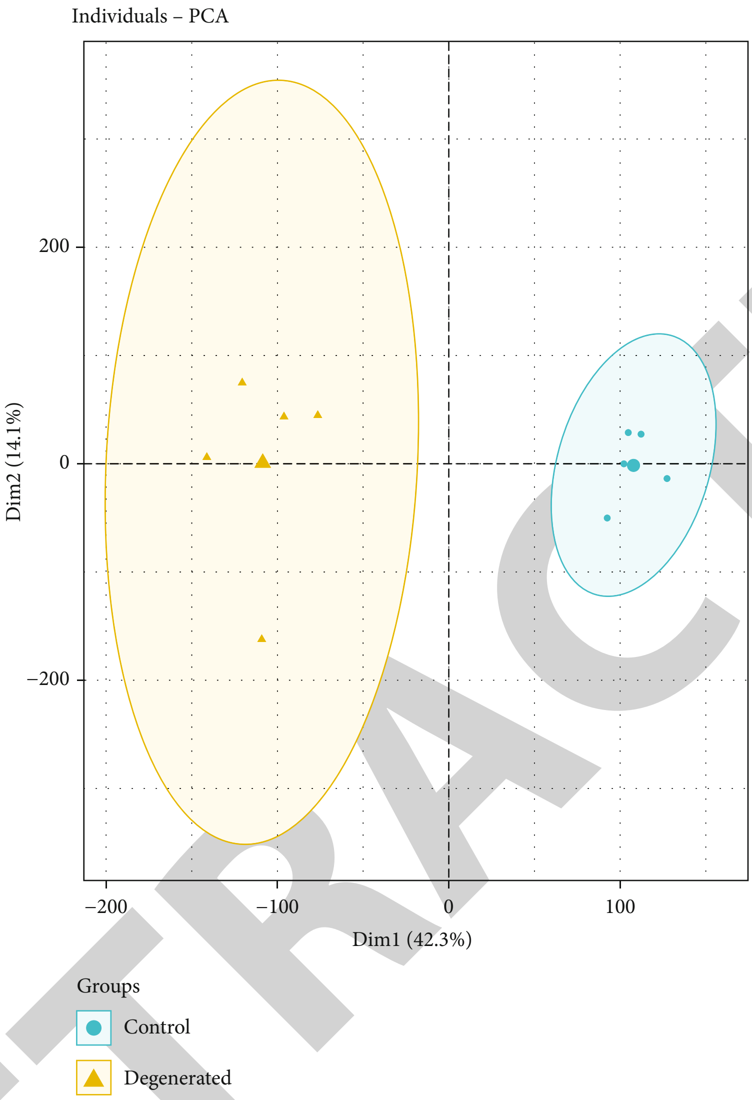
2.6. The Protein-Protein Interaction (PPI) Network. The Search Tool for the Retrieval of Interacting Genes (STRING) (string db.org) v10.0 database has been widely applied to analyze the interactions between proteins [22]. The mRNAs in the ceRNA networks were imported into the STRING database to study the interactions between the proteins encoded by genes. Their relationships were visualized by the Cytoscape v3.7.2 software (http://www.cytoscape.org/) [21].

2.7. Functional Annotation Analysis. Gene ontology (GO), covering biological process, cellular component, molecular function, and Kyoto Encyclopedia of Genes and Genomes (KEGG) were analyzed by using the Database for Annotation, Visualization, and Integrated Discovery (DAVID) online tool database v6.8 (http://david.abcc.ncifcrf.gov) [23]. $p < 0.05$ was significantly enriched.



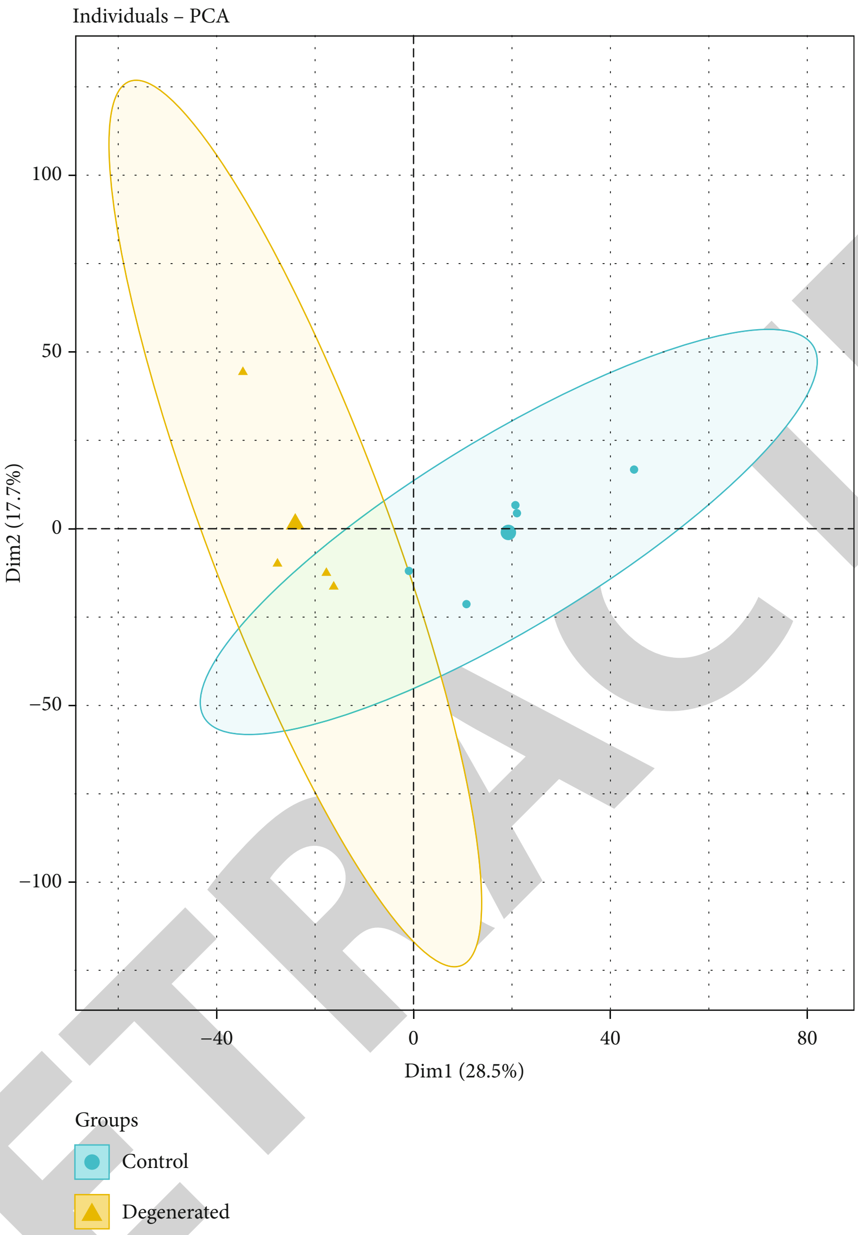
(a)

Figure 1: Continued.



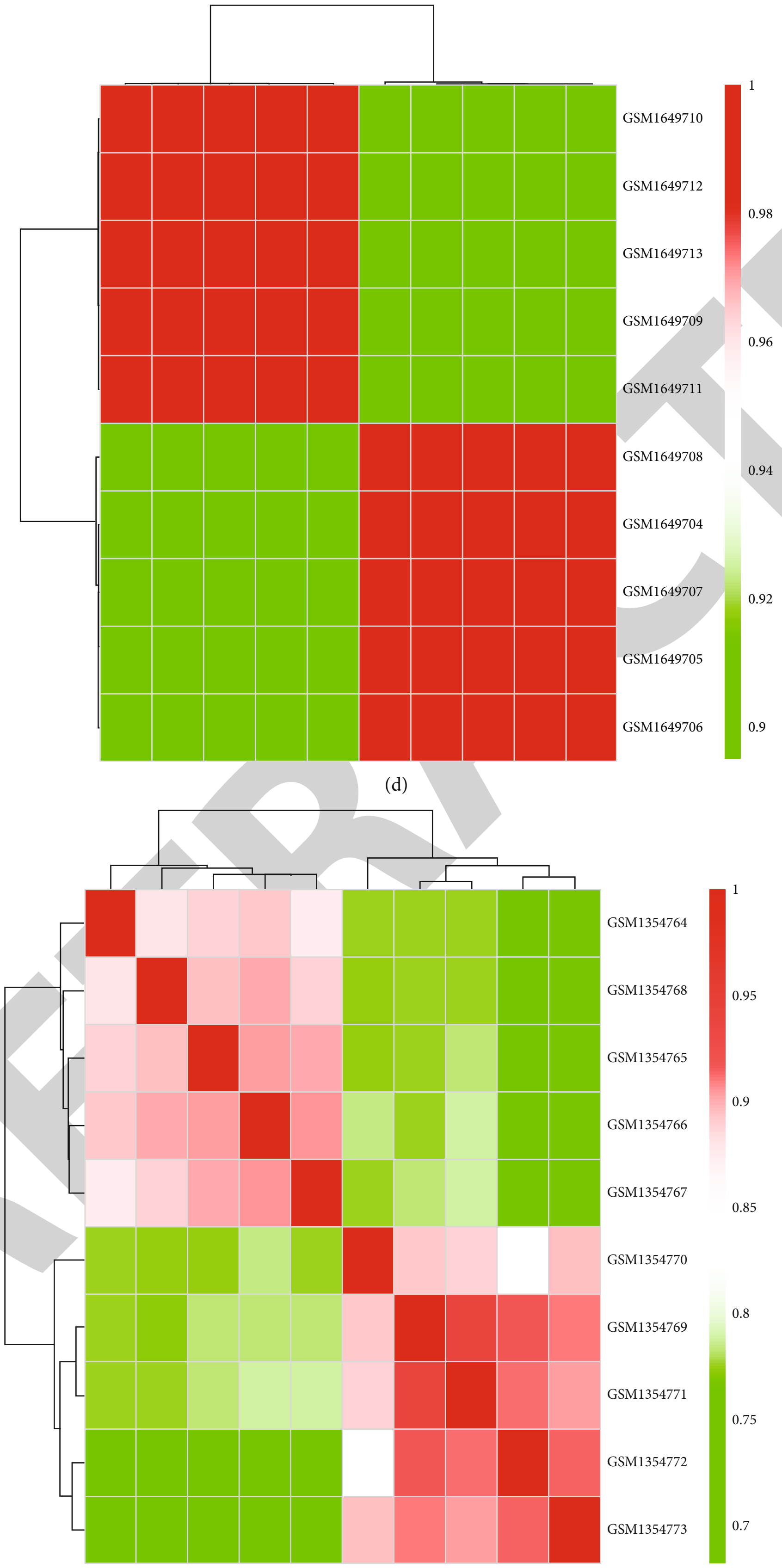
(b)

Figure 1: Continued.



(c)

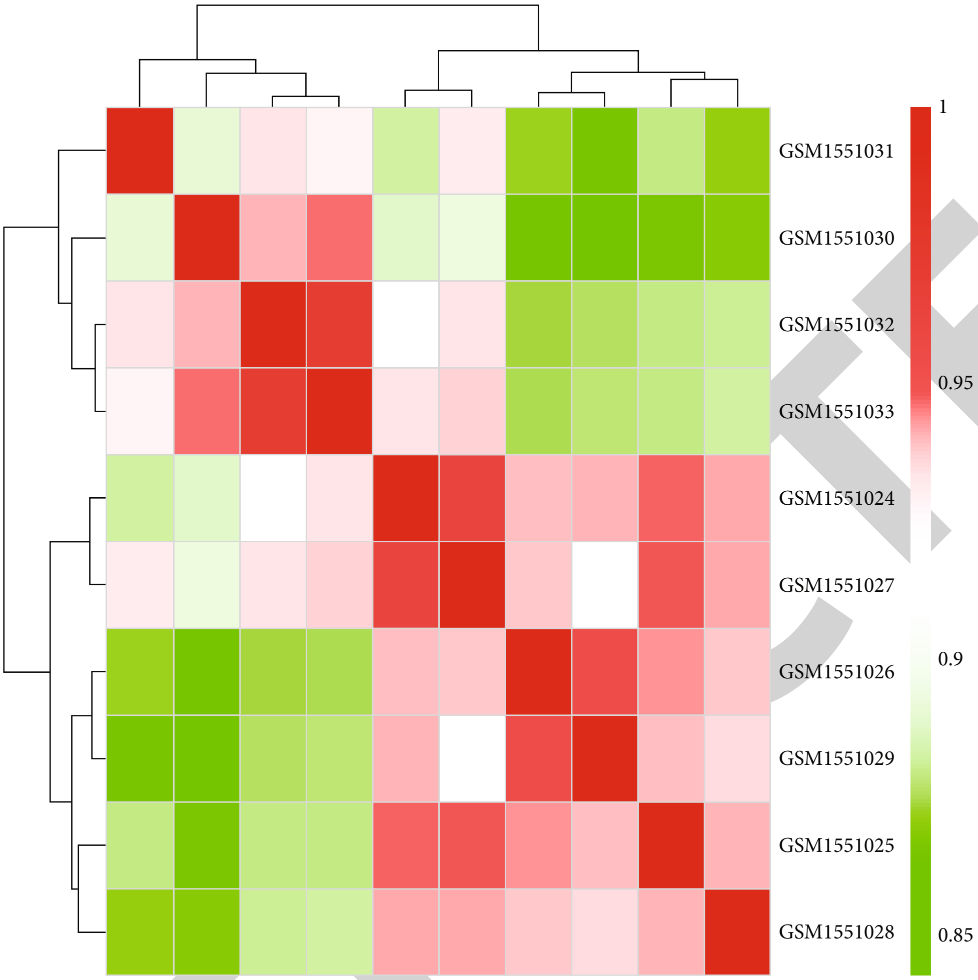
Figure 1: Continued.



(d)

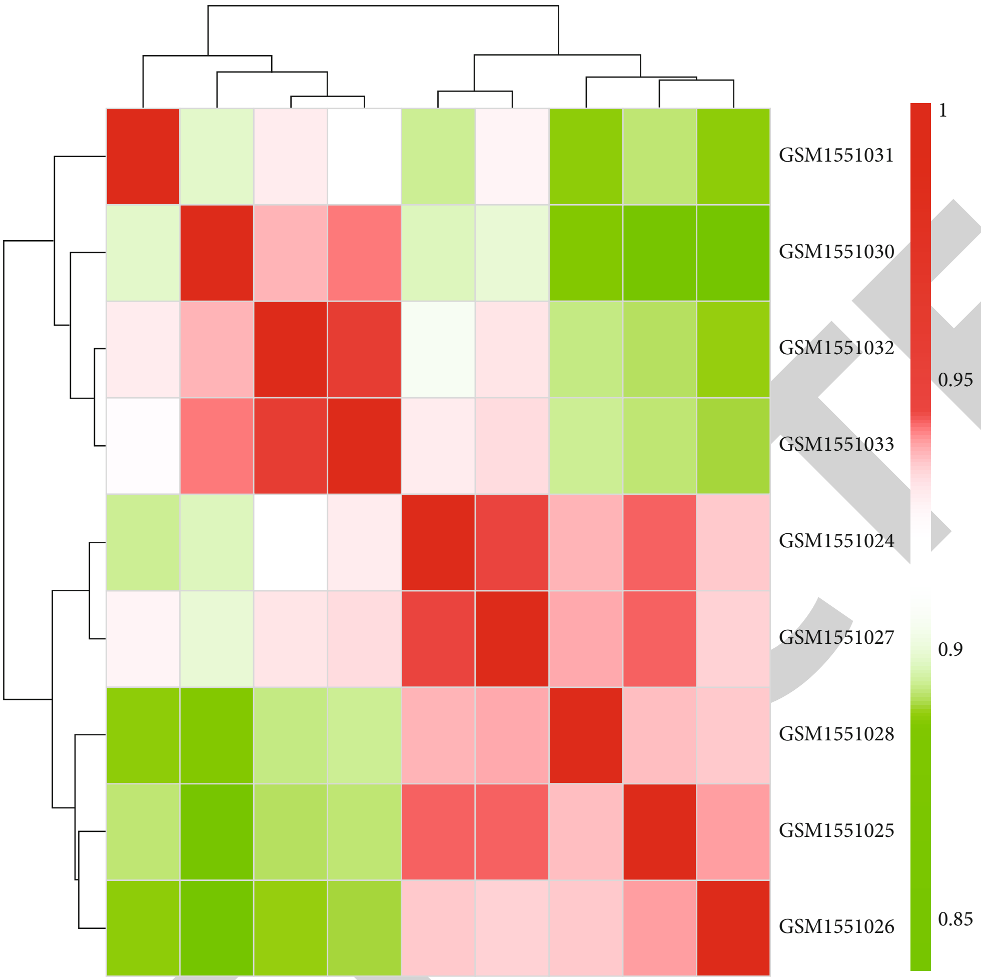
(e)

Figure 1: Continued.



(f)

Figure 1: Continued.



(g)

Figure 1: PCA and Pearson's correlation analysis. (a) PCA results for samples of circRNA expression profiles in the GSE67566 dataset. (b) PCA results for samples of mRNA and lncRNA expression profiles in the GSE56081 dataset. (c) PCA results for samples of miRNA expression profiles in the GSE63492 dataset. Blue is indicative of control samples and yellow is indicative of IVDD samples. Each point represents a sample. Sample correlation results in the (d) GSE67566, (e) GSE56081, and (f) GSE63492 datasets. (g) Sample correlation results in the GSE63492 dataset after excluding the GSM1551029 sample. The more the color tends to be redder, the stronger the correlation; on the contrary, the more the color tends to be greener, the weaker the correlation.

3. Results

3.1. The Intragroup Data Repeatability Test. This study retrieved the circRNA, lncRNA/mRNA, and miRNA expression datasets from human nucleus pulposus tissues separately derived from 5 IVDD patients and 5 normal control in the GEO database. To avoid the existence of abnormal samples from adversely affecting subsequent analysis, PCA was presented to visualize the clustering of samples of microarrays. The results showed that no abnormal samples were found in the GSE67566 (Figure 1(a)) and GSE56081 (Figure 1(b)). All samples from the two datasets were obviously divided into control and IVDD groups. In the GSE63492 dataset, there was one outlier sample (GSM1551029) that was removed (Figure 1(c)). Next, we performed correlation analysis on 10 samples in the three datasets. In the GSE67566 circRNA expression dataset, there were high consistencies between the samples in the IVDD group or the control group (Figure 1(d)). Furthermore, the correlation of samples between the IVDD group was higher than the correlation between the IVDD group and the control group in the GSE56081 lncRNA/mRNA dataset (Figure 1(e)). In Figure 1(f), we found that there was a case of an IVDD sample (GSM1551029) that was highly correlated with the samples in the control group from the GSE63492 miRNA expression dataset. Thus, the GSM1551029 sample was eliminated in the subsequent analysis. Figure 1(g) showed the correlation between the samples in the IVDD and control groups.

3.2. Identification of Differentially Expressed circRNAs, lncRNAs, mRNAs, and miRNAs for IVDD Nucleus Pulposus Tissues. Under the criteria of adjusted $p < 0.05$ and $|\log^2 \text{FC}| > 2$, we carried out differential expression analysis in nucleus pulposus tissues between IVDD and control groups. In Figure 2(a), there were 105 DEcircRNAs (47 up- and 58 down-regulated) between IVDD and control groups (Supplementary

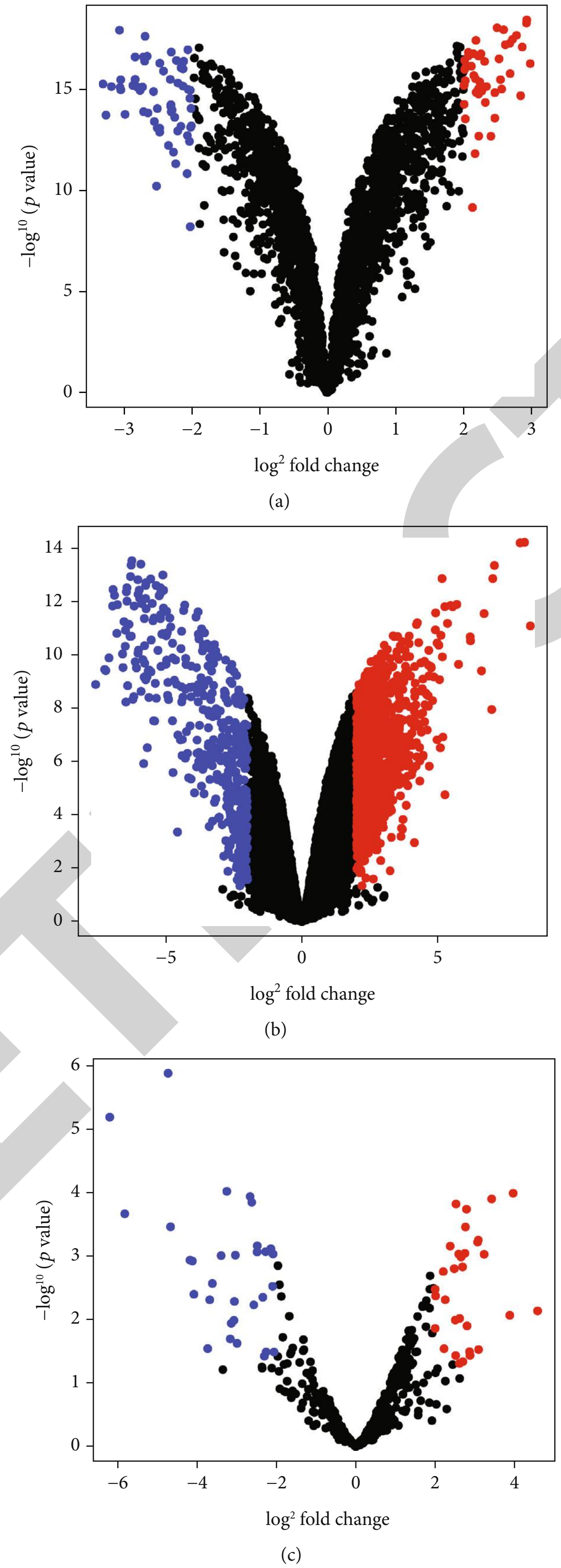
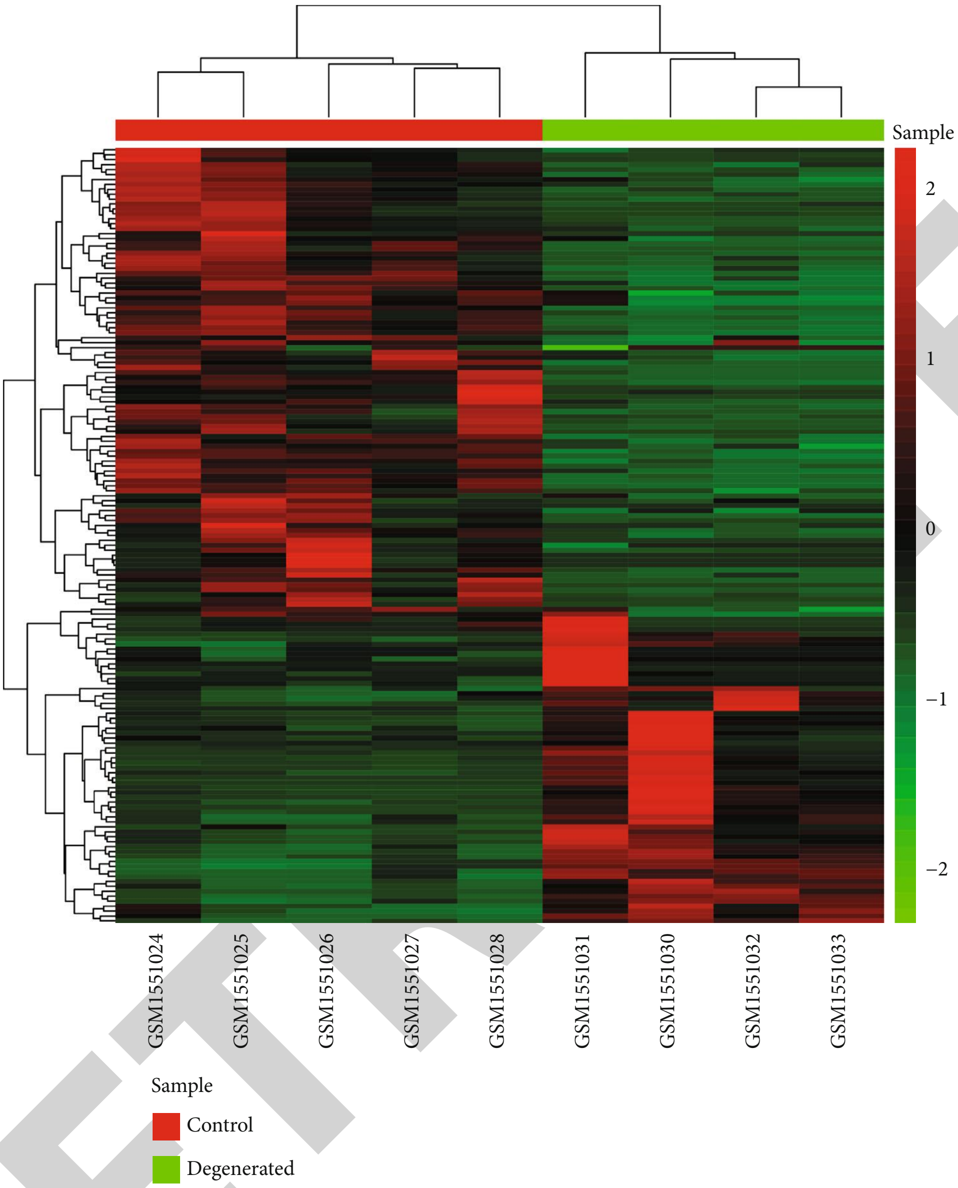
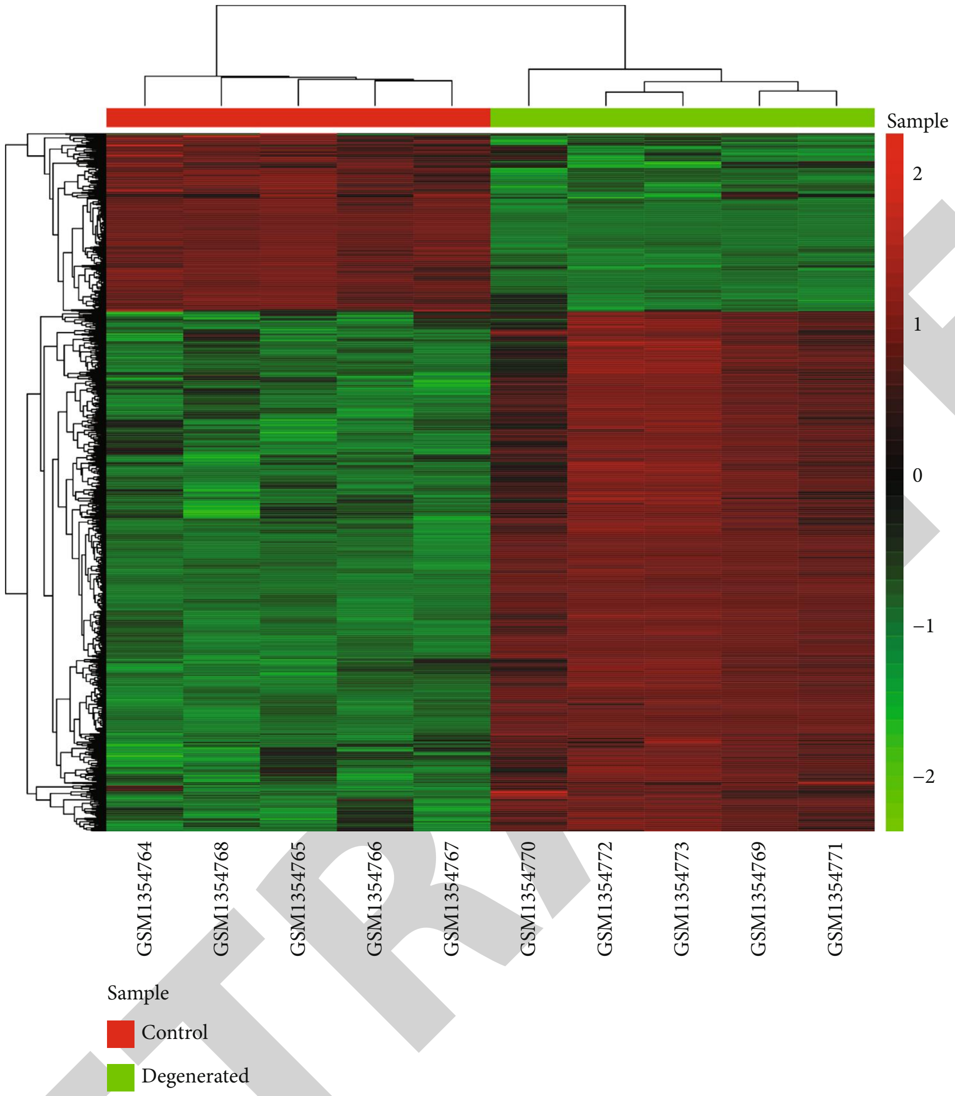


Figure 2: Continued.



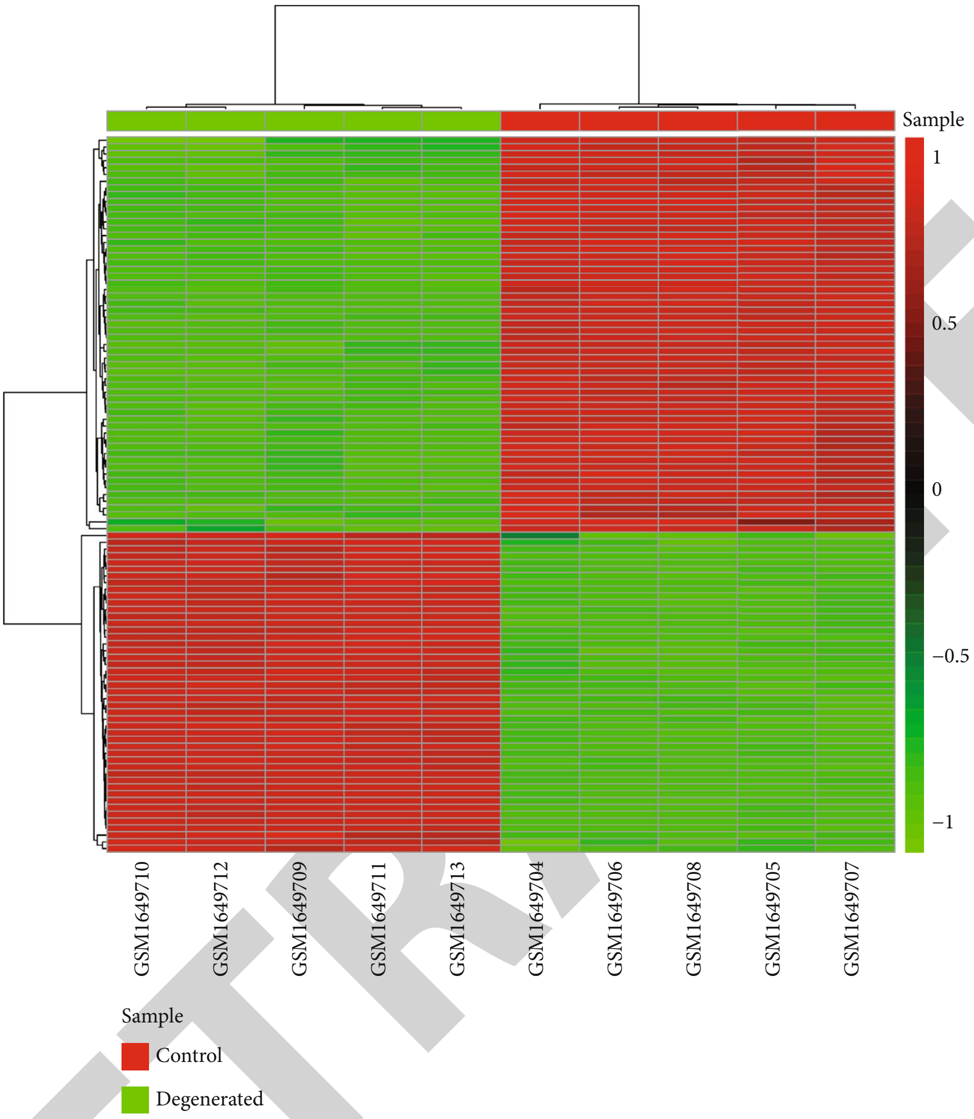
(d)

Figure 2: Continued.



(e)

Figure 2: Continued.

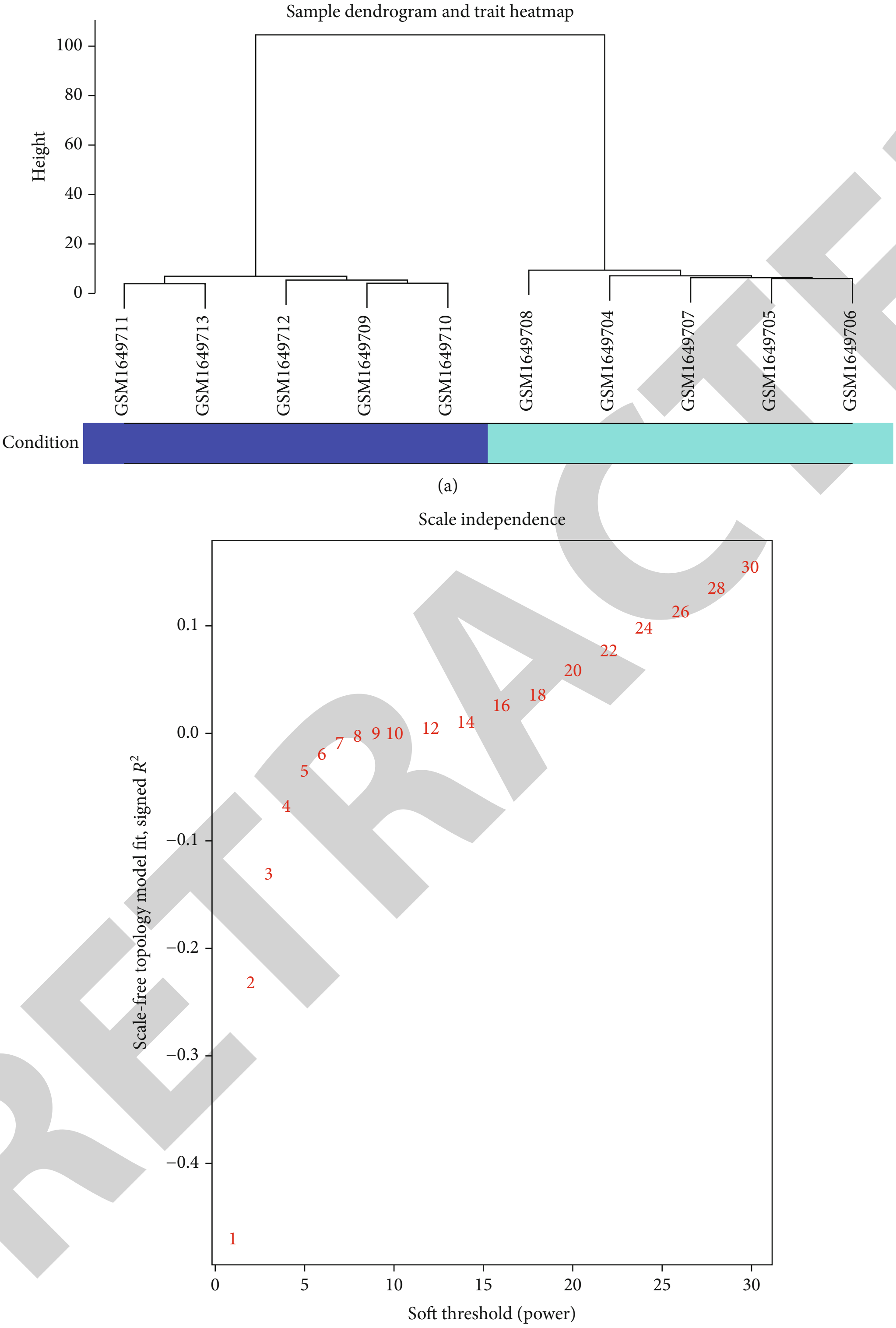


(f)

Figure 2: Identification of differentially expressed circRNAs, lncRNAs, mRNAs, and miRNAs for IVDD nucleus pulposus tissues. Volcano plots depicting all differentially expressed (a) circRNAs, (b) lncRNAs/mRNAs, and (c) miRNAs in nucleus pulposus tissues between IVDD and control groups. Blue is indicative of downregulation and red signifies upregulation in the IVDD group. Hierarchical clustering analysis results for differentially expressed (d) circRNAs, (e) lncRNAs/mRNAs, and (f) miRNAs between IVDD and control groups. Green is indicative of downregulation and red signifies upregulation in the IVDD group.

Table 1). Furthermore, 131 DElncRNAs (96 up- and 35 downregulated) and 928 DEmRNAs (687 up- and 241 downregulated) were identified in nucleus pulposus tissues between IVDD and control groups (Figure 2(b); Supplementary Table 2). 62 DEmiRNAs (31 up- and 31 downregulated) were screened in nucleus pulposus tissues between IVDD and control groups (Figure 2(c); Supplementary Table 3). Hierarchical clustering analysis revealed that DEcircRNAs (Figure 2(d)), DElncRNAs/mRNAs (Figure 2(e)), and DEmiRNAs (Figure 2(f)) could distinctly distinguish IVDD samples from control samples.

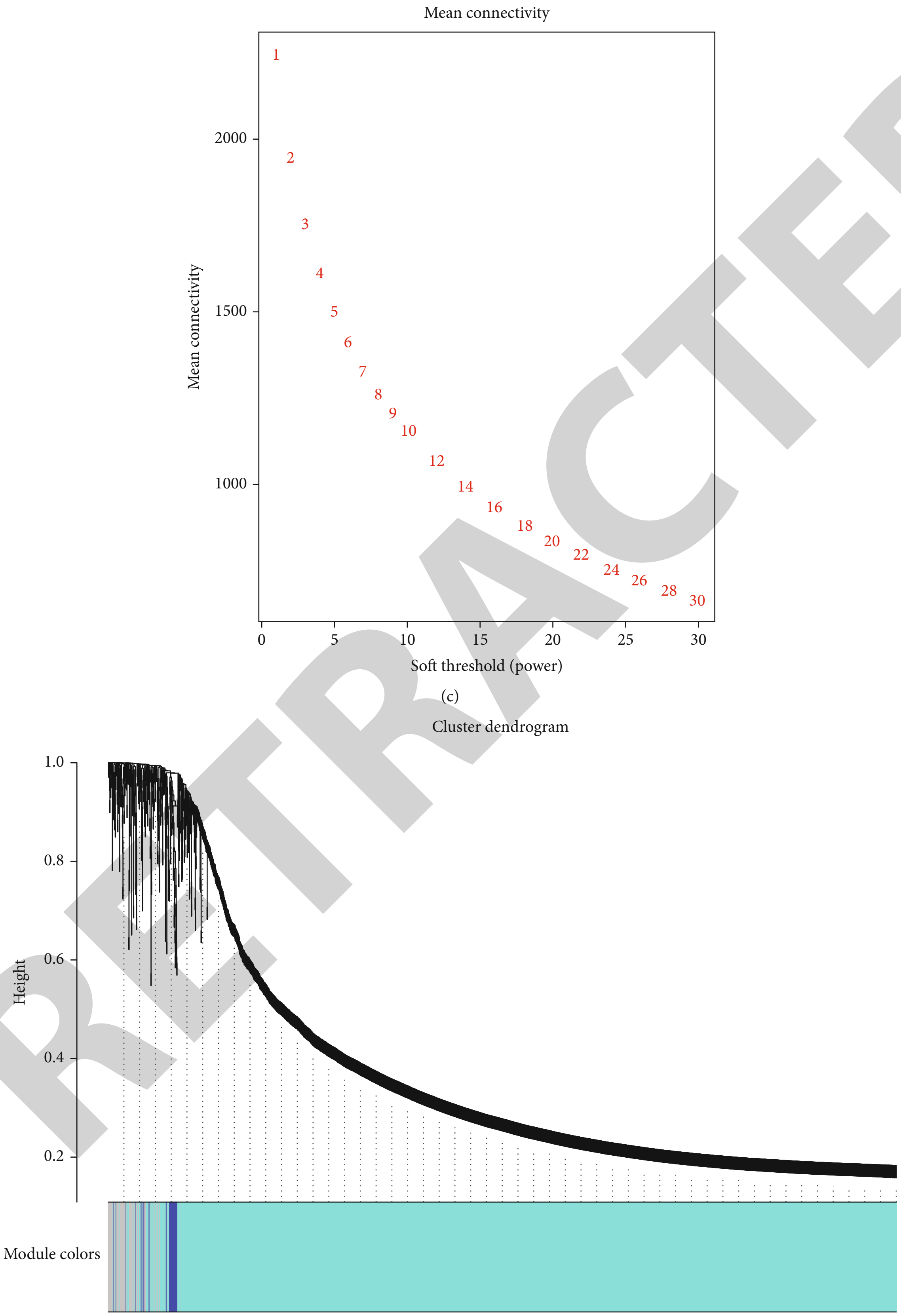
3.3. An IVDD-Related Coexpression Network in the GSE67566 Dataset. 5 IVDD nucleus pulposus samples and control samples in the GSE67566 dataset were used for WGCNA. In Figure 3(a), no outlier samples were detected among them, which was consistent with PCA results. Using the pickSoftThreshold function, we calculated the degree of scale-free topology model fit (Figure 3(b)) and mean connectivity (Figure 3(c)) when the value of β was from 1 to 30. Herein, $\beta = 18$ was chosen. Via the adjacency function, the adjacency coefficient between genes was calculated. TOM was calculated via the TOMsimilarity function, followed by the calculation of dissTOM. Based on the dissTOM, the hierarchical clustering graphs of genes were generated. In this study, the dynamic tree cutting method was utilized to identify and merge 3 coexpressed gene modules with similar expression patterns (Figure 3(d)). As shown in Figure 3(e),



(a)

(b)

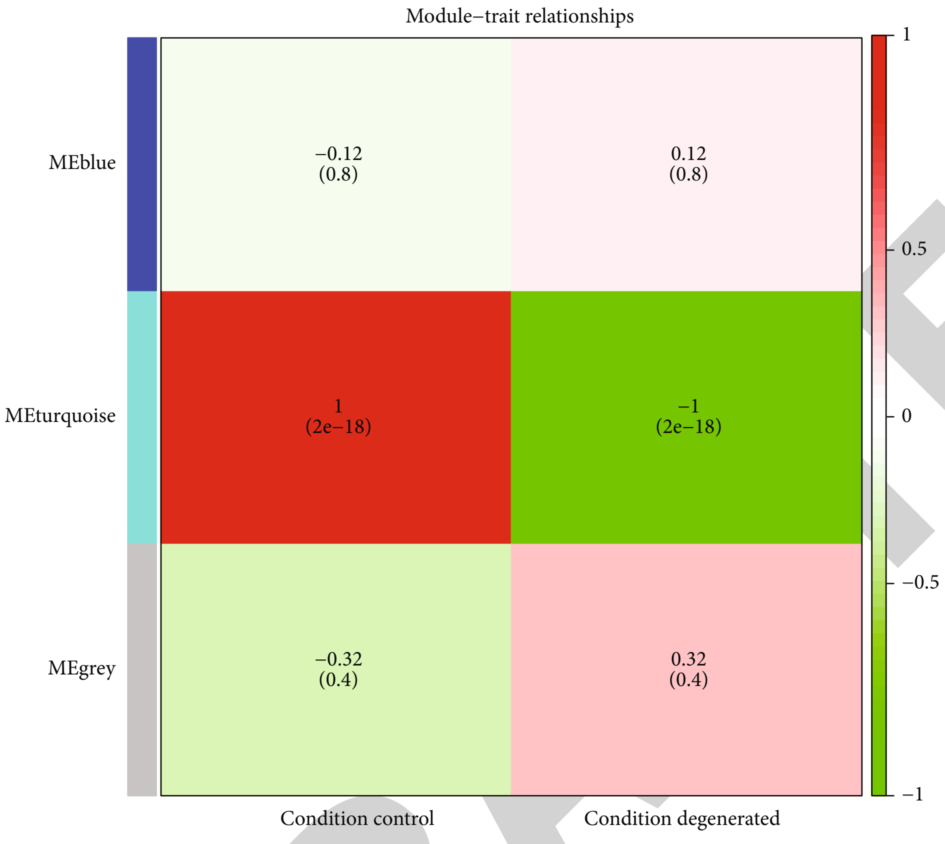
Figure 3: Continued.



(c)

(d)

Figure 3: Continued.



(e)

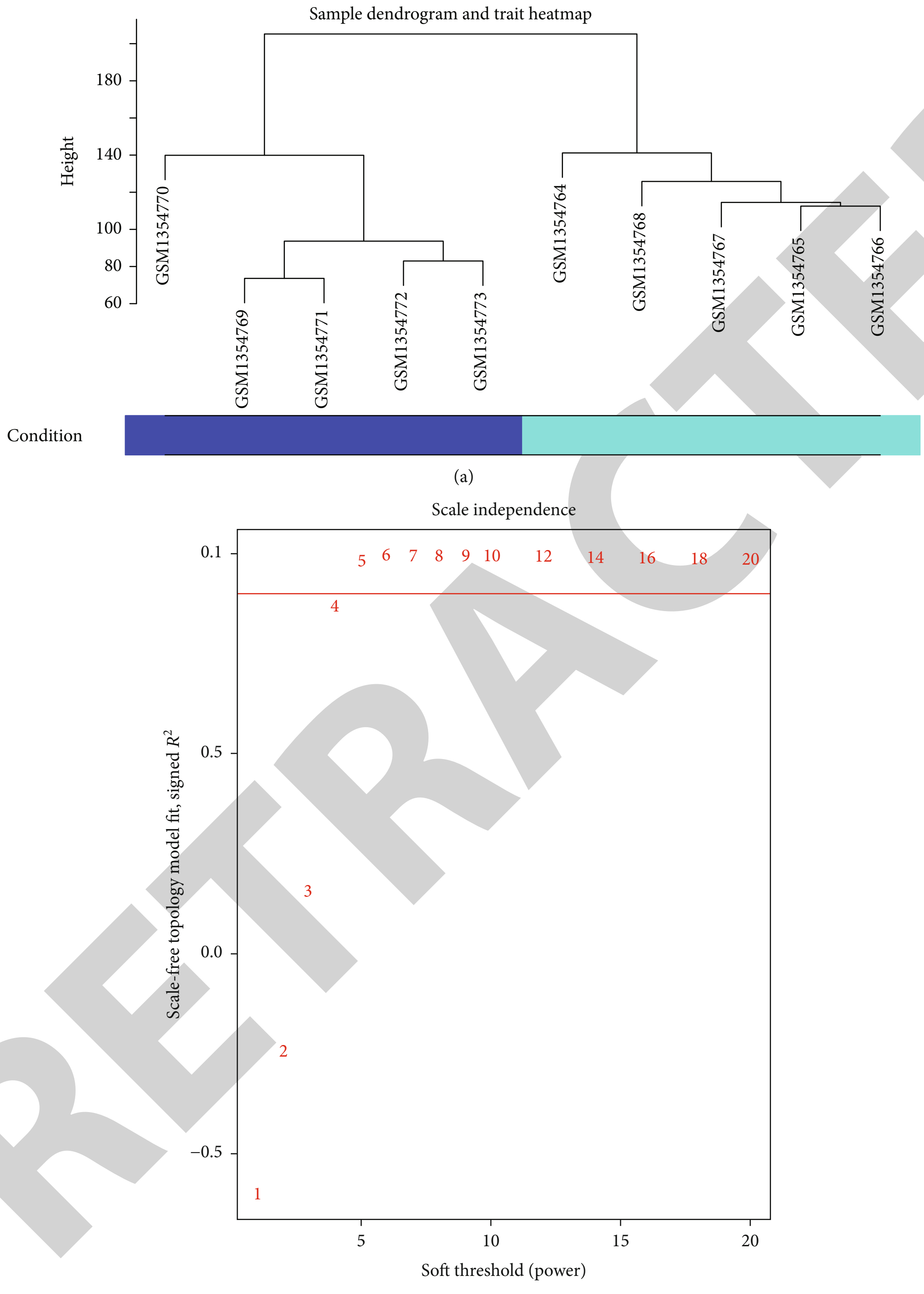
Figure 3: An IVDD-related coexpression network in the GSE67566 dataset. (a) Sample hierarchical clustering tree and corresponding sample types (IVDD or control). The height of the clustering tree was set as 100. Dark-blue represents the IVDD group and cyan-blue suggests the control group. (b) Scale-free topology model fit R^2 under different β values. (c) Mean connectivity under a series of β values. (d) Coexpressed network construction. (e) Correlation between coexpression modules and IVDD. The darker the color, the greater the correlation. Red indicates positive correlation and green is indicative of negative correlation.

the genes in the turquoise module were significantly related to IVDD ($p = 2e - 18$ and $R = -1$). There were 2726 feature genes in the turquoise module (Supplementary Table 4), which contained all 105 DEcircRNAs.

3.4. An IVDD-Related Coexpression Network in the GSE56081 Dataset. Consistent with our PCA results, there was no outlier sample in the GSE56081 dataset (Figure 4(a)). The optimal soft threshold β was determined under different values (1–20). Combining the scale-free topology model fit R^2 (Figure 4(b)) and mean connectivity (Figure 4(c)), $\beta = 4$ was determined. In the hierarchical clustering tree, 9 coexpression modules were merged via the dynamic cutting tree method (Figures 4(d)–4(f)). Pearson's correlation between modules and IVDD was analyzed. In Figure 4(g), the genes in the blue ($p = 0.002$ and $R = -0.86$) and turquoise ($p = 6e - 06$ and $R = 0.97$) modules had a significant correlation with IVDD. The feature genes in the blue ($n = 10946$) and turquoise ($n = 11717$) modules were listed in Supplementary Tables 5 and 6.

3.5. An IVDD-Related Coexpression Network in the GSE63492 Dataset. 9 samples in the GSE63492 dataset were used for coexpression analysis. No outliers were found, as shown in Figure 5(a). On the basis of scale-free topology model fit R^2 (Figure 5(b)) and mean connectivity (Figure 5(c)), 10 was the optimal threshold of β. In Figure 5(d), 4 coexpression modules were conducted by the dynamic cutting method. Among them, the genes in the blue module were distinctly correlated to IVDD ($p = 0.003$ and $R = -0.86$) in Figure 5(e). The blue module contained a total of 537 feature genes (Supplementary Table 7).

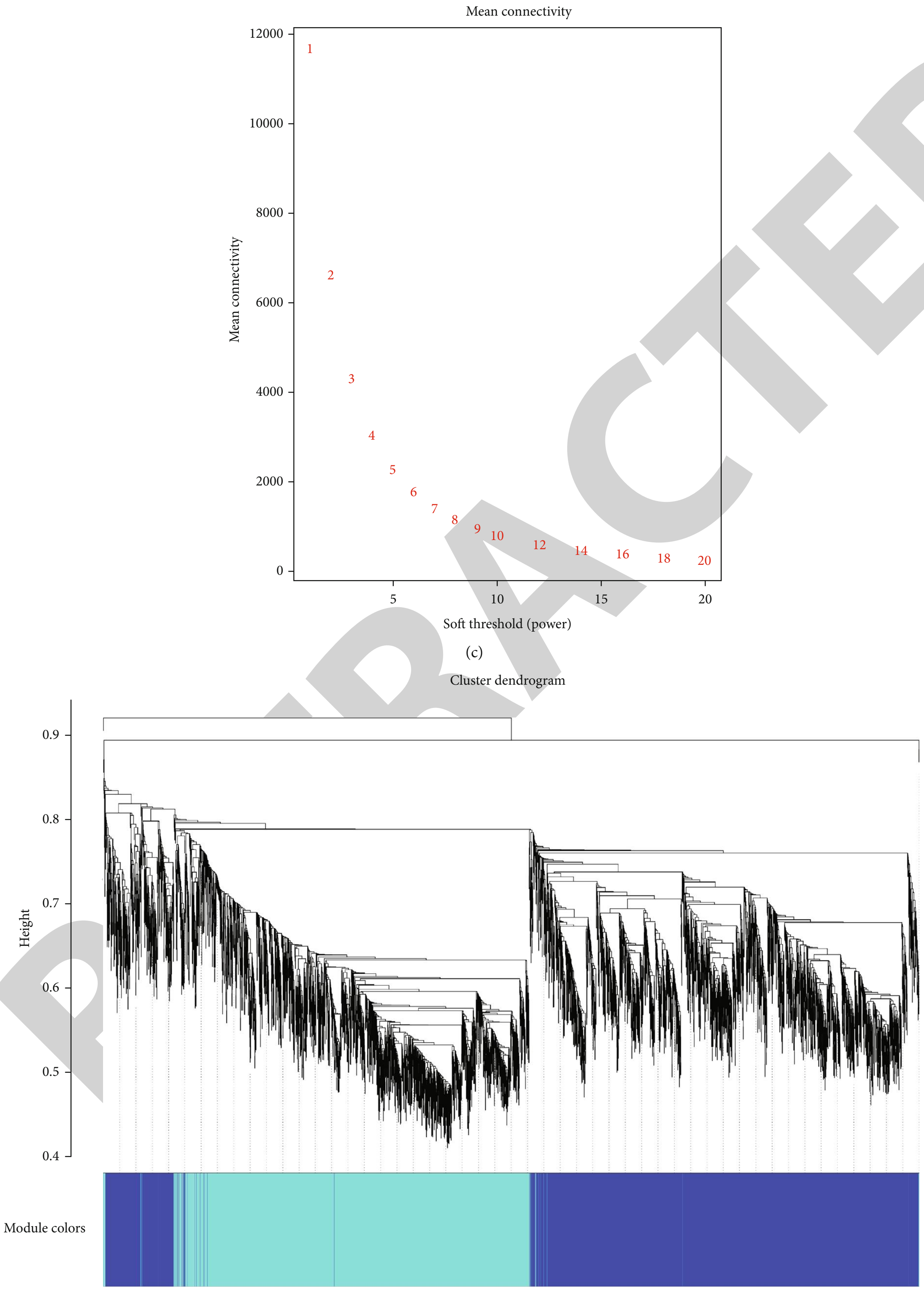
3.6. Construction of a circRNA-miRNA-mRNA Network. After overlapping feature genes in the blue and turquoise modules and DEmRNAs in the GSE56081 dataset, 1456 feature DEmRNAs were obtained (Figure 6(a)). Furthermore, a total of 35 feature DEmiRNAs were obtained by the intersection of feature genes in the blue module and DEmiRNAs in the GSE63492 dataset (Figure 6(b)). 105 DEcircRNAs were all included in the feature circRNAs. Through the StarBase database, the target mRNAs of these feature DEmiRNAs were predicted. As a result, we only obtained the target mRNAs of 8 miRNAs (hsa-miR-519d-3p, hsa-miR-489-3p, hsa-miR-486-5p, hsa-miR-431-5p, hsa-miR-4306, hsa-miR-3196, hsa-miR-193a-5p, and hsa-miR-155-5p), as shown in Figure 6(c). Then, we took the intersection between the target mRNAs of 8 miRNAs



(a)

(b)

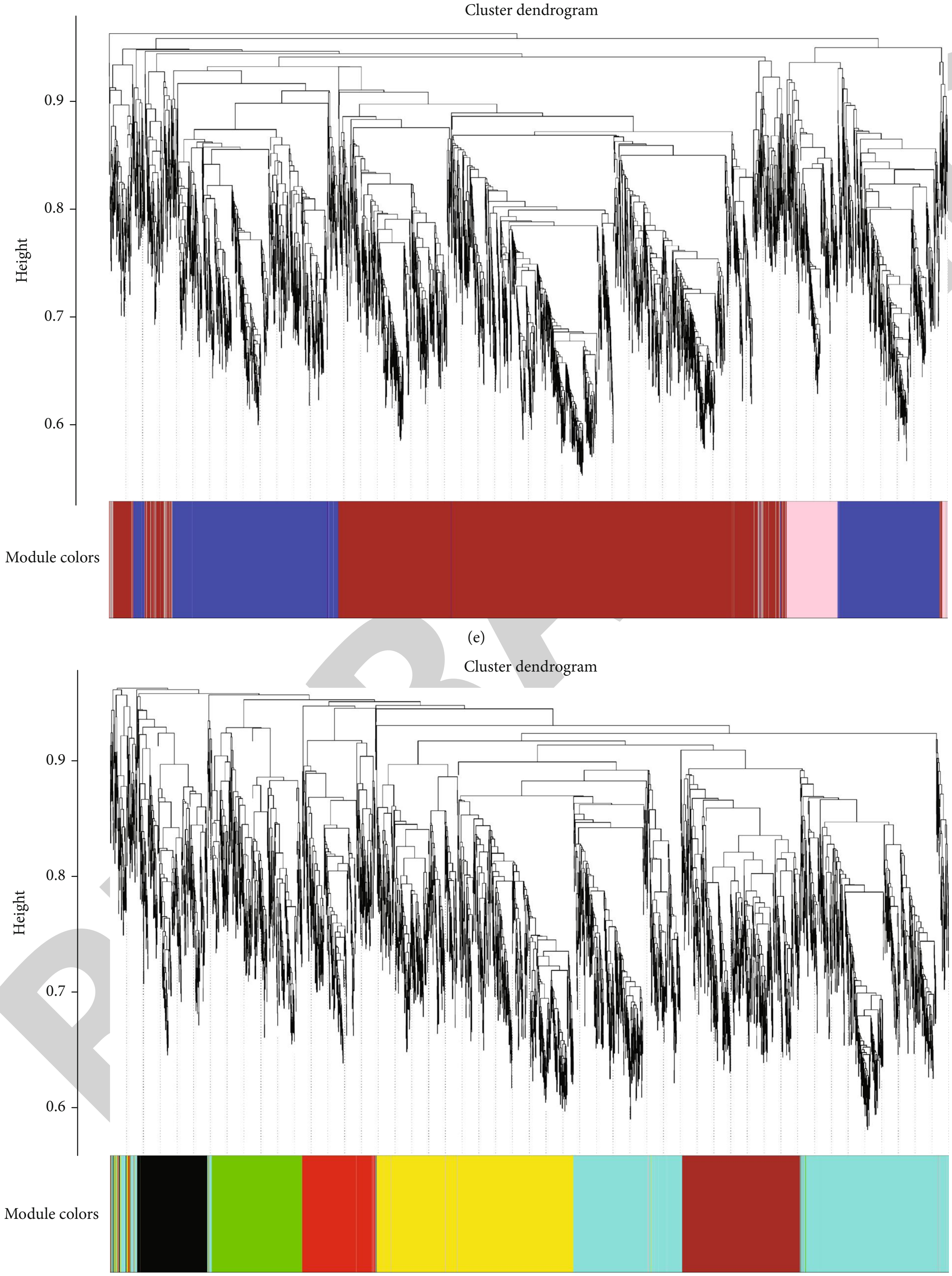
Figure 4: Continued.



(c)

(d)

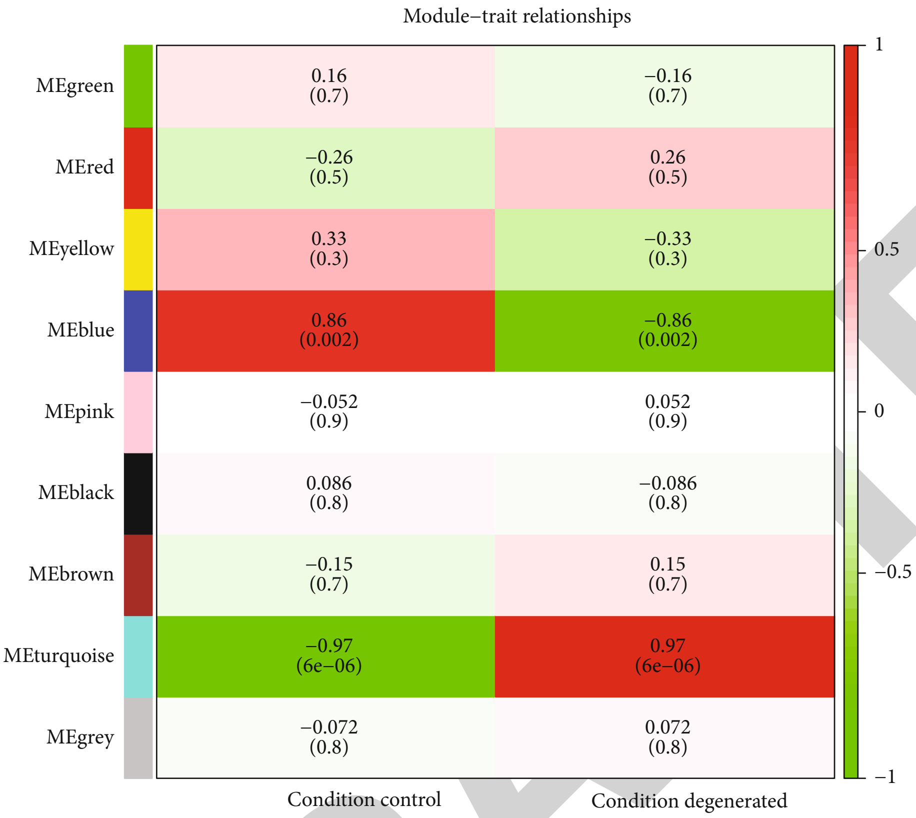
Figure 4: Continued.



(e)

(f)

Figure 4: Continued.



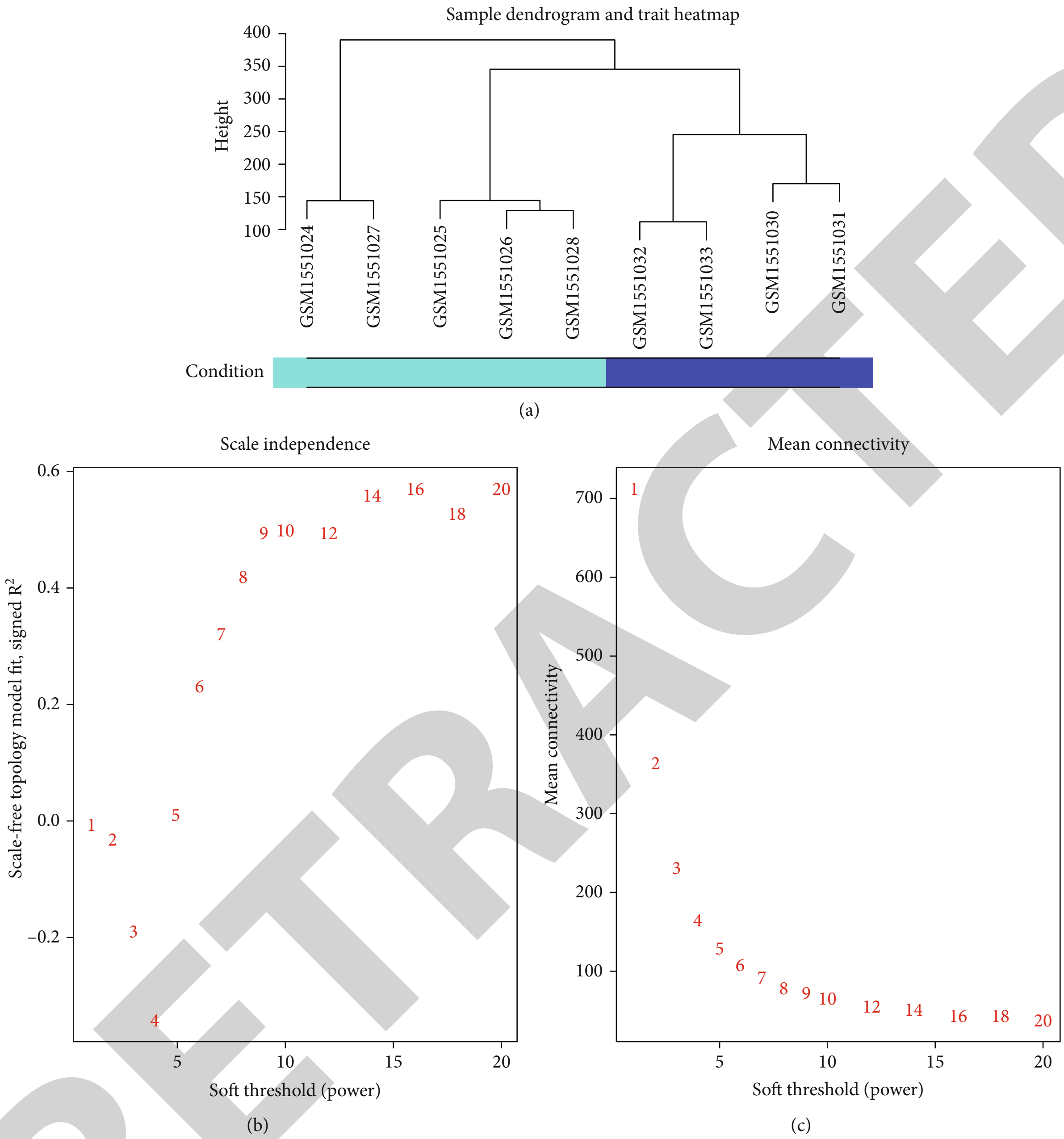
(g)

Figure 4: An IVDD-related coexpression network in the GSE56081 dataset. (a) Sample dendrogram and corresponding sample types (IVDD or control). The height of the clustering tree was set as 200. Dark-blue represents the IVDD group and cyan-blue suggests the control group. (b) Scale-free topology model fit R^2 with distinct β values (1–20). The red line indicates an R^2 value of 0.90. (c) Mean connectivity with distinct β values. (d–f) A coexpressed network containing 9 modules. (g) Pearson's correlation between coexpression modules and IVDD. The darker the color, the greater the correlation. Red suggests positive correlation and green suggests negative correlation.

and DEmRNAs. Consequently, 462 target DEmRNAs were identified, as shown in Figure 6(d). Moreover, circRNAs that targeted the 8 miRNAs were predicted using the circBase database. In Figure 6(e), 68 DEcircRNAs were intersected, which could be as the sponge of the 8 miRNAs. Based on the circRNA-miRNA and miRNA-mRNA pairs, we conducted a circRNA-miRNA-mRNA ceRNA network, composed of 68 DEcircRNAs, 8 DEmiRNAs, and 462 DEmRNAs (Figure 6(f)).

3.7. A PPI Network and Function Enrichment Analysis for DEmRNAs in the circRNA-miRNA-mRNA Network. To explore the intersections between proteins, we conducted a PPI network via the STRING database based on the DEmRNAs in the circRNA-miRNA-mRNA network. There were 212 nodes in the network (Figure 7(a)). Among them, 167 DEmRNAs were upregulated and 45 were downregulated in IVDD in comparison to controls. Furthermore, we probed into the biological functions of these DEmRNAs in the ceRNA network. In Figure 7(b), these genes were significantly related to key biological processes such as regulation of transcription, protein ubiquitination, and cell proliferation. They had a close relationship with multiple cellular components like nucleus, cytoplasm, and cytosol (Figure 7(c)). As shown in Figure 7(d), these genes exhibited distinct molecular functions such as protein binding, DNA binding, RNA binding, and transcription binding. For KEGG enrichment analysis results, critical pathways were markedly enriched like ubiquitin-mediated proteolysis, ribosome, FoxO signaling pathway, cell cycle, TNF signaling pathway, TGF-β signaling pathway, and HIF signaling pathway (Figure 7(e)).

3.8. Construction of a lncRNA-miRNA-mRNA Network for IVDD. To construct an IVDD-related lncRNA-miRNA-mRNA network, we analyzed the DElncRNA-DEmiRNA and DEmiRNA-DEmRNA pairs. Firstly, we obtained 15 feature DElncRNAs by overlapping DElncRNAs in the GSE56081 dataset and predicted lncRNAs of 8 DEmiRNAs (Figure 8(a)). In Figure 8(b), a total of 437 DEmRNAs that were targeted by 8 DEmiRNAs were identified by intersection of the targeted mRNAs of 8 DEmiRNAs and DEmRNAs in the GSE56081 dataset. Following the integration of lnRNA-miRNA and miRNA-mRNA pairs, a lncRNA-miRNA-mRNA network was built for IVDD (Figure 8(c)), composed of 15 lncRNAs, 7 miRNAs, and 437 mRNAs. A PPI network was then constructed through DEmRNAs in

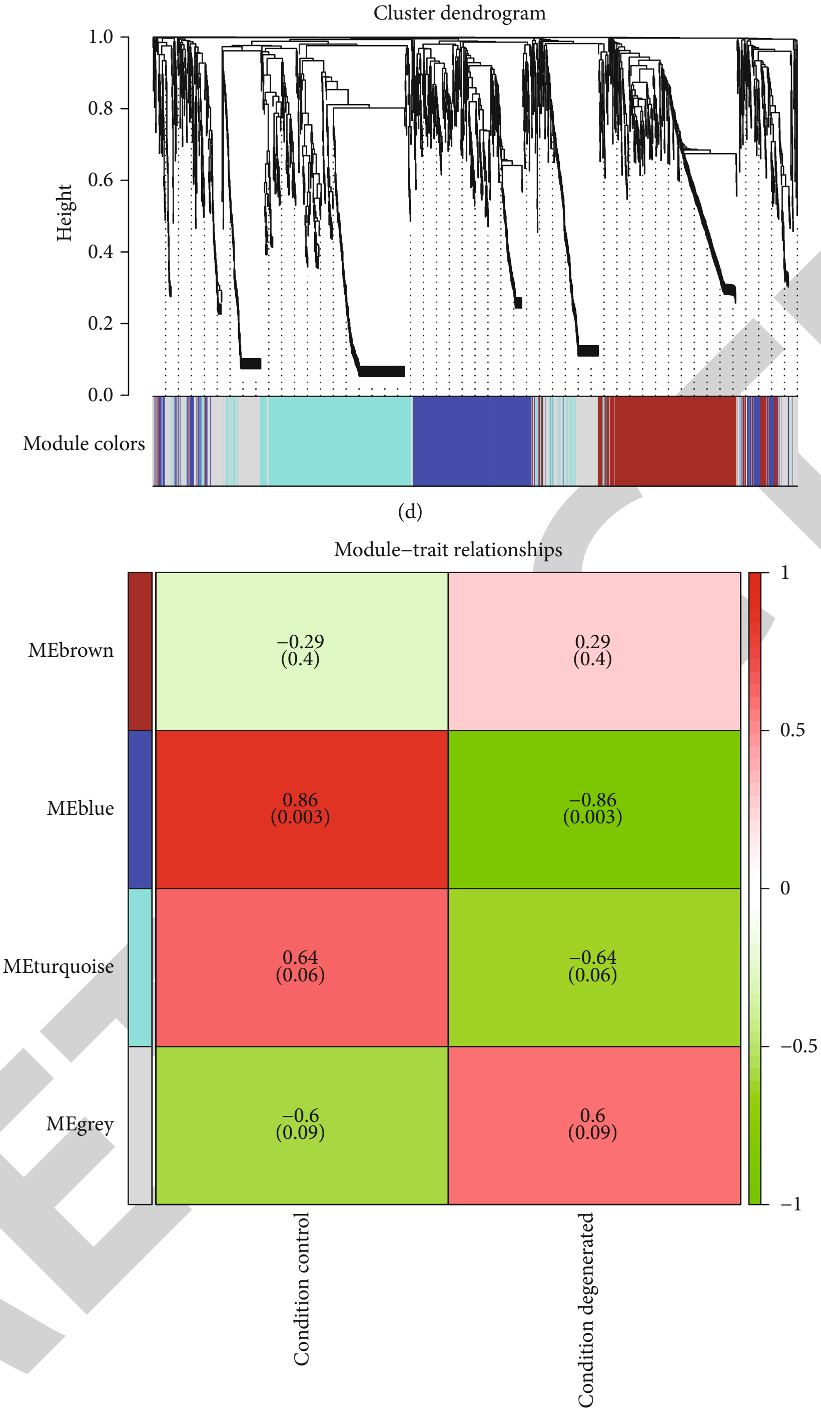


(a)

(b)

(c)

Figure 5: Continued.



(d)

(e)

Figure 5: An IVDD-related coexpression network in the GSE63492 dataset. (a) Sample hierarchical clustering tree and corresponding sample types (IVDD or control). The height of clustering tree was set as 100. Dark-blue represents the IVDD group and cyan-blue suggests the control group. (b) Scale-free topology model fit R^2 under different β values. (c) Mean connectivity under a series of β values. (d) Coexpressed network construction. (e) Correlation between coexpression modules and IVDD. The darker the color, the greater the correlation. Red indicates positive correlation and green is indicative of negative correlation.

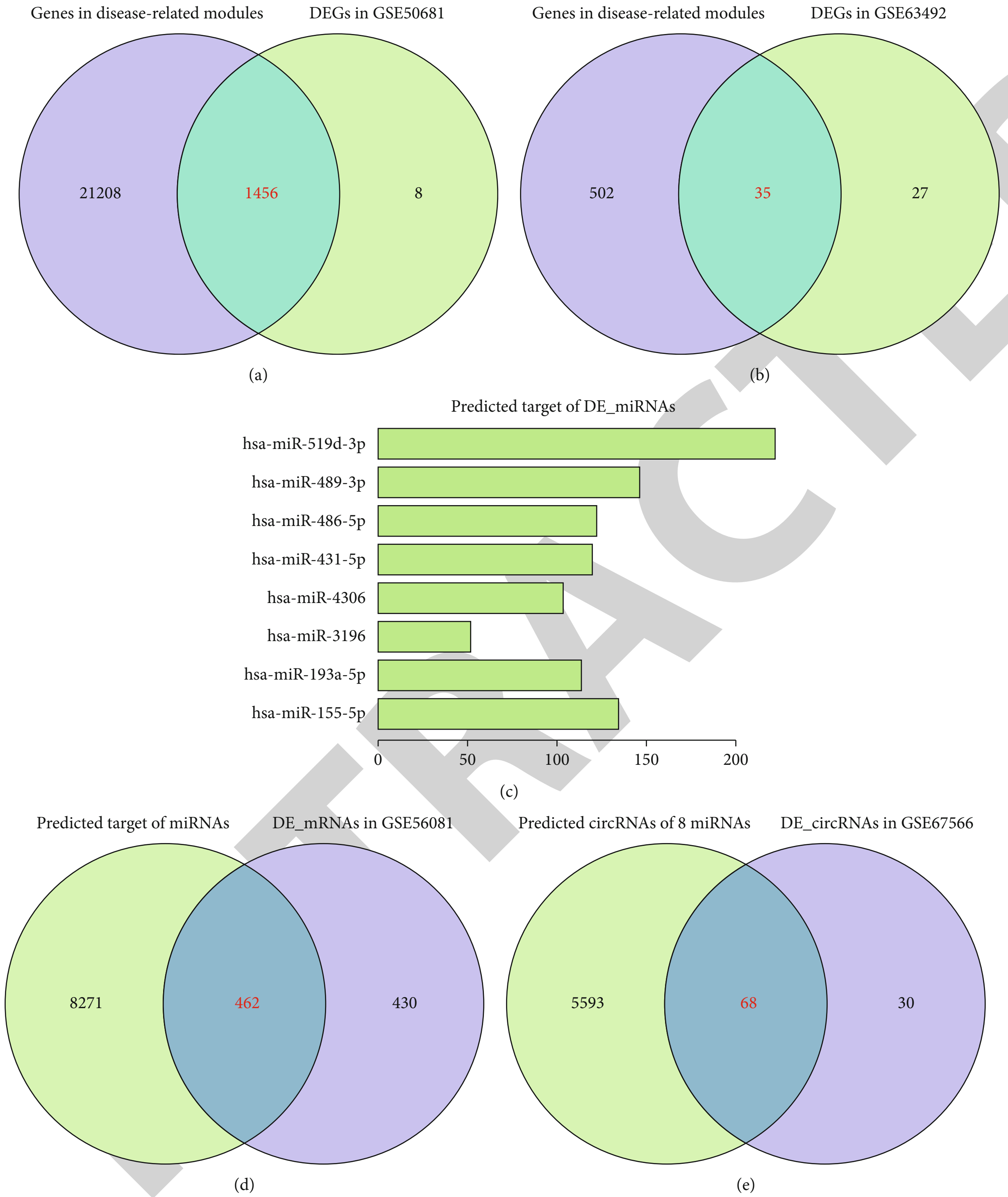
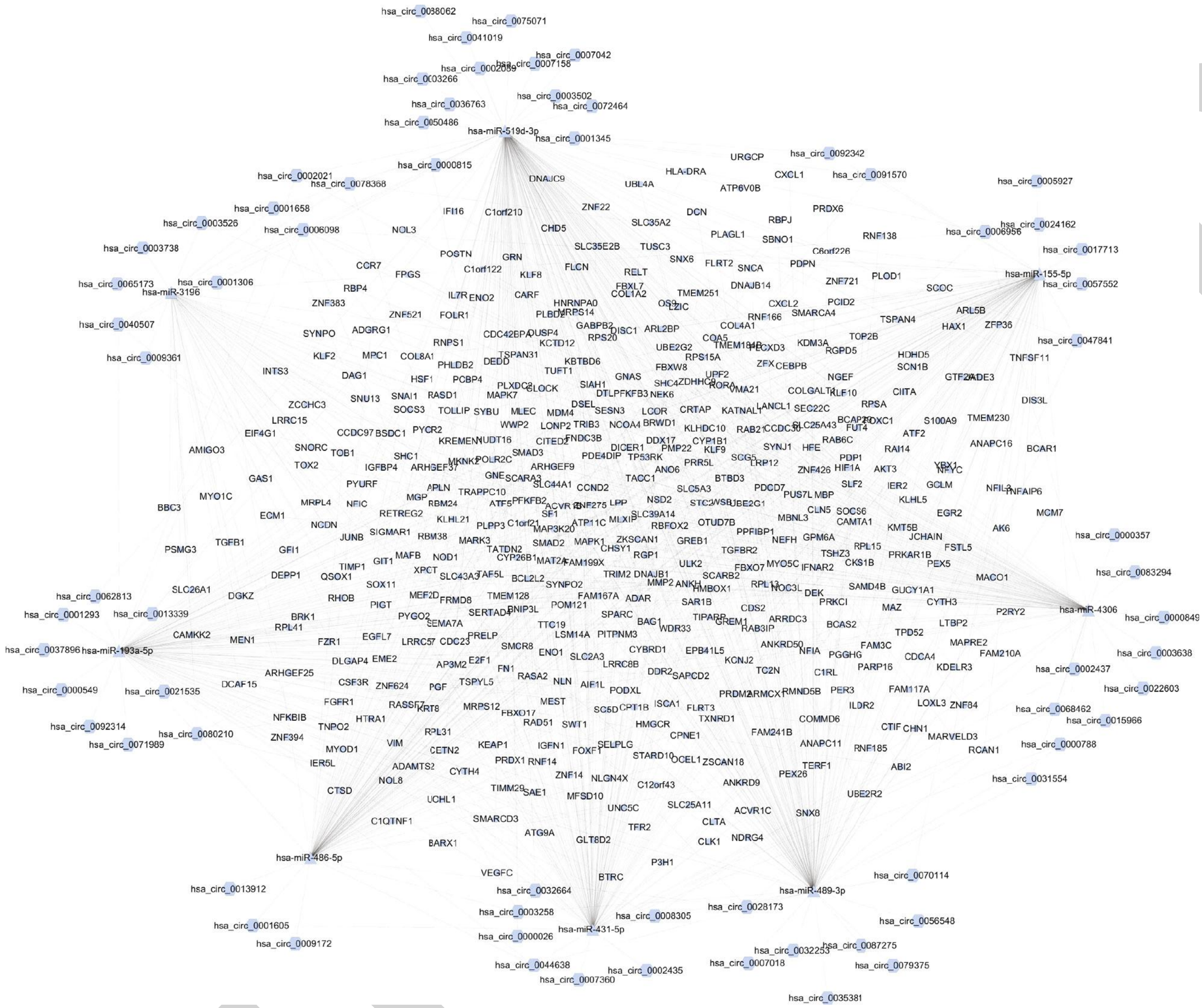


Figure 6: Continued.



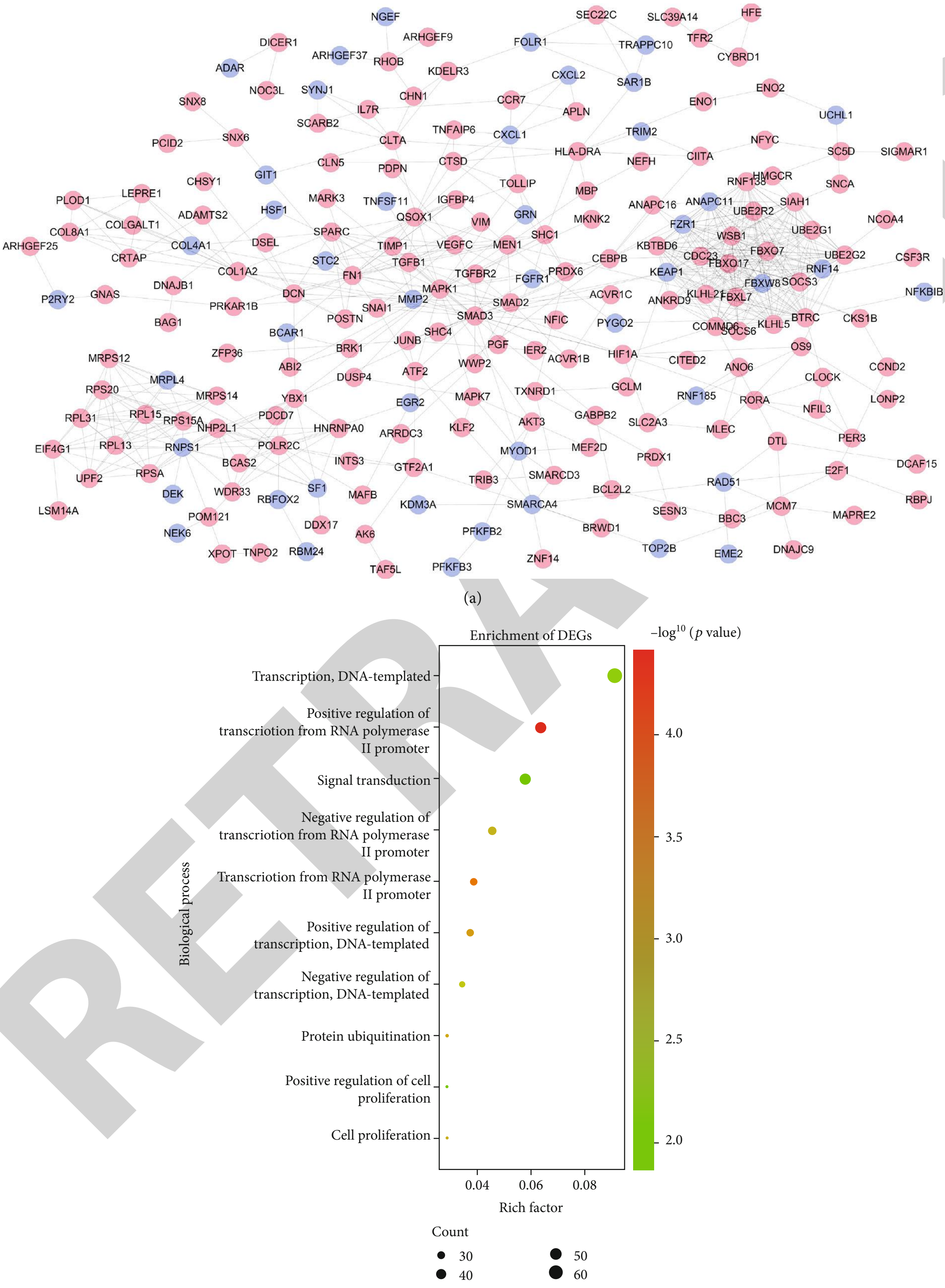
(f)

Figure 6: Construction of a circRNA-miRNA-mRNA network. (a) Venn diagram showing the overlap between feature genes in the blue and turquoise modules and DEmRNAs in the GSE56081 dataset. (b) Venn diagram depicting the overlap between feature genes in the blue module and DEmiRNAs in the GSE63492 dataset. (c) The number of targeted mRNAs of 8 DEmiRNAs. (d) The intersection between targeted mRNAs of 8 DEmiRNAs and DEmRNAs. (e) The intersection between circRNAs that targeted DEmiRNAs and DEcircRNAs. (f) The circRNA-miRNA-mRNA network for IVDD. The hexagon represents circRNA, the triangle represents miRNA, and the circle represents mRNA.

the lncRNA-miRNA-mRNA network (Figure 8(d)). In the network, there were 206 nodes, including 163 up- and 43 downregulated mRNAs in IVDD. The underlying functions of these DEmRNAs in the ceRNA network were then analyzed. Consequently, several critical biological processes were markedly enriched including the nodal signaling pathway, cell-cell junction organization, protein ubiquitination, and cell proliferation and transcription (Figure 8(e)). Moreover, they could be involved in the regulation of various cell components like cytoplasm, nucleoplasm, and nucleus (Figure 8(f)). Their molecular functions were predicted such as protein binding, transcription factor activity, and RNA binding (Figure 8(g)). KEGG enrichment analysis results suggested that these genes were distinctly related to IVDD-related pathways including the FoxO signaling pathway, TNF signaling pathway, cell cycle, TGF-β signaling pathway, and HIF signaling pathway (Figure 8(h)).

4. Discussion

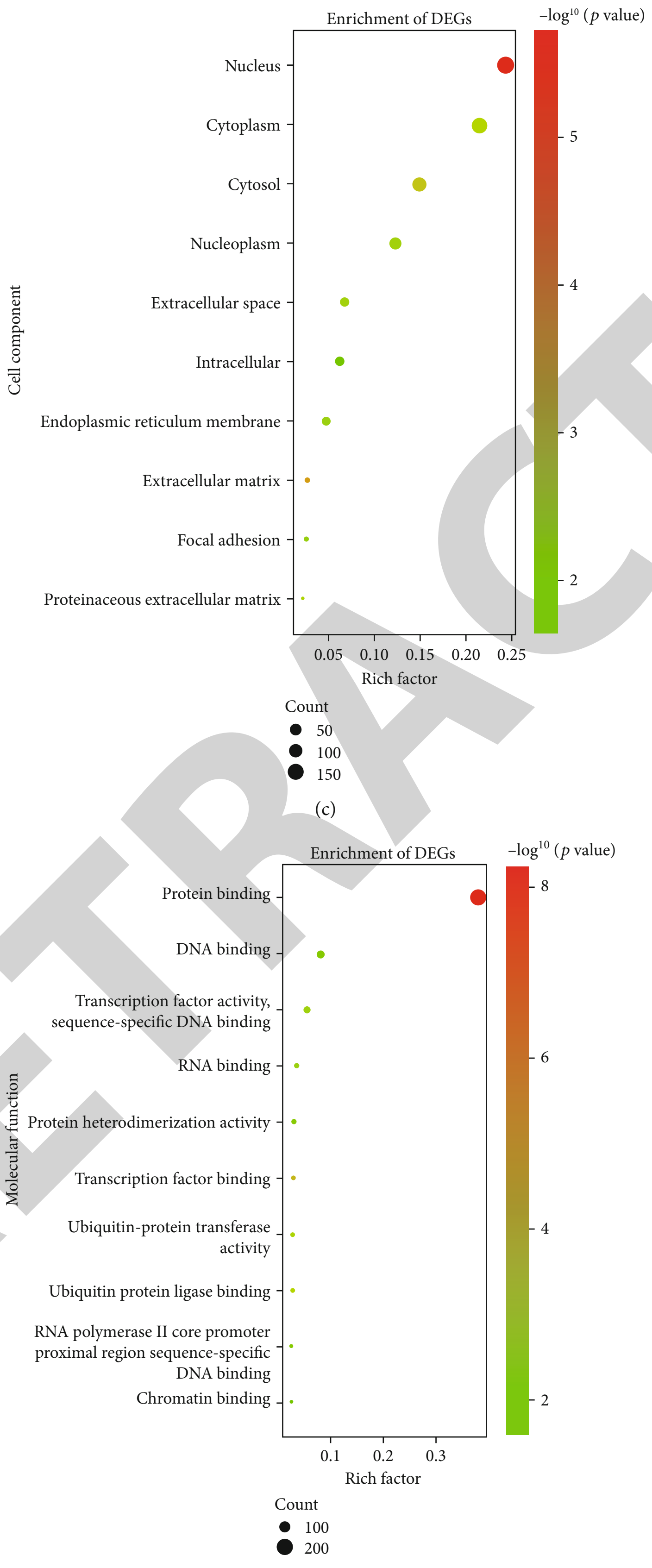
The pathological mechanisms of IVDD remain ill-defined [24]. There is currently no effective therapy for reversal of IVDD progression [25]. Thus, it is of necessity to develop new therapeutic targets for IVDD. This study comprehensively illustrated the biologically significant RNAs in nucleus pulposus for IVDD, which could be potential targets for IVDD. Their regulatory networks were conducted, revealing underlying molecular mechanisms during the progression of IVDD.



(a)

(b)

FIGURE 7: Continued.



(c)

(d)

Figure 7: Continued.

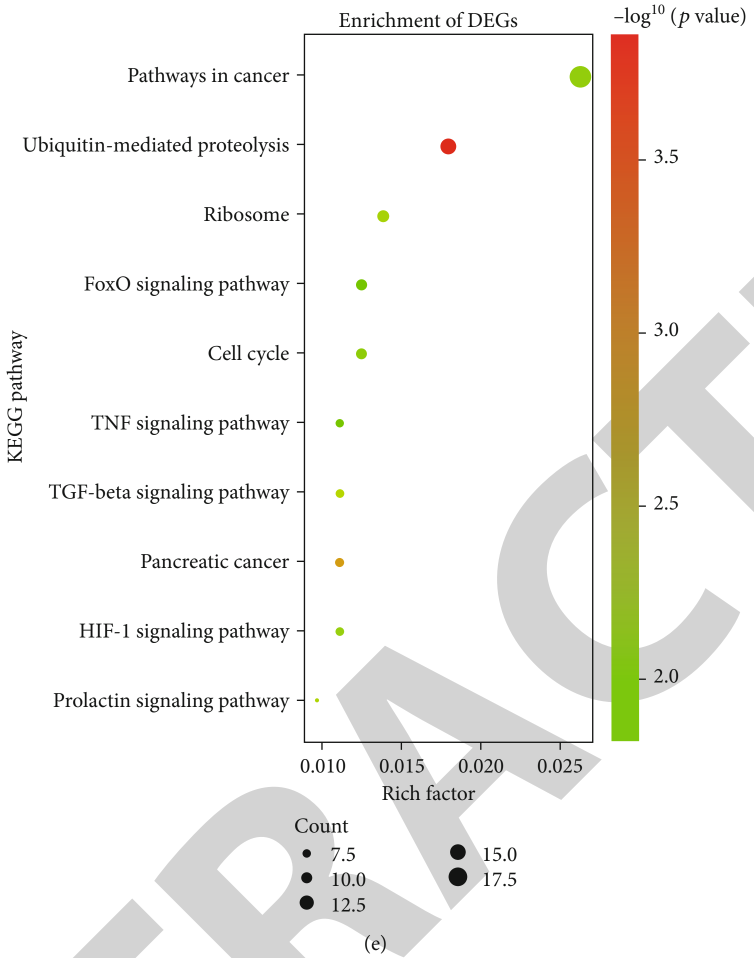


Figure 7: A PPI network and function enrichment analysis for DEmRNAs in the circRNA-miRNA-mRNA network. (a) A PPI network based on DEmRNAs in the circRNA-miRNA-mRNA network. The red circle is indicative of upregulated mRNA in IVDD and blue is indicative of downregulated mRNA in IVDD. (b) The top ten biological processes of DEmRNAs in the ceRNA network. (c) The top ten cellular components of DEmRNAs in the ceRNA network. (d) The top ten molecular functions of DEmRNAs in the ceRNA network. (e) The top ten KEGG pathways of DEmRNAs in the ceRNA network. The more the color tends to be redder, the smaller the p value. The size of the circle is proportional to the number of enriched genes.

The progression of IVDD is an extremely complex biological process. Studies have shown that mRNAs, lncRNAs, miRNAs, and circRNAs are all involved in the development of IVDD [12]. This study explored differentially expressed RNAs in nucleus pulposus tissues between IVDD and control samples based on transcriptomics [12]. With the criteria of adjusted $p < 0.05$ and $|\log^2 \mathrm{FC}| > 2$, 105 DEcircRNAs, 131 DElncRNAs, DEmRNAs, and 62 DEmiRNAs were identified for IVDD, which represented key genes during the progression of IVDD.

The construction of coexpression modules through the WGCNA algorithm to find potential key genes has been widely applied in various diseases [26–28]. For example, through WGCNA, key modules and genes related to non-small-cell lung cancer have been identified [29]. The genes that are divided into the same module exhibiting the same molecular functions. By calculation of the correlation between the module and the trait, we can determine the modules related to the trait and lock the feature genes. In this study, we applied WGCNA to find feature circRNAs, lncRNAs, mRNAs, and miRNAs related to IVDD development. By integration of differentially expressed circRNAs, lncRNAs, mRNAs, and miRNAs, feature RNAs have been explored for IVDD. It has been confirmed that circRNAs and lncRNAs could serve as the sponge of miRNAs, thereby regulating the expression of downstream target genes [30]. In the circRNA-miRNA-mRNA ceRNA network, there were 68 DEcircRNAs, 8 DEmiRNAs, and 462 DEmRNAs. For the lncRNA-miRNA-mRNA network, there were 15 DElncRNAs, 7 DEmiRNAs, and 437 DEmRNAs. Their regulatory mechanisms should be validated in cellular levels in depth.

We further explored the biological functions of DEmRNAs in the circRNA- and lncRNA-miRNA-mRNA networks. Our results showed that these mRNAs were mainly enriched in transcription, cell proliferation, TNF, TGF-β, and HIF signaling pathways. The inflammatory process exacerbated by TNF-α is a key mediator of IVDD [31]. TNF-α participates in distinct pathological processes of IVDD like inflammation, apoptosis, and proliferation [32]. TGF-β pathway activation is

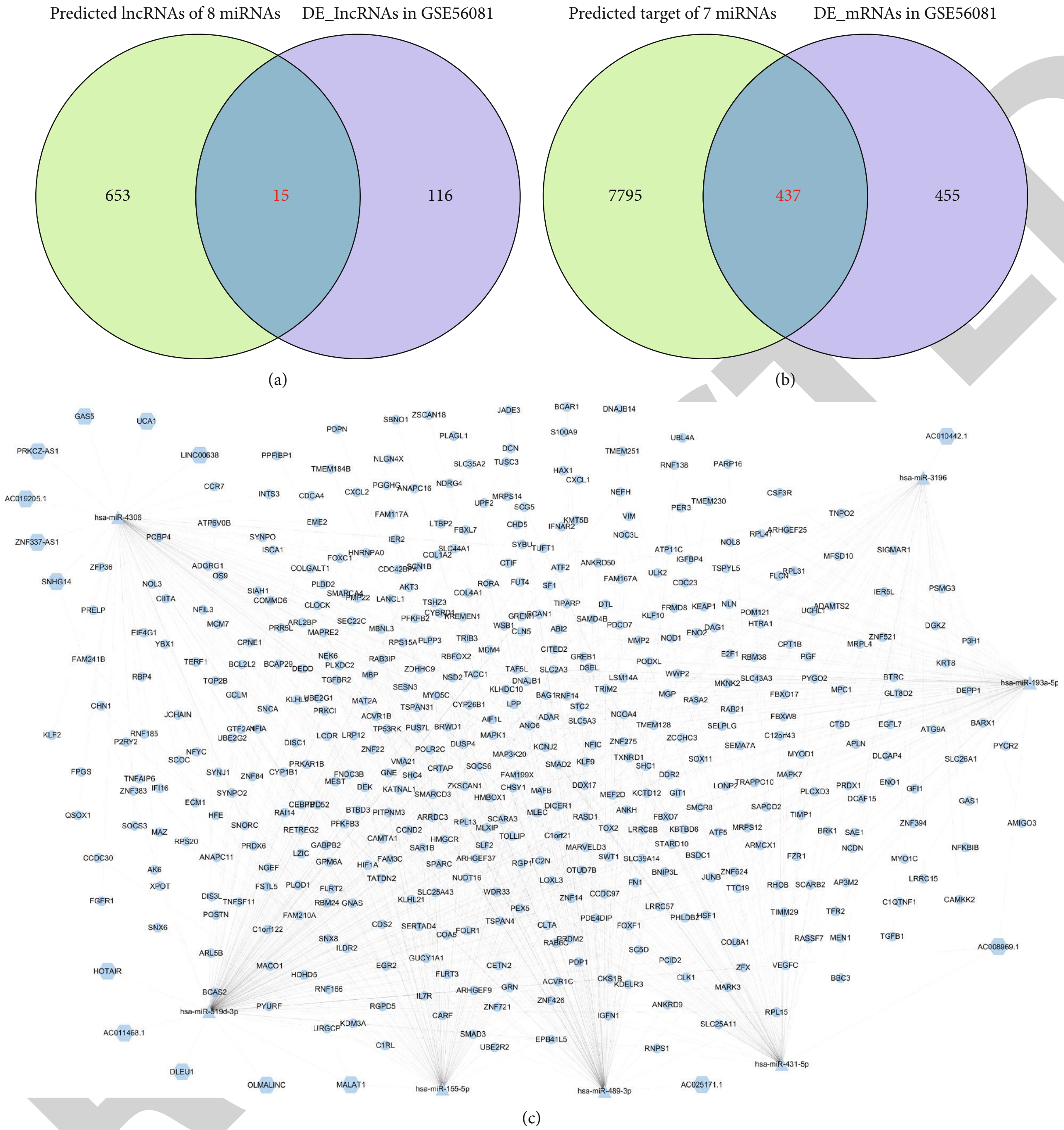
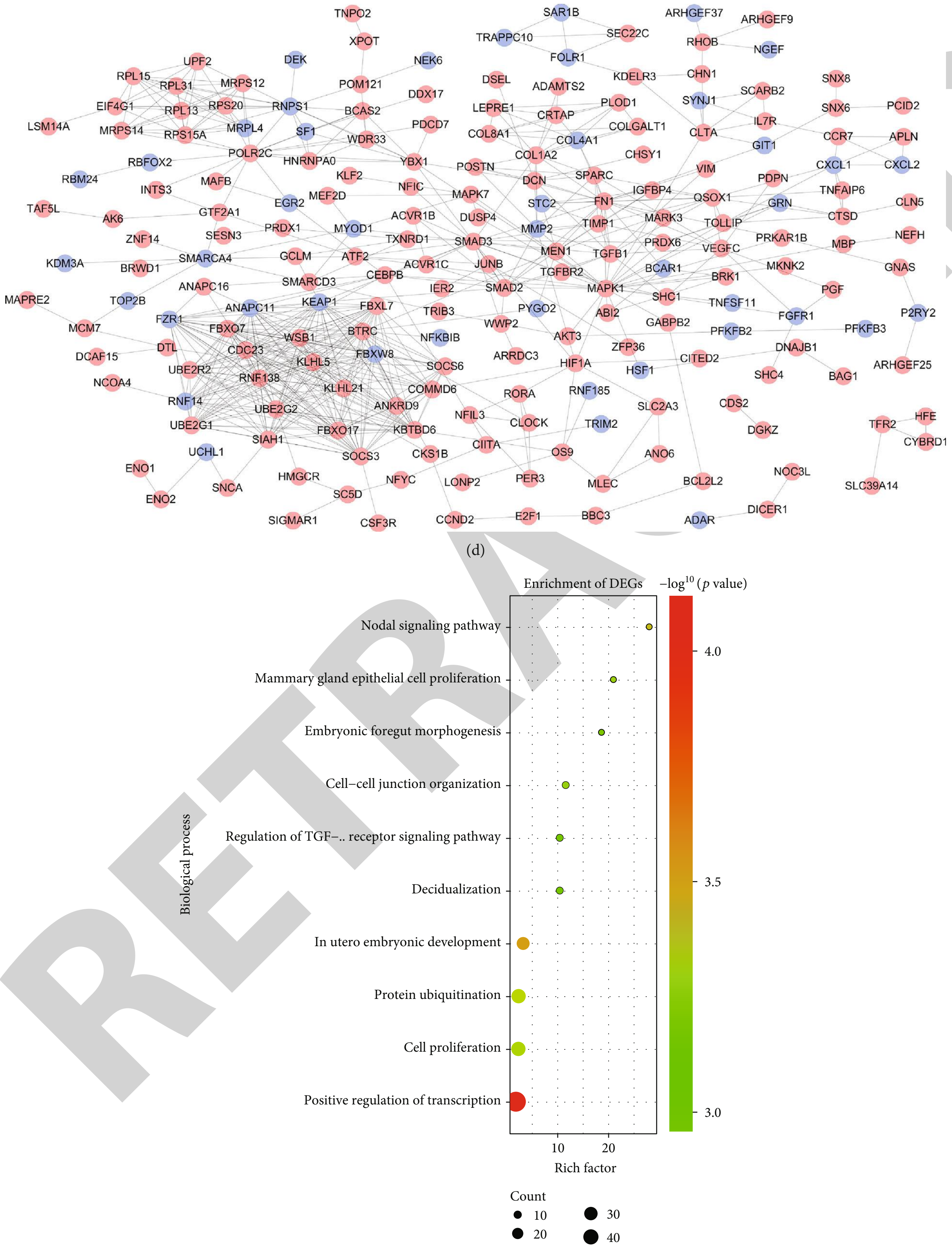


Figure 8: Continued.



(d)

(e)

Figure 8: Continued.

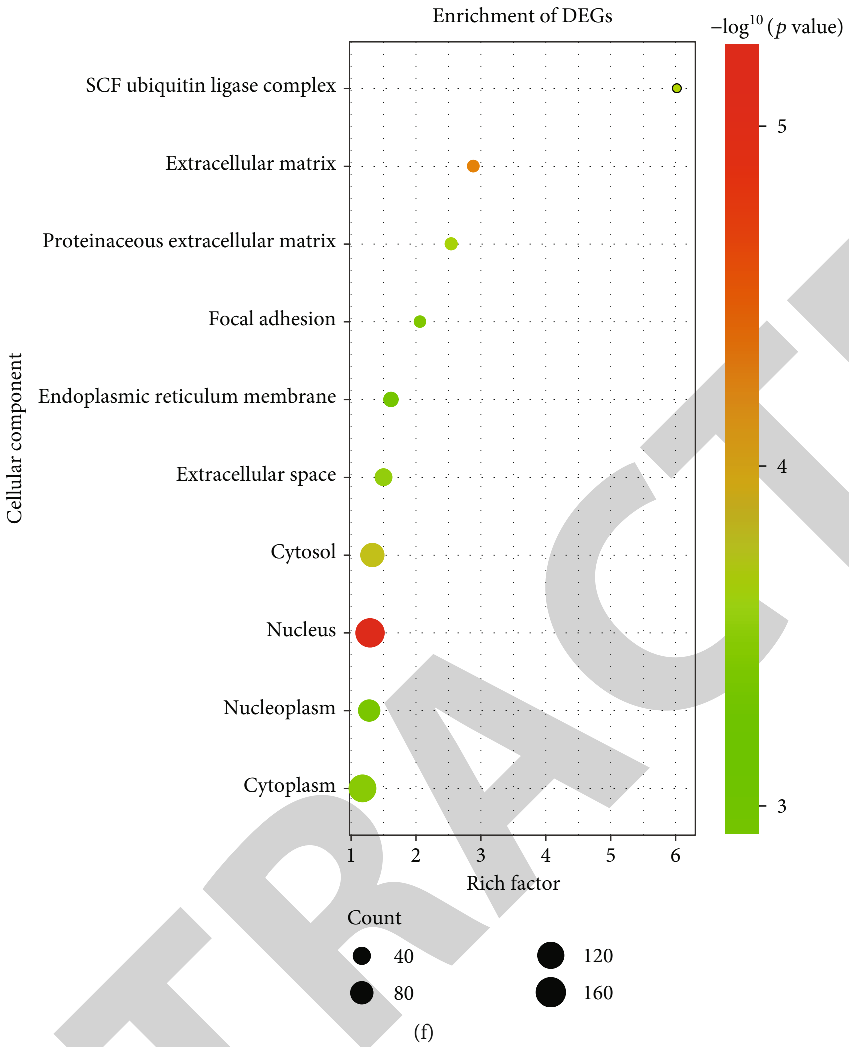
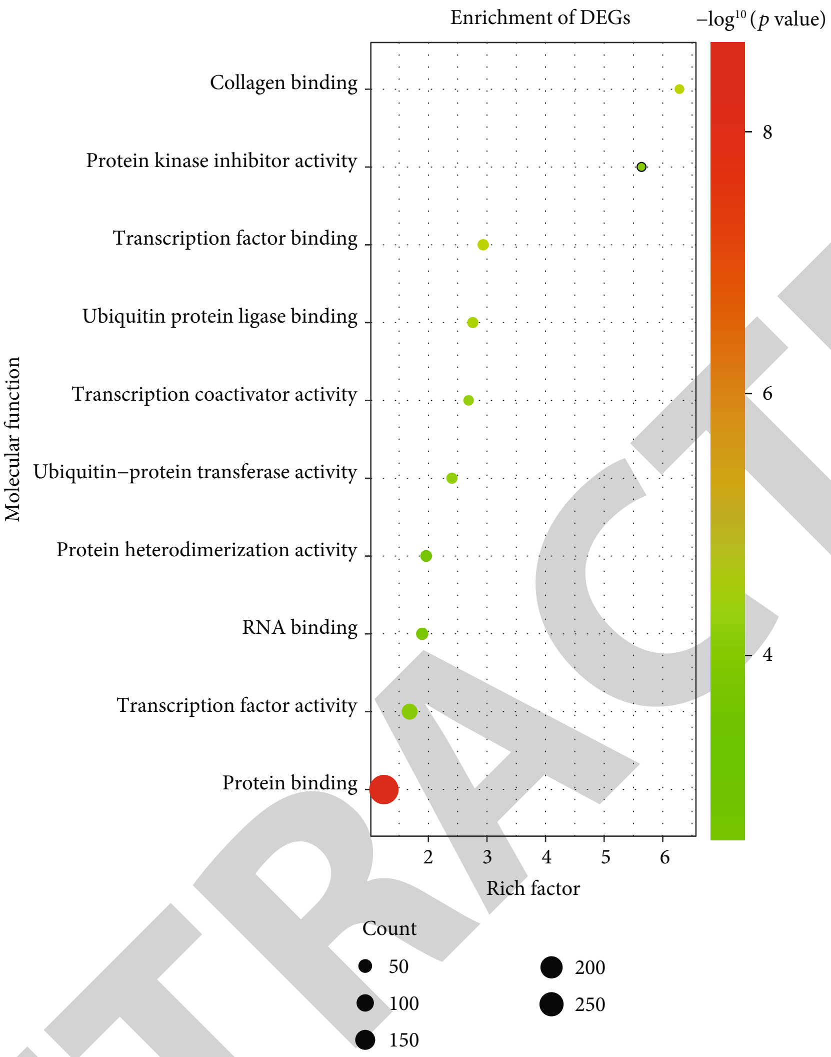
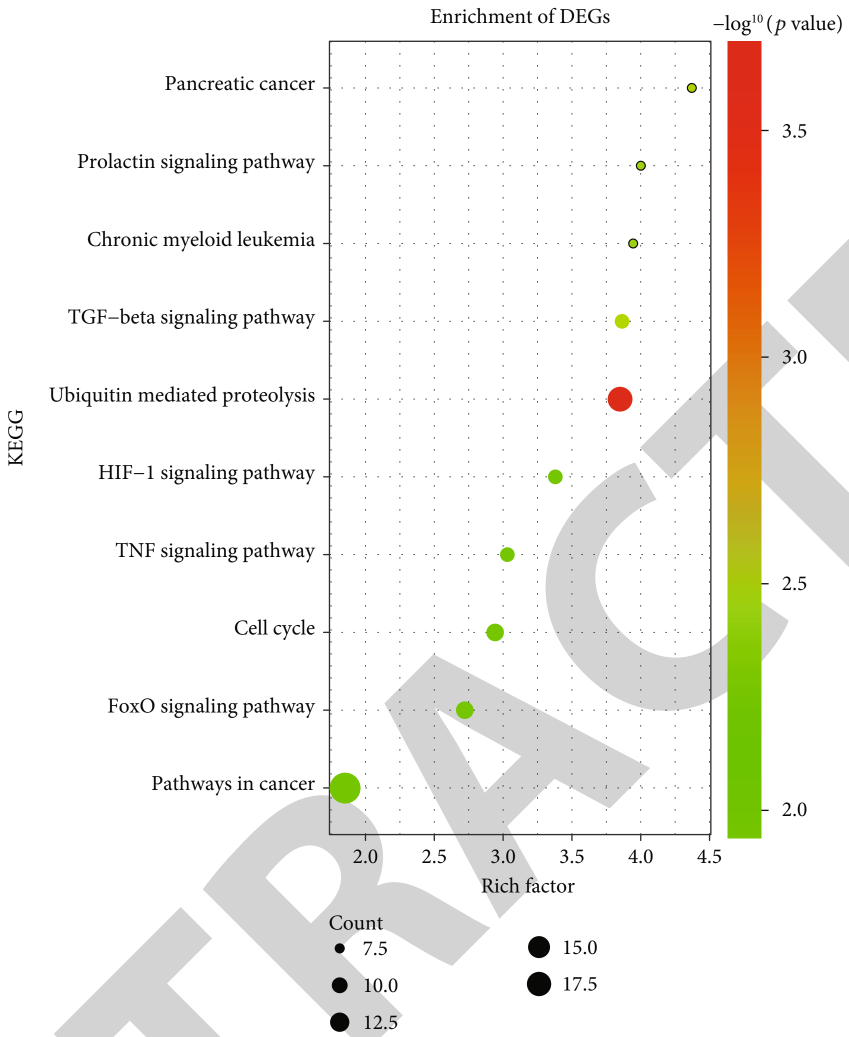


FIGURE 8: Continued.



(g)

Figure 8: Continued.



(h)

Figure 8: Construction of a lncRNA-miRNA-mRNA network for IVDD. (a) Venn diagram showing the intersection between targeted lncRNAs of 8 DEmiRNAs and DElncRNAs in the GSE56081 dataset. (b) The overlap between predicted mRNAs of 8 DEmiRNAs and DEmRNAs in the GSE56081 dataset. (c) A lncRNA-miRNA-mRNA network for IVDD. (d) A PPI network based on the DEmRNAs in the ceRNA network. The red circle suggests upregulated mRNA in IVDD and the blue circle indicates downregulated mRNA in IVDD. (e) The top ten biological processes of DEmRNAs in the ceRNA network. (f) The top ten cellular components of DEmRNAs in the ceRNA network. (g) The top ten molecular functions of DEmRNAs in the ceRNA network. (h) The top ten KEGG pathways of DEmRNAs in the ceRNA network. The more the color tends to be redder, the smaller the *p* value. The size of the circle is proportional to the number of enriched genes.

a promising treatment strategy for IVDD. However, its excessive activation could accelerate IVDD progression. Our findings offered more clues on its specific molecular mechanisms [33]. Furthermore, hypoxia affects the synthesis of intervertebral disc cells and weakens the ability to support the extracellular matrix of the intervertebral disc [34]. Hence, these DEmRNAs may be involved in the progression of IVDD via various biological processes and pathways.

However, several limitations should be pointed out in this study. Firstly, the sample size was not insufficient. In our future studies, the above findings of biologically significant noncoding RNAs (circRNAs, lncRNAs, and miRNAs) and mRNAs will be validated in a larger IVDD cohort. Secondly, our study is a preliminary screening study. Hence, we will verify the biological functions and interactions of these noncoding RNAs and mRNAs in IVDD by a series of experiments. Collectively, our findings identified IVDD-related RNAs and constructed two complex ceRNA networks, which offered potential therapeutic targets as well as molecular mechanisms for IVDD.

5. Conclusion

Taken together, this study identified biologically significant noncoding RNAs (circRNAs, lncRNAs, and miRNAs) and mRNAs. The regulatory networks involved in them were conducted for IVDD, indicating the complex molecular mechanisms. These DEmRNAs participated in distinct IVDD-related biological processes or pathways. Hence, these RNAs could be promising therapeutic targets for IVDD.

Abbreviations

IVDD:	Intervertebral disc degeneration
GEO:	Gene Expression Omnibus
WGCNA:	Weighted gene correlation network analysis
PPI:	Protein-protein interaction
ECM:	Extracellular matrix
NPC:	Nucleus pulposus cells
miRNAs:	MicroRNAs
lncRNA:	Long noncoding RNA
circRNAs:	Circular RNAs
limma:	Linear Models for Microarray Data
PCA:	Principal component analysis
DEcircRNAs:	Differentially expressed circRNAs
DElncRNAs:	Differentially expressed lncRNAs
DEmRNAs:	Differentially expressed mRNAs
DEmiRNAs:	Differentially expressed miRNAs
ceRNA:	Competing endogenous RNA
STRING:	Search Tool for the Retrieval of Interacting Genes
GO:	Gene ontology
KEGG:	Kyoto Encyclopedia of Genes and Genomes
DAVID:	Database for Annotation, Visualization, and Integrated Discovery.

Data Availability

The data used to support the findings of this study are included within the supplementary information files.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This work was funded by Shanghai Science and Technology Commission Project (19JC1411702) and Retrospective Clinical Research Project of Shanghai Sixth People's Hospital (ynhg201901).

Supplementary Materials

Supplementary 1. Supplementary Table 1. 105 differentially expressed circRNAs in nucleus pulposus tissues between IVDD and control groups.

Supplementary 2. Supplementary Table 2. 131 differentially expressed lncRNAs and 928 differentially expressed mRNAs in nucleus pulposus tissues between IVDD and control groups.

Supplementary 3. Supplementary Table 3. 62 differentially expressed miRNAs in nucleus pulposus tissues between IVDD and control groups.

Supplementary 4. Supplementary Table 4. 2726 genes in the turquoise module for the GSE67566 dataset.

Supplementary 5. Supplementary Table 5. 2726 genes in the blue module for the GSE56081 dataset.

Supplementary 6. Supplementary Table 6. 2726 genes in the turquoise module for the GSE56081 dataset.

Supplementary 7. Supplementary Table 7. 537 feature genes in the blue module in the GSE63492 dataset.

References

[1] W. Hua, S. Li, R. Luo et al., "Icariin protects human nucleus pulposus cells from hydrogen peroxide-induced mitochondria-mediated apoptosis by activating nuclear factor erythroid 2-related factor 2," *Biochimica et Biophysica Acta - Molecular Basis of Disease*, vol. 1866, p. 165575, 2020.

[2] P. Sampara, R. R. Banala, S. K. Vemuri, G. R. AV, and S. GPV, "Understanding the molecular biology of intervertebral disc degeneration and potential gene therapy strategies for regeneration: a review," *Gene Therapy*, vol. 25, no. 2, pp. 67–82, 2018.

[3] G. Zheng, Z. Pan, Y. Zhan et al., "TFEB protects nucleus pulposus cells against apoptosis and senescence via restoring autophagic flux," *Osteoarthritis and Cartilage*, vol. 27, no. 2, pp. 347–357, 2019.

[4] W. Fang, X. Zhou, J. Wang et al., "Wogonin mitigates intervertebral disc degeneration through the Nrf2/ARE and MAPK signaling pathways," *International Immunopharmacology*, vol. 65, pp. 539–549, 2018.

[5] Y. Zhang, F. He, Z. Chen et al., "Melatonin modulates IL-1β-induced extracellular matrix remodeling in human nucleus pulposus cells and attenuates rat intervertebral disc degeneration and inflammation," *Aging (Albany NY)*, vol. 11, no. 22, pp. 10499–10512, 2019.

[6] Z. Buser, A. S. Chung, A. Abedi, and J. C. Wang, "The future of disc surgery and regeneration," *International Orthopaedics*, vol. 43, no. 4, pp. 995–1002, 2019.

[7] S. Chen, M. Luo, H. Kou, G. Shang, Y. Ji, and H. Liu, "A review of gene therapy delivery systems for intervertebral disc degeneration," *Current Pharmaceutical Biotechnology*, vol. 21, no. 3, pp. 194–205, 2020.

[8] Z. Li, X. Li, C. Chen et al., "Long non-coding RNAs in nucleus pulposus cell function and intervertebral disc degeneration," *Cell Proliferation*, vol. 51, no. 5, article e12483, 2018.

[9] M. L. Ji, H. Jiang, X. J. Zhang et al., "Preclinical development of a microRNA-based therapy for intervertebral disc degeneration," *Nature Communications*, vol. 9, no. 1, p. 5051, 2018.

[10] X. Cheng, L. Zhang, K. Zhang et al., "Circular RNA VMA21 protects against intervertebral disc degeneration through targeting miR-200c and X linked inhibitor-of-apoptosis protein," *Annals of the Rheumatic Diseases*, vol. 77, no. 5, pp. 770–779, 2018.

[11] Z. Li, X. Chen, D. Xu, S. Li, M. T. Chan, and W. K. Wu, "Circular RNAs in nucleus pulposus cell function and intervertebral disc degeneration," *Cell Proliferation*, vol. 52, article e12704, 2019.

[12] J. Zhu, X. Zhang, W. Gao, H. Hu, X. Wang, and D. Hao, "lncRNA/circRNA-miRNA-mRNA ceRNA network in lumbar intervertebral disc degeneration," *Molecular Medicine Reports*, vol. 20, no. 4, pp. 3160–3174, 2019.

[13] L. Xie, W. Huang, Z. Fang et al., "CircERCC2 ameliorated intervertebral disc degeneration by regulating mitophagy and

Hindawi
BioMed Research International
Volume 2021, Article ID 6676171, 14 pages
https://doi.org/10.1155/2021/6676171

Research Article

A Secure Storage and Sharing Scheme of Stroke Electronic Medical Records Based on Consortium Blockchain

Qiuli Qin,[1] **Biyuan Jin**,[1] **and Yanqing Liu**[2]

[1]*School of Economics and Management, Beijing Jiaotong University, Beijing 100000, China*
[2]*Health Service Center, Beijing Jiaotong University, Beijing 100000, China*

Correspondence should be addressed to Qiuli Qin; qlqin@bjtu.edu.cn

Received 29 December 2020; Revised 11 January 2021; Accepted 20 January 2021; Published 3 February 2021

Academic Editor: Parameshachari B. D.

The maintenance and sharing of electronic medical records are one of the essential tasks in the medical treatment combination. Traditional cloud-based electronic medical record storage system is difficult to realize data security sharing. The tamper resistance and traceability of blockchain technology provide the possibility for the sharing of highly sensitive medical data. This paper proposes a safe sharing scheme of stroke electronic medical records based on the consortium blockchain. The scheme adopts the storage method of ciphertext of medical records stored in the cloud and index of medical records stored on the blockchain. The privacy protection mechanism proposed in this paper innovatively combines proxy reencryption and searchable encryption which supports patient pseudoidentity search. The mechanism could achieve controllable sharing of medical records and precise search. According to the organizational characteristics of the stroke medical treatment combination, this paper proposes an improved Practical Byzantine Fault Tolerance mechanism to reach a consensus between consensus nodes. Then, the proposed scheme is analyzed and evaluated from three aspects of medical record integrity, user privacy, and data security. The results show that the scheme can not only ensure the privacy of patient identity information and private key data but also resist the tampering and deletion attacks of internal and external malicious nodes on the medical record data. Therefore, the proposed scheme is conducive to the improvement of the timeliness of stroke treatment and the safe sharing of electronic medical records in stroke medical treatment combination.

1. Introduction

Stroke is a major chronic noncommunicable disease that seriously endangers the health of Chinese people. It is the first cause of death and disability among adults in China. It has five characteristics: high morbidity, high disability, high mortality, high recurrence, and high economic burden [1]. The burden of cerebrovascular disease in China ranks first in the world. Since stroke is an emergency, whether it can be treated in a professional stroke center within the prime time is a key factor in treatment. "Healthy China 2030 Plan" put forward a request for medical institutions above the secondary level to establish a scientific regional collaborative medical treatment system for acute cerebrovascular diseases and promote the construction of stroke centers [2]. The stroke medical treatment combination is to form a hierarchical treatment system based on the stroke center in the region and realize the standardized chronic disease management of stroke disease through the community-second-level hospital-tertiary hospital linkage and cooperation model [3]. The key role of the medical treatment combination lies in the realization of referral treatment, which involves the storage and sharing of clinical and diagnosis and treatment multimode heterogeneous data such as patient medical record homepage information and electronic medical records, which contains the standard medical and clinical data gathered by physicians [4]. The rapid identification, diagnosis, and treatment of stroke determine the prognosis to a large extent. After the higher-level hospital receives the referral patient, it is essential to obtain the patient's past medical history and physical examination, blood routine examination, vascular ultrasound examination, and other risk factors screening collected in the lower-level hospital for the treatment of patients with stroke. In addition to the above check items, the electronic medical record also contains the patient's personal privacy information, such as name, date of birth, and address. The highly sensitive feature requires that the electronic medical record can not only be able to be shared authentically and credibly among medical institutions in the medical association but also meet the access control rights of the data owner and meet the needs of privacy protection [5].

The traditional stroke medical treatment combination uses cloud storage technology to build an information-sharing platform. The advantages of cloud storage technology include fast data transmission, high storage capacity, low cost, easy access to information, and dynamic communication [6, 7]. However, centralized storage is vulnerable to single-point attacks, there is a high risk of electronic medical record data leakage and tampering, and the security, integrity, and immutability of electronic medical records cannot be guaranteed. Blockchain technology is expanding its use from the parts that can be applied to transactions that exclude intermediary agencies [8]. The development of blockchain technology provides new opportunities to solve the above problems. Blockchain relies on public key encryption and hashing mechanisms to track and record the historical transactions of data on the chain. Data copies are distributed to each participating node in the network to ensure that records will not be lost or mistakenly modified, altered, or accessed by unauthorized users. And the blockchain can reach a consensus between distributed entities without relying on a single trusted party, thereby building a shared platform with trust, reliability, and transparency.

In recent years, blockchain has received extensive attention in the medical field. Medical data storage, sharing, and privacy protection have become hot research issues for scholars. To improve the storage scalability of massive, multimode, and heterogeneous medical big data, Azaria et al. [9] proposed MedRec, a scheme based on public blockchain for electronic medical records, and designed three Ethereum smart contracts to achieve fine-grained access control for patients to medical records. However, using public blockchains to store medical data is too expensive and cannot restrict the identity of network participants. Besides public blockchains, scholars have found that consortium blockchains are more suitable for medical data storage. Zhang et al. [10] proposed a secure storage and sharing scheme for medical records based on dual-blockchain and cloud server. Medical institutions form a consortium blockchain to store medical metadata. Each institution deploys a private chain to store digital abstracts of records. However, the cost of the dual-chain solution is high, and there is no guarantee that each medical institution has the economic and technical capabilities to build their own private chain. Kumar et al. [11] proposed a distributed medical data chain storage system that combines the IPFS system and blockchain technology. The scheme stores the index of medical records in the blockchain and the medical records in the IPFS distributed storage system. Research on the privacy protection of medical data based on blockchain is mainly focused on two aspects: cryptography and access control. In the field of cryptography, [12, 13] proposed medical record sharing schemes based on blockchain that use a symmetric key to encrypt medical data and then use the patient's public key to encrypt the symmetric key. In the field of access control, [14–16] proposed to use an identity-based access control model for electronic medical record sharing. The role determines the rights of the data resource. However, this method has drawbacks in fine-grained control. To solve the problem, scholars combine cryptography and access control schemes to achieve fine-grained access control. The BPDS model designed by Liu et al. [17] uses a joint scheme of CP-ABE-based access control mechanism and content extraction to achieve fine-grained access control. Niu et al. [18] proposed a secure searchable electronic medical record sharing scheme based on blockchain, which supports multiuser search and uses the ciphertext-based attribute encryption mechanism to achieve fine-grained access control. Luo et al. [19] proposed to combine distributed key generation technology with identity-based proxy reencryption technology and selected proxy nodes in the blockchain to reencrypt EHR to ensure the correctness of user private key generation. It can be found that most of the existing solutions are applied to the scenarios of patient life-cycle electronic medical record management and clinical scientific research sharing, which storage and access control methods are not suitable for single-disease electronic medical record sharing in regional referral scenarios.

Aiming at the problems of centralized storage of electronic medical records, weak interoperability, and difficulty in safe sharing in the traditional stroke medical treatment combination, this paper proposes a blockchain-based electronic stroke sharing scheme in referral scenarios to achieve targeted and accurate referrals. To meet the storage requirements, the scheme only uses the consortium blockchain for access control. The consortium blockchain stores medical metadata blocks containing medical record hash values, indexes, hospital signatures, and other information, while the original data of electronic medical records is encrypted and randomly stored in the cloud, thereby reducing the main chain pressure. To achieve the security and privacy requirements of electronic medical records sharing, this paper proposes a blockchain data privacy protection mechanism based on searchable encryption and proxy reencryption and realizes the privacy protection of patients' identity by setting anonymity. Searchable encryption technology is used to realize the encryption of key data and the ciphertext search of medical records. The searchable encryption in this study allows the search for the pseudoidentity of the patient. The proxy reencryption technology is used to realize the decryption of the requested data by the data visitor without revealing the patient's private key. Patients can designate the hospital entity to access the authorized personal data within a given time frame. In the scheme in this paper, all interactions between users and users and the blockchain are encrypted with digital signatures, and identity verification is performed to ensure system security. Aiming at the application scenario of the stroke medical treatment combination, this paper proposes to improve the preselection node rules of the Practical Byzantine Fault Tolerance (PBFT) consensus mechanism according to the classification of medical institutions in the medical treatment combination to improve the reliability and safety of the system.

2. Materials and Methods

2.1. Blockchain Technology. Blockchain is a decentralized distributed data ledger connected by a series of ordered blocks. Nakamoto first promoted the blockchain technology that integrates cryptography and peer-to-peer communication technology in the Bitcoin white paper [20]. Blockchain is usually managed cooperatively by a peer-to-peer network, which reaches consensus by following an agreement for authenticating new blocks into the blockchain. There is no fixed central node in the blockchain network. All nodes in the network store a copy of the blockchain information. The data on a single node is tampered with or destroyed, which will not affect the data stored on the blockchain. Each block in the blockchain ledger is linked by a cryptographic hash value, that is, each block contains the hash value of the previous block content, and the irreversible hash function is used as the link mechanism to verify the integrity of the previous block [16]. This feature makes the blockchain as a decentralized distributed database immutable and traceable. The typical block structure of the Bitcoin is shown in Figure 1. A block is mainly composed of a block header and block body. The block header is made up of six components. Block header contains the hash of the previous block, the Merkle root hash value generated by the transaction ID, the timestamp when the block was generated, the version which indicates the validation rules to follow for a particular data type, and the target difficulty and random number used for consensus node to execute consensus mechanism. All the transaction orders are saved in the block body. The blocks are connected by the previous hash value to become a chain. The structure of the block ensures that the block content could not be modified or tampered with.

Blockchain networks can be divided into public chain, private chain, and consortium blockchain according to the scope of the network. The public chain is an open blockchain that allows any user to participate, and there is no identity authentication and permission setting. The transactions on the chain are completely open and transparent. All users can obtain a complete account book in the chain. Typical public chain platforms include Bitcoin and Ethereum [21]. The consortium blockchain stipulates that only approved network members can join, and it is usually managed by several institutions or organizations, and the processing speed is faster than that of the public chain. Private chains have centralized ownership and management rights and are only used within the organization. The blockchain system in the medical treatment combination scenario requires multilevel medical institutions, patients, doctors, and other entities to interact with each other, and medical record data is only accessible within the medical treatment combination, so the consortium chain is selected as the blockchain type of this paper.

2.2. Searchable Encryption. The searchable encryption originated from the development of cloud storage. In the cloud storage mode, cheap computing and considerable capacity attract increasingly users to outsource private data to cloud servers to save local storage and maintenance costs. However, considering the centralized characteristics of the cloud environment and the semitrusted and semihonest attributes of cloud servers, data is vulnerable to theft and loss. This problem can be solved by encrypting the data before uploading, but at the same time, there will be difficulties in ciphertext retrieval. In 2000, Song [22] first proposed the concept of searchable encryption to realize the search for ciphertext keywords without revealing user privacy. Therefore, in the medical record sharing system with cloud chain and storage proposed in this paper, searchable encryption technology is introduced, which allows searching for patients' pseudoidentities and medical record keywords to achieve precise matching.

Searchable encryption technology can be divided into symmetric searchable encryption (SSE) and asymmetric searchable encryption (ASE) according to the encryption method. SSE uses the same key for encryption and decryption, and the data owner does not need to interact with the data requester. ASE, that is, public key searchable encryption, is used to solve the problem of untrustworthy server routing. The encryption process involves two keys. The public key is used to encrypt the matching target ciphertext of the plaintext keyword information, and the private key is used to generate keyword trapdoors. In the searchable encryption mechanism, the keyword trapdoor controls the search matching behavior. Once the blockchain node receives the search trapdoor, it can perform search matching. In the medical treatment combination referral scenario, the patient has the right to use and ownership of the data, and the keyword trapdoor should be generated by the patient using private key encryption. Therefore, the scheme proposed adopts the asymmetric searchable encryption mechanism to prevent visitors from using searchable encryption public keys to generate the desired keyword trapdoors infinitely under the symmetric searchable encryption mechanism. The public-key encryption with keyword search (PEKS) algorithm first proposed by Boneh et al. [23] is described as follows:

(1) *Initialization Algorithm.* Shown as Equation (1). Enter the security parameters 1^λ to obtain the private key sk and public key pk

$$\mathrm{Setup}\left(1^\lambda\right) \longrightarrow (\mathrm{sk}, \mathrm{pk}). \tag{1}$$

(2) *Keyword Encryption Algorithm.* Shown as Equation (2). Enter the public key pk and keywords w of documents, and output the ciphertext of document keywords c

$$\mathrm{Encrypt}(\mathrm{pk}, w) \longrightarrow c. \tag{2}$$

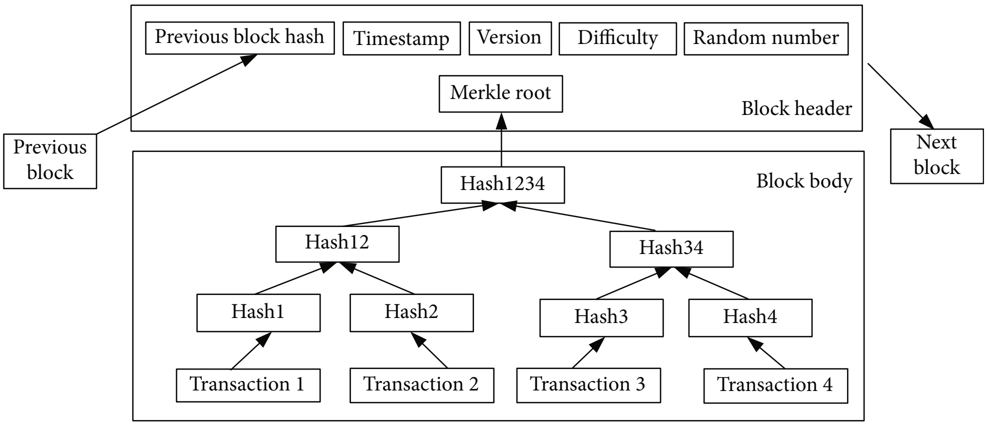


Figure 1: Structure of block.

(3) *Trapdoor Generation Algorithm*. Shown as Equation (3). Enter the private key sk and search keyword w, and output the trapdoor td corresponding to the search keyword

$$\text{Tropdoor}(\text{sk}, w) \longrightarrow \text{td}. \quad (3)$$

(4) *Matching Algorithm*. Shown as Equation (4). Input the generated public key pk, trapdoor td, and ciphertext c, and output Boolean variable b. When the trapdoor and ciphertext correspond to the same keyword, $b = 1$, otherwise, $b = 0$

$$\text{Test}(\text{td}, c, \text{pk}) \longrightarrow b. \quad (4)$$

2.3. Proxy Reencryption. Proxy reencryption is a cryptographic concept proposed by Blaze et al. in 1998 [24]. It is a mechanism for converting ciphertexts, which solves the problem of sharing the private key of the data owner when transferring encrypted records between nodes. A trusted third party or a semihonest agent is usually entrusted as an agent to reconstruct the encrypted message in some way. Even if another user did not encrypt the message with his associated public key, he can still use his private key to decrypt the message. Agent reencryption can ensure that even if the agent has the conversion key, he cannot obtain the plaintext information, thereby enhancing the reliability and security of the data. Proxy reencryption can be divided into one-way proxy reencryption and two-way proxy reencryption according to the direction of ciphertext conversion; according to the number of times of reencryption key conversion, it can be divided into single-hop ciphertext conversion and multihop ciphertext conversion [25]. In this paper, one-way one-hop proxy reencryption technology is adopted to ensure the privacy of the patient's private key during the medical record sharing process and reduce the number of user interactions. The master node of the consortium blockchain acts as an agent to reencrypt the ciphertext of the medical record. The workflow of proxy reencryption is as follows:

(1) *Data Owner*. Encrypt the plaintext M with his own public key PK_a to form a ciphertext C_{pka}, where M is the file that the data owner wants to give to the data requester

(2) *Data Owner*. The reencryption key $\text{RK}_{a\rightarrow b}$ is formed by encrypting and calculating his own public key PK_a and the public key of the data requester PK_b

(3) *Agent*. Use the key $\text{RK}_{a\rightarrow b}$ generated by the data owner to convert the ciphertext C_{pka} into the ciphertext C_{pkb} that can be decrypted by the private key of the data requester and forward it to the data requester

(4) *Data Requester*. Use the personal public key PK_b to decrypt the plaintext M

3. Results

To promote the information construction of the three-level diagnosis and treatment model of stroke medical treatment combination and to meet the need for doctors to quickly obtain patients' past medical history in stroke referral scenarios, this paper designs a blockchain-based stroke electronic medical record model. The purpose of the model construction includes (a) clarify the patient's ownership of medical record data and realize strict access control rights of patients to medical record data, (b) realize the privacy protection of patient identity and medical record data in the process of storage and sharing, (c) construct an efficient and secure sharing protocol to reduce user redundant operation, and (d) realize the secure storage of massive high-privacy electronic medical record data.

3.1. System Model. The stroke electronic medical record sharing model based on the medical treatment combination referral treatment scenario proposed in this paper is displayed in Figure 2, which describes how the proposed scheme is used to store and share electronic medical records and how the entities interact to implement each phase of the scheme.

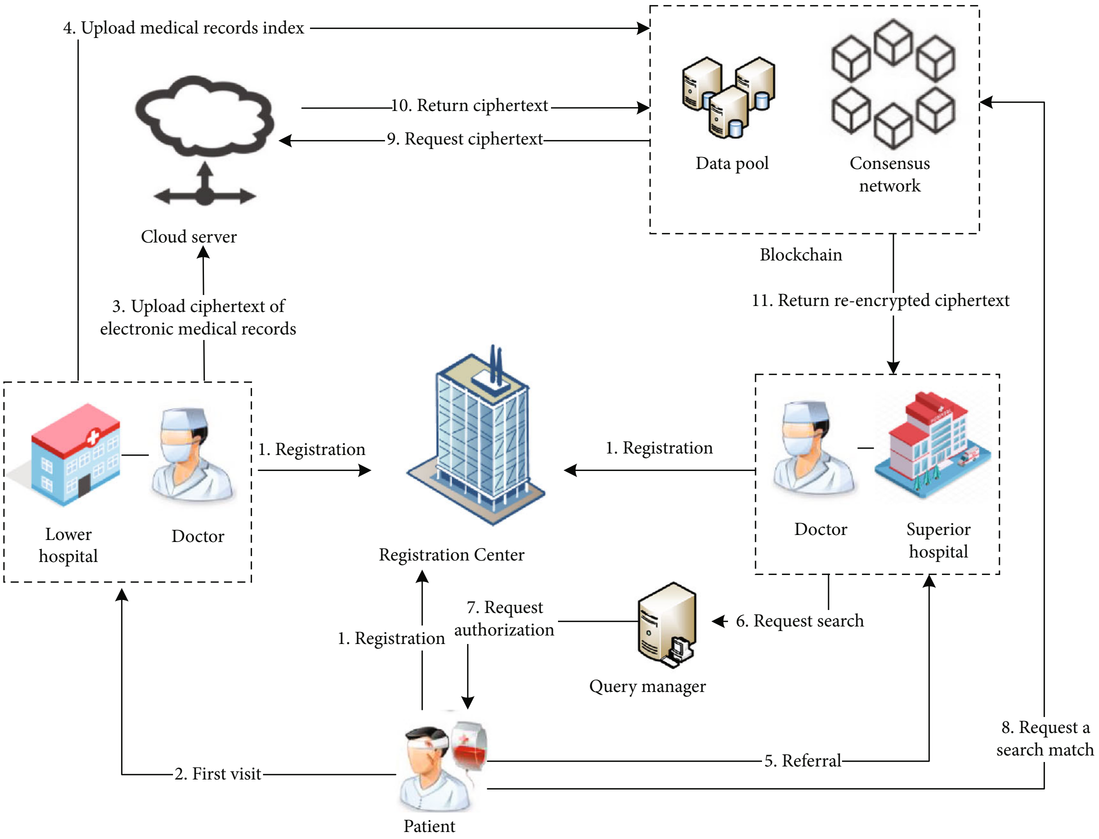


Figure 2: Blockchain-based sharing model for stroke electronic medical record.

The scheme contains six entities, which are registration center, query manager, patient, medical institution, consortium blockchain, and cloud server. These entities interact with each other to provide data control and protection service in the exchange of medical record information. There are four phases in the scheme proposed, user registration, data storage, request access, and request processing. In Figure 2, interaction 1 and interaction 2 are the user registration phase. Interactions 3 and 4 are the data storage phase, interactions 5, 6, 7, and 8 are the request access phase, and interactions 9, 10, and 11 are the request processing interaction.

(1) *Registration Center*. The entity is used to generate and store keys. The registration center is responsible for generating the public parameters of the system, that is, the master key when the system is initialized. In addition, when a user sends a request to join the blockchain to the registration center, the entity generates a corresponding key for the user requesting registration and sends it to the user through a secure channel to authenticate the user as a legitimate member of the system. Users' identity materials will be safely stored in the registration center, and the SHA256 hash of the public key will be transmitted to the consortium blockchain for backup.

(2) *Query Manager*. The entity receives a request and authenticates the user to determine that he is a legitimate user of the system. Meanwhile, the entity formats the request in a standard manner and then forwards it to other users or the consortium blockchain network.

(3) *Patient*. Patients should register in the system when they first visit the hospital for stroke, and the registration center allocates a public-private key pair for encrypting electronic medical records and a symmetric key for generating pseudonyms for each patient. During consultation and treatment, doctors generate pseudoidentities and electronic medical records for patients. When a patient goes to a higher-level hospital for referral treatment and needs to provide the hospital with his past electronic medical records, the patient authorizes the doctor to visit, generates a search trapdoor and reencryption key for the doctor, and sends them to the master node of consortium blockchain to request a search match. In the scheme, the right to use personal medical data is completely controlled by patients. Patients can grant doctors access to relevant data, set the time limit for record access, and revoke his authorization at any time.

(4) *Medical Institution*. The entity is responsible for uploading electronic medical records and is subject to the supervision and management of the regulatory

authorities. When medical institutions are registered in the blockchain, strict audit standards are required. The client of each medical institution is operated by doctors to create electronic medical records for patients. And then doctors encrypt, sign, and finally send the records to the cloud server. Medical record indexes and abstracts are sent to the consortium blockchain for storage on the chain. When a doctor at a higher-level hospital requests a medical record search, he needs to be authenticated to get patient authorization.

(5) *Cloud Server*. Due to the practical limitations of cost, storage capacity, and other factors, large-scale medical data is encrypted and stored outside the blockchain. The cloud server is used to store the ciphertext of electronic medical records uploaded by the doctor from the client, thereby reducing the storage pressure of the blockchain. When the user requests to search for data, the cloud server interacts with the blockchain. After receiving the ciphertext request from the blockchain, the ciphertext of the electronic medical record is returned to the blockchain master node for proxy reencryption.

(6) *Consortium Blockchain*. The consortium blockchain network is composed of nodes of various medical institutions. The stroke medical treatment combination includes multiple levels of medical institutions, including tertiary hospitals, second-level hospitals, and community hospitals. Each hospital acts as a node with different functions in the consortium blockchain network according to the level. They receive data requests and process them by assisting the query manager to verify the request. The consortium blockchain network uses an improved Practical Byzantine Fault Tolerant (PBFT) consensus mechanism to reach consensus among the nodes and add transaction orders to the blockchain distributed ledger. In the process of sharing medical records, after receiving the search trapdoor sent by the patient, the main node of the consortium blockchain performs search matching and sends a ciphertext request to the cloud server according to the matched index address. After receiving the ciphertext data returned by the cloud server and the reencryption key transmitted by the patient, the consortium chain master node acts as an agent to reencrypt the ciphertext data and convert it into a ciphertext that the doctor user can decrypt with the private key.

3.2. Scheme Phase

3.2.1. User Registration. The user registration flow chart is shown in Figure 3. A user sends a request to join the blockchain network from the client to the registration center. The registration center generates the corresponding keys for the entity and sends them to the user through a secure channel. The registration center generates a unique identification code ID for medical institutions in the consortium blockchain network, generates a public-private key pair for signatures for doctors, and generates a public-private key pair for encrypting the original data of electronic medical records and a symmetric key for generating pseudonyms for patients. When the patient visits a doctor, the doctor generates a pseudoidentity for the patient. The user's identity material will be safely stored in the registration center, and the SHA256 hash of the public key will be transmitted to the blockchain for backup, authenticating the user as a legal member of the system. Considering that there is a situation where the patient is seriously ill and cannot operate the system, this scheme allows the family members of the patient to act as agents.

3.2.2. Data Storage. The data storage flow chart is shown in Figure 4. The electronic medical records storage is divided into two steps: on-chain storage and off-chain storage. First, when a patient first visits a lower-level hospital, the attending doctor A generates an electronic medical record, then encrypts it with the patient's public key, and attaches it to the signature of his private key, and transmits the record to the cloud server for storage. After receiving the storage address returned by the cloud server, doctor A extracts the medical record keywords, the pseudoidentity pseudonym of the patient, and the IP storage address of the medical record in the cloud, and then executes a searchable encryption algorithm to generate a secure index. Doctor A combines the index, his signature, personal public key, and other information to form a transaction order and sends it to the consortium blockchain for broadcasting. The node network completes the transaction on the chain through the improved PBFT consensus mechanism and realizes the data synchronization between the nodes in the chain. Doctor A who generated the medical record has read and written authority to the medical record, and there is no need to obtain a request from the patient during access.

3.2.3. Request Access. Request access flow chart is shown in Figure 5. When a patient is referred to a higher-level hospital, the attending doctor B sends a medical record access request signed with his private key from the client. After the query manager receives the message, it combines with the blockchain to perform hybrid verification. The query manager retrieves the doctor's public key from the blockchain and verifies the signature on the request. If the signature verification is successful, the query manager forwards the access request and doctor B's public key to the patient, and the patient authorizes doctor B to access the medical records of the lower-level hospital and sets the access period. If the patient has been treated in a community hospital and a second-level hospital and referred to a third-level hospital, the doctors in the third-level hospital should be granted access to the medical records of the two lower-level hospitals. Subsequently, the patient executes an encryption algorithm to generate a search trapdoor and uses his private key and doctor B's public key to generate a reencryption key. The patient signs and packages the trapdoor, reencryption key, and

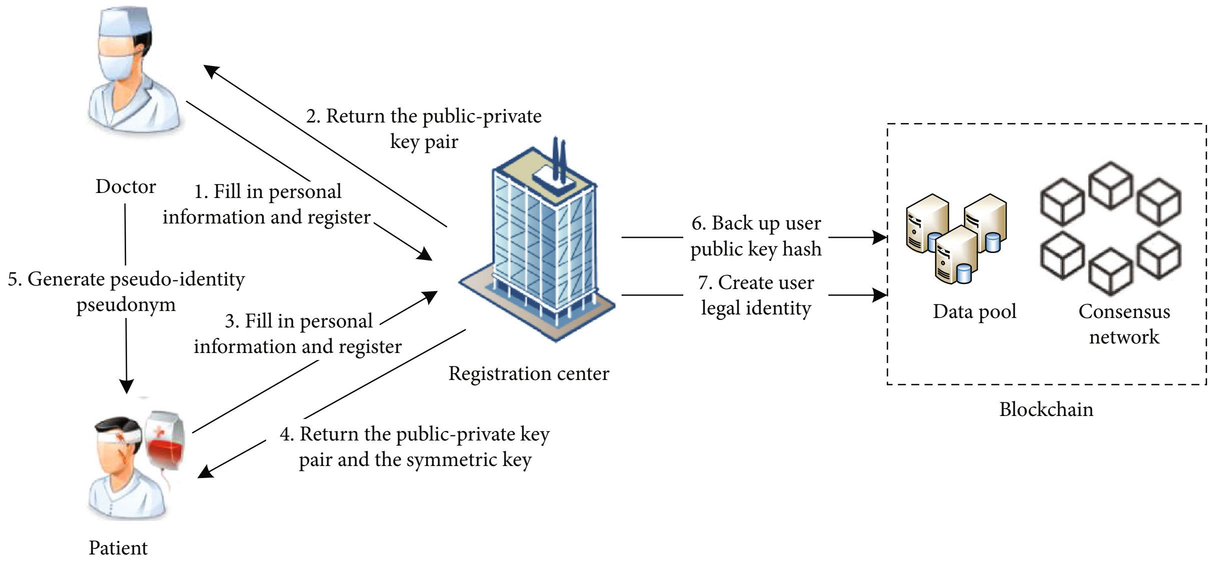

2. Return the public-private key pair
Doctor
1. Fill in personal information and register
5. Generate pseudo-identity pseudonym
3. Fill in personal information and register
Registration center
6. Back up user public key hash
7. Create user legal identity
Data pool
Consensus network
Blockchain
4. Return the public-private key pair and the symmetric key
Patient

Figure 3: User registration.

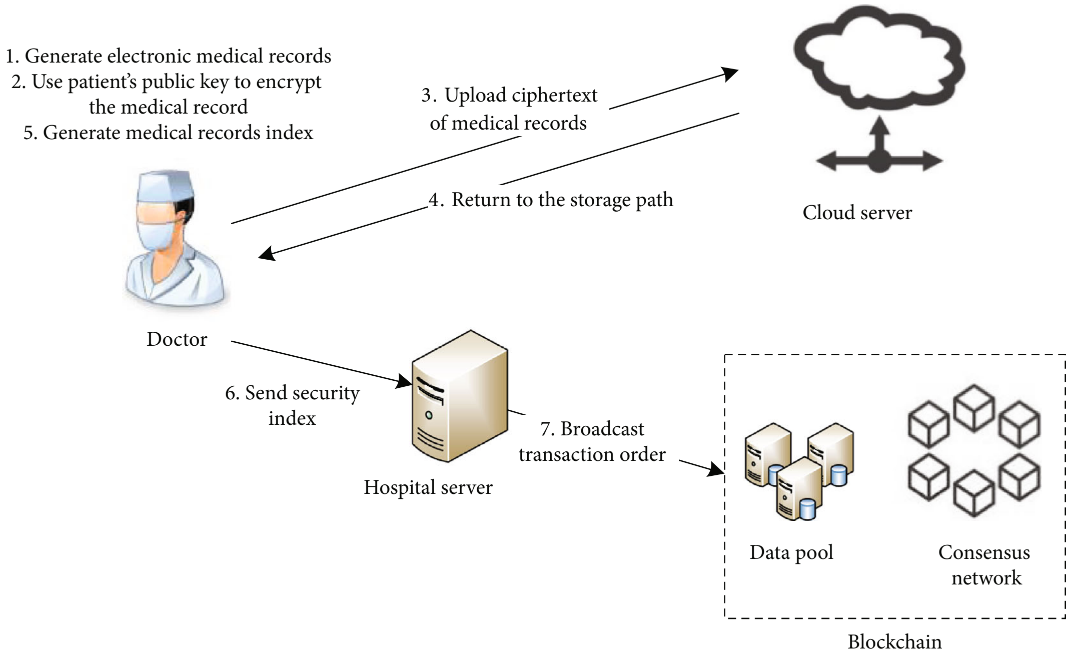

1. Generate electronic medical records
2. Use patient's public key to encrypt the medical record
5. Generate medical records index
3. Upload ciphertext of medical records
4. Return to the storage path
Cloud server
Doctor
6. Send security index
Hospital server
7. Broadcast transaction order
Data pool
Consensus network
Blockchain

Figure 4: Data storage.

doctor's request and then sends them to the master node of the consortium blockchain for search matching and proxy reencryption. The master node verifies the requester's authority and executes the next stage of request processing operations.

3.2.4. Request Processing. Request processing flow chart is shown in Figure 6. The master node of the consortium blockchain is responsible for processing the request made by system users. After receiving the search trapdoor and visit request, the master node executes the matching algorithm and sends a data request containing the specified IP storage address to the cloud server according to the search result. After receiving the returned medical record ciphertext, the master node performs a two-step verification process. First, it verifies whether the medical record returned has been authorized to be accessed by doctor B. If authorized, it verifies whether the medical record is complete. The master node hashes the ciphertext of the medical record. If the value matches the digital digest in the block transaction sheet, then it is confirmed that the medical record file has not been tampered with, and the reencryption operation can be performed. Then, the master node uses the proxy reencryption key to reencrypt the verified ciphertext of the medical record

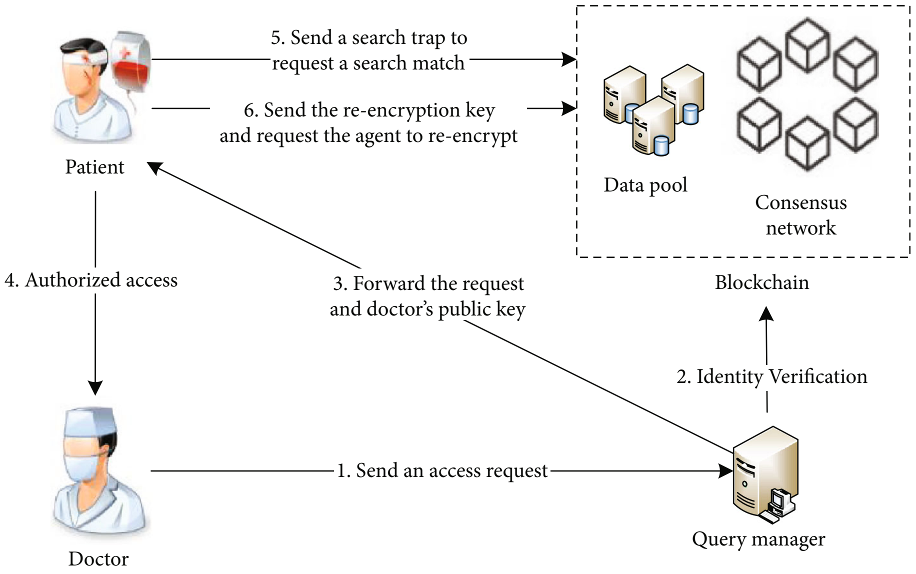

5. Send a search trap to request a search match
6. Send the re-encryption key and request the agent to re-encrypt
Patient
Data pool
Consensus network
Blockchain
4. Authorized access
3. Forward the request and doctor's public key
2. Identity Verification
1. Send an access request
Doctor
Query manager

Figure 5: Request access.

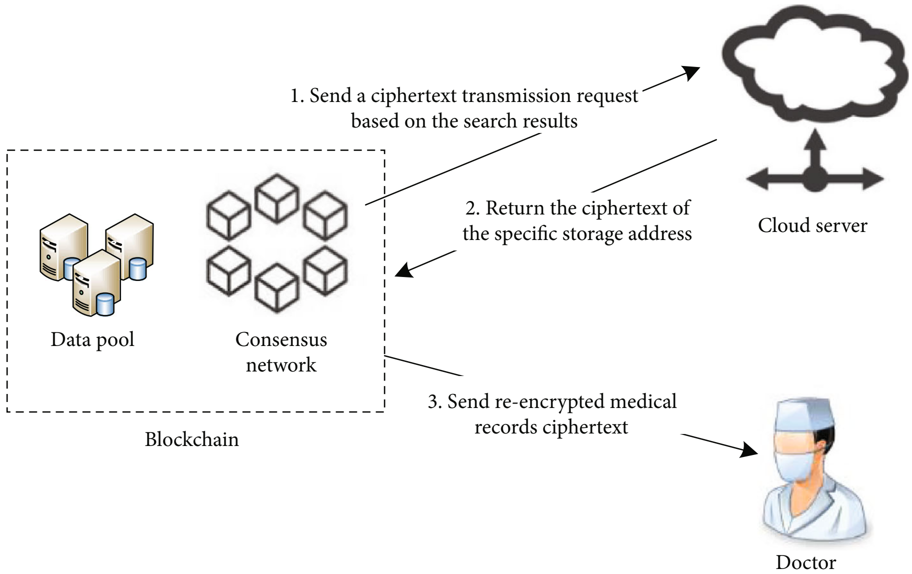

1. Send a ciphertext transmission request based on the search results
2. Return the ciphertext of the specific storage address
Cloud server
Data pool
Consensus network
Blockchain
3. Send re-encrypted medical records ciphertext
Doctor

Figure 6: Request processing.

and transmit it to doctor B who requests access. When the patient has finished treatment in a tertiary hospital, he can choose to withdraw doctor B's permission to view the medical records of the lower-level hospital. The details of all transactions are formed into blocks and transmitted to the consortium blockchain.

3.3. System Consensus Mechanism. Practical Byzantine Fault Tolerance (PBFT) algorithm, as a state machine copy replication algorithm, can provide $(n-1)/3$ fault tolerance (n is the total number of nodes in the blockchain network). The mechanism can not only start and run on fewer nodes but also does not require a lot of computing power to maintain. Considering that the number of medical institutions in the medical treatment combination is small, compared with Proof of Work (POW), Proof of Stake (POS), and Delegated Proof of Stake (DPOS) that need to rely on tokens, the PBFT mechanism is more suitable for the application scenarios of the medical treatment combination blockchain.

The consortium blockchain node in this scheme is composed of medical institutions in the medical treatment combination, which is divided into two types of functional nodes according to the hospital level. High-level medical institutions such as tertiary hospitals and secondary hospitals will serve as accounting nodes due to their high server computing capabilities. Accounting nodes package the requests submitted by the medical institutions into medical data blocks and sign them with their own private keys. Small hospitals such as community hospitals are used as verification nodes to verify data blocks submitted to the consortium blockchain network and supervise behaviors related to medical data blocks. Each node reaches a consensus agreement through the improved PBFT consensus mechanism. According to the actual needs of the medical treatment combination, the improved PBFT mechanism based on the original PBFT mechanism changes the selection method of the main node from the whole network selection to fixed node rotation to improve the reliability and safety of the system. After the implementation of the improved mechanism, the secondary and tertiary hospital nodes in the medical treatment combination take turns as the master node, responsible for registering and storing data blocks in the blockchain. The improved PBFT principle is shown in Figure 7.

The improved PBFT algorithm includes five processes: request, prepreparation, preparation, confirmation, and reply. The description of each process is as follows.

(1) *Request*. When the doctor uploads the medical record metadata to the blockchain, he sends a signed data upload request message from the client to the master node of the consortium blockchain and submits his personal public key as an identifier.

(2) *Prepreparation*. To verify whether the signature in the request message is correct, the master node extracts the doctor's public key, decrypts the doctor's signature in the transaction sheet to obtain the ciphertext hash value of the electronic medical record, and compares it with the digital abstract value in the transaction sheet. If the comparison results are consistent, it means that the electronic medical records are stored completely and have not been forged or tampered with. After successful verification, the master node numbers the request message and broadcasts it to all slave node members.

(3) *Preparation*. After receiving the message verification request sent by the master node, the slave node needs to pass the verification and broadcast the preparation message to other nodes except itself. If the comparison result is different, it means that the data has been tampered with, and the node will not broadcast the preparation message.

(4) *Confirmation*. The slave node sends a confirmation message to all nodes except itself and enters the reply stage after receiving $2f + 1$ confirmation messages including itself.

(5) *Reply*. The node sends a reply message to the doctor. Only when $f + 1$ nodes receive the reply message, the request is considered executed successfully. If it is less than $f + 1$, the verification failure result will return to the user. The node checks the number of transactions in the block every one minute. If the number reaches the specified number, the node needs to form a data block and calculates the Merkle root of the block. When the specified time is reached, the node anchors the Merkle roots of all newly generated blocks to the blockchain.

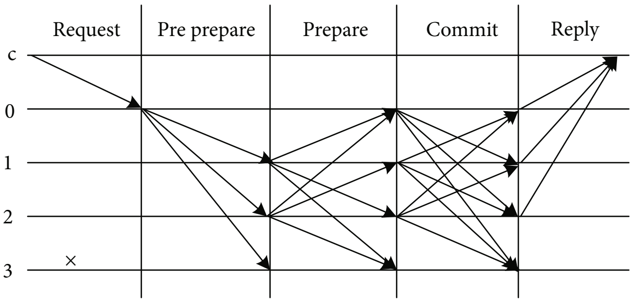


Figure 7: PBFT consensus mechanism.

3.4. Block Structure. The block of the consortium blockchain proposed in this paper is composed of block header and block body. The block structure is shown in Figure 8. The block header includes block ID, block size, hash of the previous block, hash value of Merkle tree root, timestamp, and digital signature. The root hash value of the Merkle tree is the SHA256 hash value of the transaction content to ensure that the transaction sheet has not been modified. The digital signature is the signature of the block producer, confirming that the block has passed the verification; the timestamp is used to display the time when the block was generated. The content of the block body is the transaction information of the block, including transaction ID, medical record index, digital abstract, doctor's signature, and hospital server's signature. The doctor's signature refers to the digital signature information of the medical staff who generates the electronic medical record for the patient, which is used to ensure that the data can achieve accountability. The digital abstract is the hash value of the ciphertext of the medical record, which is used to ensure that the files stored in the distributed ledger cannot be forged and tampered with. The doctor's public key is the asymmetric encryption public key of the doctor who generates the electronic medical record for the patient and is used to decrypt the digital signature for block verification. The medical record index is generated by an encryption algorithm, and the information includes the IP address of the medical record in the cloud server, medical record keywords, and patient pseudoidentity ID.

3.5. Sharing Protocol. The sharing protocol proposed in this paper consists of four stages: system initialization, user registration, data upload, and data sharing. Table 1 describes the symbols and their meanings in the protocol.

3.5.1. System Initialization. The system initialization algorithm is shown as Equation (5). The algorithm is executed by the registration center. The registration center enters the security parameter λ and selects two multiplicative cyclic groups G_0 and G_1 with prime order p, g is the generator of

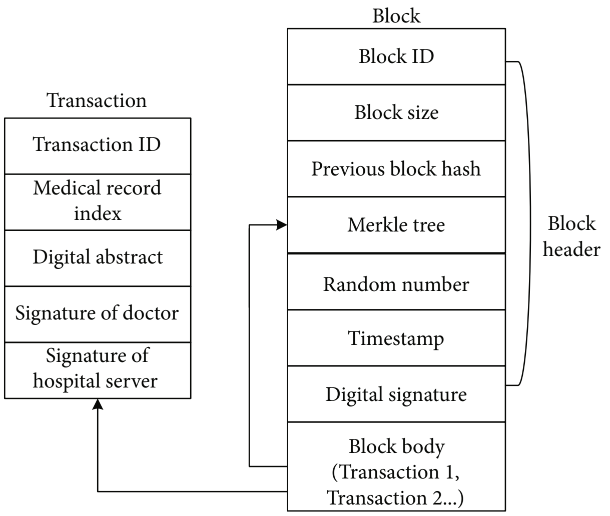


Figure 8: Block structure of medical consortium blockchain.

Table 1: Scheme symbol.

Symbol	Description
PP	System public parameters
MSK	System master private key
P_k, S_k	Public-private key pair of users
$\mathrm{SK}_{\mathrm{aes}}$	Symmetric key of user
P_{k_peks}, S_{k_peks}	Searchable encryption public-private key pair
UID	Patient ID
MD	Plain text of electronic medical record
EMD	Ciphertext of electronic medical record
x_i	Digital signature of doctor
W	Electronic medical record keyword set
PID_a	Pseudoidentity
I	Electronic medical record security index
IP	Cloud storage path of ciphered medical records
T_w	Search trapdoor
RK	Reencryption key
EMD$'$	Reencrypted ciphertext

G_0, and then defines a bilinear mapping, $e : G_0 * G_0 \longrightarrow G_1$. Finally, the equation outputs the system public parameters PP and the system master private key MSK.

$$(\mathrm{PP}, \mathrm{MSK}) = \mathrm{Setup}\left(1^{\lambda}\right). \quad (5)$$

3.5.2. User Registration

(1) *User Key Generation*. The algorithm is shown as Equation (6). It is executed by the registration center. The registration center inputs the public parameters PP, the master private key MSK, and outputs the public and private key pair P_{ka}, S_{ka} assigned to the patient, the symmetric key $\mathrm{SK}_{\mathrm{aes}}$ used to generate pseudonyms, and the public and private key pair P_{ki}, S_{ki} assigned to the doctor for signature. After the user requests registration and completes the personal information on the system client, the registration center sends the key to the user client through a secure channel, and at the same time hashes the SHA256 of the public key and sends it to the blockchain for copy storage

$$(P_k, M_k) = \mathrm{KeyGen}(\mathrm{PP}, \mathrm{MSK}). \quad (6)$$

(2) *Searchable Encryption Key Generation*. The algorithm is shown as Equation (7). It is executed by the patient. The patient enters the security parameter λ and outputs the searchable encrypted public key P_{k_peks} and the searchable encrypted private key S_{k_peks}. Then, the patient uploads the searchable encrypted public key P_{k_peks} to the registration center for storage, and the copy is transmitted to the blockchain for backup

$$\left(P_{k\_\mathrm{peks}}, S_{k\_\mathrm{peks}}\right) = \mathrm{KeyGen}(\lambda). \quad (7)$$

(3) *Pseudoidentity Generation*. The algorithm is shown as Equation (8). It is executed by the doctor. The patient encrypts the symmetric key with the doctor's public key and transmits it to his doctor, and the doctor uses the private key to decrypt the symmetric key. The doctor enters the patient's symmetric key $\mathrm{SK}_{\mathrm{aes}}$ and the identity code UID generated by the hospital server when the patient is registered and performs a series hash operation to output the pseudoidentity pseudonym of the patient

$$\mathrm{PID}_a = \mathrm{NameGen}(\mathrm{SK}_{\mathrm{aes}}, \mathrm{UID}). \quad (8)$$

3.5.3. Data Upload

(1) *Document Encryption*. The algorithm is shown as Equation (9). It is executed by the doctor. Doctor inputs the system public parameters P_K, the plaintext document MD, the patient's public key P_{ka}, and outputs the ciphertext medical record document EMD. And then the doctor uploads the ciphertext medical record to the cloud server, and the cloud server generates a unique ID value for each ciphertext medical record

$$\mathrm{EMD} = \mathrm{MDEnc}(P_K, \mathrm{MD}, P_{\mathrm{ka}}). \quad (9)$$

(2) *Signature Generation*. The algorithm is shown as Equation (10). It is executed by the doctor. The doctor inputs the patient's ciphertext medical record document EMD and his own private key S_{ka} and outputs the doctor's digital signature x_i. The encryption process is specifically to perform a hash calculation on the ciphertext to extract its digital digest and then use the doctor's private key to encrypt the digital digest to form a digital signature

$$x_i = \mathrm{SignGen}(\mathrm{EMD}, S_{\mathrm{ka}}). \tag{10}$$

(3) *Index Generation*. The algorithm is shown as Equation (11). It is executed by the doctor. The doctor enters the storage path IP of the encrypted medical record EMD, the searchable encrypted public key P_{k_peks}, the medical record keyword set W, and the patient's pseudoidentity PID_a and output the security index I. This process first performs a hash operation on the medical record keyword W and the patient's pseudoidentity PID_a to obtain $H(W)$ and $H(\mathrm{PID}_a)$ and then selects a random number $r \longleftarrow Z_p$ to calculate the hash result and the searchable encryption public key to obtain $H(W)^r$, $H(\mathrm{PID}_a)^r$, and $P_{k\_\mathrm{peks}}{}^r$, and then the index I is calculated. The index is used for keyword search and matching in the data sharing stage

$$I = \mathrm{IndexGen}\left(P_{k\_\mathrm{peks}}, W, \mathrm{PID}_a, \mathrm{IP}\right). \tag{11}$$

(4) *Data Upload*. Hospital server extracts the security index I, doctor's digital signature x_i, and doctor's asymmetric encryption public key P_{ki} to construct a new transaction sheet, and adds data such as timestamp, medical record hash value, random number, and hospital's signature, and finally packs it into blocks for broadcasting

3.5.4. Data Sharing

(1) *Trapdoor Generation*. The algorithm is shown as Equation (12). It is executed by the patient. At the user registration stage, the patient obtains the searchable encrypted private key S_{k_peks}. Patients enter the public parameter PP, the query keyword K_w, the patient's pseudoidentity K_{PID}, and the searchable encryption private key S_{k_peks}. The algorithm first hashes the keywords or the patient's pseudoidentity, then selects a random number, and obtains the search trapdoor T_w after encryption calculation. The patient sends a search request to the consortium blockchain master node after generating a search transaction sheet

$$T_w = \mathrm{TokenGen}\left(PP, K_w, S_{k\_\mathrm{peks}}\right). \tag{12}$$

(2) *Search Match*. The algorithm is shown as Equation (13). It is executed by the main node of the consortium blockchain. The main node extracts the trapdoor T_w after receiving the search request sent by the patient and traverses the matching trapdoor T_w and the security index I of the blockchain database. If the matching is successful, the IP of the medical record containing the keyword C_W is obtained. The master node interacts with the cloud server and requests the cloud server to return the ciphertext of the medical record of the IP storage path

$$\mathrm{IP} = \mathrm{Query}\left(T_w, I, P_{k\_\mathrm{peks}}\right). \tag{13}$$

(3) *Reencryption Key Generation*. The algorithm is shown as Equation (14). It is executed by the patient. Patient inputs his private key S_{ka} and the public key P_{ki} of the doctor requesting access, outputs the reencryption key RK, and sends it to the main node of the consortium blockchain

$$\mathrm{RK} = \mathrm{RKGen}(S_{\mathrm{ka}}, P_{\mathrm{ki}}). \tag{14}$$

(4) *Proxy Reencryption*. The algorithm is shown as Equation (15). It is executed by the main node of the consortium blockchain. The main node inputs the reencryption key RK and the original ciphertext medical record document EMD and outputs the reencrypted ciphertext EMD$'$. The reencryption process converts the EMD that needs to be decrypted with the patient's private key into EMD$'$, which can be directly decrypted with the doctor's private key. The master node sends EMD$'$ to the doctor requesting access

$$\mathrm{EMD}' = \mathrm{ReEnc}(\mathrm{RK}, \mathrm{EMD}). \tag{15}$$

(5) *Data Decryption*. The algorithm is shown as Equation (16). It is executed by the doctor requesting access. The doctor inputs the reencrypted ciphertext EMD$'$ and his own private key P_{ki} and outputs the plaintext electronic medical record document MD

$$\mathrm{MD} = \mathrm{Dec}\left(\mathrm{EMD}', P_{\mathrm{ki}}\right). \tag{16}$$

4. Discussion

4.1. Scheme Comparison. The paper selects the existing electronic medical record sharing scheme based on blockchain

Table 2: Comparison and analysis of schemes.

Properties	Pournaghi [6]	Zhang [14]	Luo [19]	Xu [26]	Chen [27]	Wang [28]	This paper
Application scenario	Personal life cycle management	Patient-oriented medical record sharing	Sharing of medical records among hospitals	Cross-domain visits	Clinical and scientific research	Sharing of medical records among hospitals	Sharing of referral medical records in medical treatment combination
Based on blockchain	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fine-grained access control	Yes	Yes	Yes	No	No	No	No
Patient anonymous	No	Yes	No	Yes	No	No	Yes
Keywords searchable	No	No	No	No	Yes	Yes	Yes
Consensus mechanism	PBFT	PBFT	DPOS	—	—	Improved PBFT	Improved PBFT
Main chain pressure	Small	Big	Small	Small	Small	Small	Small
Smart contract	Yes	No	No	Yes	Yes	No	No

to compare with the scheme proposed in this paper and analyze the advantages and disadvantages of the schemes. Table 2 analyzes and evaluates the selected schemes from eight aspects. Compared with the paper [14] that stores all medical electronic records on the blockchain, this paper uses cloud storage technology to reduce the pressure on blockchain storage and meet the actual deployment requirements. Compared with the DPOS consensus adopted in paper [19], the improved PBFT algorithm adopted in this paper has a low CPU occupancy rate, and it does not require a large amount of computing power to maintain. It is suitable for early exploration and later expansion of the medical blockchain. The application scenarios of this paper are similar to those in paper [26], which are sharing electronic medical records during patient referral and treatment. In this scenario, data access personnel are limited to the attending doctor, and there is no special requirement for fine-grained access control. In the process of data access, paper [27] requires the patient to transmit the private key to the doctor. The transmission process cannot guarantee the security of the private key, and it may be obtained by malicious nodes. This paper adopts the proxy reencryption method; the doctor can use the personal private key to decrypt the matched file, avoiding the possibility of private key leakage. Compared with the paper [6, 28], this paper supports the anonymity of patients and supports the search for anonymous pseudoidentities. Through accurate retrieval, it can quickly match medical records and resist node guessing attacks, thereby improving user identity security. Through comparison, it is found that this paper can be improved in the application of smart contracts. In the future work, we will consider the introduction of smart contracts to automatically execute the processing and response to requests, thereby improving transaction efficiency.

4.2. Scheme Analysis and Evaluation. This section analyzes whether the proposed scheme can resist internal and external attacks from three perspectives of medical record integrity, user privacy, and data security.

4.2.1. Medical Record Integrity. The scheme proposed in this paper can resist electronic medical record forgery and deletion attacks. In traditional cloud-based medical record storage schemes, the cloud itself is an untrusted third party, and electronic medical records are vulnerable to forgery and erasure attacks, which cannot be detected. In the real world, to improve process efficiency, patients usually do not need to sign electronic medical records, but instead authorize doctors to directly generate and store electronic medical records. Therefore, doctors try to forge or delete electronic medical records that have been outsourced to cloud servers to cover up their medical accidents. In the blockchain-based electronic medical record sharing scheme, the electronic medical record index generated by the doctor is integrated into the transaction of the basic blockchain, and the transaction is accompanied by the hash value of the medical record ciphertext. Any change to the original data will result in a change in the hash value, thereby ensuring that the electronic medical record is unchangeable and traceable. Suppose that doctors collude with cloud servers to replace the real documents stored in the system with forged medical records. If doctors want to modify the current medical record hashes stored in the blockchain system at the same time, they must imitate the main chain like the source chain so that the blockchain transactions containing the transaction corresponding to the hash of the forged medical record can be accepted by most nodes. Due to the considerable computing power, this scenario is almost impossible to achieve. Without modifying the corresponding medical record hash, the master node will hash the data before being accessed. According to the hash rules, hash values obtained by hashing two different files are different; therefore, the forged files will not be shared and accessible. For file tampering attacks, unless the blockchain is threatened by 51% attacks, the data stored in the blockchain is immutable. The main attributes of the blockchain ensure the correctness and completeness of the electronic medical records in the cloud server.

4.2.2. User Privacy. In the scheme proposed in this paper, three ways are used to ensure that user data privacy is not violated.

The first way is to adopt the encryption scheme that combines the proxy reencryption mechanism and the searchable encryption mechanism. In the scheme, the electronic medical record is encrypted by the patient's asymmetric encryption public key, and it is randomly stored anywhere in the cloud server. The blockchain only stores the index address. Therefore, to access the plain text of the electronic medical record, the requester must obtain the patient's private key and IP storage path of medical records in the cloud. In this paper, the method of proxy reencryption is adopted in the process of medical record sharing. The master node of the consortium blockchain converts the ciphertext of the medical record encrypted with the patient's public key into the ciphertext that the doctor can decrypt with his personal private key. The sharing process will not reveal the patient's private key at all. Without the patient's private key, any ciphertext information in the blockchain network cannot be encrypted. The data storage path is encrypted and stored in the consortium blockchain. Only entities with search trapdoors generated by the patient can access the medical record storage path in the cloud. Through the searchable encryption mechanism, patients have complete control over the access rights of their electronic medical record files.

The second way is to set up the access control mechanism. Access to all medical records on the medical consortium blockchain is managed by personnel, which can prevent malicious access to medical information from the source. When a doctor sends an access request, the query manager strictly verifies the identity of the visitor and prevents unauthorized behavior. Only authenticated and authorized users can access and retrieve medical record data from a specific path.

The third way is to set up pseudoidentity pseudonyms for patients. This allows all patient data to be associated with the pseudonym generated using his symmetric key, and the generated electronic medical records will not have any connection with the patient's actual identity. Therefore, malicious nodes cannot obtain the true identity of the data owner during the data sharing process. Since the pseudoidentity pseudonym generated by each attending doctor for the patient is random, the attacker cannot determine that multiple medical records are from the same patient, and the relationship between different electronic medical record data cannot be established, which ensures that the true identity of the patient cannot be traced.

4.2.3. Data Security. The searchable encryption mechanism adopted by the scheme in this paper needs to send information including keyword indexes and search trapdoors to the blockchain. To ensure the security of the above information, the scheme process is analyzed to ensure that it can resist attacks from malicious nodes. First, in the index generation process, it is necessary to input the keywords of the medical records. The doctor selects a random number to calculate the keyword hash and uses the searchable encryption public key to independently execute the encryption algorithm for each keyword. Due to the uncertainty of random numbers, it is impossible for a malicious attacker to derive keywords from the keyword ciphertext. Secondly, when accessing data, the patient needs to send the search trapdoor to the blockchain node. During the construction of the search trapdoor, different keywords are used to encrypt different keywords, so the keywords can be hidden. During the sharing process, no plaintext information of the electronic medical record will be displayed.

5. Conclusions

The electronic medical record of stroke is the core information resource in the referral process, and its safe sharing will effectively promote doctors to accurately grasp the patient's condition. In response to the specific needs of stroke data sharing in the medical treatment combination referral scenario, this paper has provided a secure sharing protocol for electronic medical records based on the consortium blockchain. The proposed scheme has adopted a storage method combining cloud and consortium blockchain. Cloud server is used to store the ciphertext of the original medical records, and the blockchain saves traceable log information and medical record index. The proposed scheme has classified the medical institutions in medical treatment combination and improved the preselection node mechanism of the PBFT consensus algorithm to improve the reliability and security of the system. By setting the pseudoidentity pseudonym of the patient and adopting the searchable proxy reencryption scheme based on the public key encryption scheme to realize data privacy and keyword hiding. Finally, the paper has analyzed the integrity of medical records, user privacy, and data security. The results showed that the proposed scheme can resist the tampering attack of doctors and semitrusted cloud and guessing attacks of malicious nodes on patient identity and privacy. The above research shows that the scheme in this paper has the following advantages. The scheme has increasing flexibility and scalability in storage capacity and security in data privacy protection. Meanwhile, the scheme ensures that patients have ownership of medical records and provides patients with instant revocation of the right to access in access control. The scheme in this paper can also be applied to medical treatment combination for other diseases besides stroke. However, the scheme has not yet implemented access control for individual parts of the medical record, which will be part of future research. In consideration of improving the system execution efficiency, the next step of the research work is to introduce self-executing and self-verifying smart contracts and build a stroke consortium blockchain system to further refine the functional modules of client users.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

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

Acknowledgments

This study is supported by the key project grant from the Science and Technology Commission of Beijing, China. The project is named the model and demonstration application of a medical consortium for the collaborative prevention and treatment of major diseases in Beijing's medical reform. The grant number is Z131100005613017.

References

[1] Report on stroke prevention and treatment in China Writing Group, "Summary of China stroke prevention and treatment report 2019," *Journal of Chinese Cerebrovascular Diseases*, vol. 17, no. 5, pp. 272–281, 2020.

[2] State Council of the PRC, "Outline of Healthy China 2030 Plan," *Chinese Cancer*, vol. 28, no. 10, p. 724, 2019.

[3] S. F. Zhang, X. Han, D. H. Wu et al., "The establishment of a new model of regional stroke management based on the intelligent medical consortium platform," *Journal of Fudan University (Medical Sciences)*, vol. 45, no. 6, pp. 805–810, 2018.

[4] S. M. Martínez Monterrubio, J. Frausto Solis, and R. Monroy Borja, "EMRlog method for computer security for electronic medical records with logic and data mining," *BioMed Research International*, vol. 2015, Article ID 542016, 12 pages, 2015.

[5] D. Banciu, M. Radoi, and S. Belloiu, "Information security awareness in Romanian public administration: an exploratory case study," *Studies in Informatics and Control*, vol. 29, no. 1, pp. 121–129, 2020.

[6] S. M. Pournaghi, M. Bayat, and Y. Farjami, "MedSBA: a novel and secure scheme to share medical data based on blockchain technology and attribute-based encryption," *Journal of Ambient Intelligence and Humanized Computing*, vol. 5, 2020.

[7] A. A. al-Absi, N. A. al-Sammarraie, W. M. Shaher Yafooz, and D. K. Kang, "Parallel MapReduce: maximizing cloud resource utilization and performance improvement using parallel execution strategies," *BioMed Research International*, vol. 2018, Article ID 7501042, 17 pages, 2018.

[8] Y. J. Song, "Blockchain-based power trading process," *Journal of System and Management Sciences*, vol. 9, no. 3, pp. 78–91, 2019.

[9] A. Azaria, A. Ekblaw, T. Vieira, and A. Lippman, "MedRec: using blockchain for medical data access and permission management," in *International Conference on Open & Big Data*, pp. 25–30, 2016.

[10] L. H. Zhang, F. Lan, P. P. Jiang, and T. F. Jiang, "Medical record safe storage and sharing scheme based on dual blockchain," *Computer Engineering and Science*, vol. 41, no. 9, pp. 1581–1587, 2019.

[11] R. Kumar, N. Marchang, and R. Tripathi, "Distributed off-chain storage of patient diagnostic reports in healthcare system using IPFS and blockchain," in *2020 International Conference on COMmunication Systems & NETworkS (COMSNETS)*, pp. 1–5, 2020.

[12] H. Wang and M. M. Zhou, "A safe storage model of medical information based on blockchain," *Computer Science*, vol. 46, no. 12, pp. 174–179, 2019.

[13] Y. Chen, S. Ding, Z. Xu, H. D. Zheng, and S. L. Yang, "Blockchain-based medical records secure storage and medical service framework," *Journal of Medical Systems*, vol. 1, p. 43, 2019.

[14] Y. B. Zhang, M. Cui, L. J. Zheng et al., "Research on electronic medical record access control based on blockchain," *International Journal of Distributed Sensor Networks*, vol. 15, no. 11, 2019.

[15] L. Hang, E. Choi, and D. H. Kim, "A novel EMR integrity management based on a medical blockchain platform in hospital," *Electronics*, vol. 8, no. 4, p. 467, 2019.

[16] S. Tanwar, K. Parekh, and R. Evans, "Blockchain-based electronic healthcare record system for healthcare 4.0 applications," *Journal of Information Security and Applications*, vol. 50, article 102407, 2020.

[17] J. W. Liu, X. L. Li, L. Ye, H. Zhang, X. Du, and M. Guizani, "BPDS: a blockchain based privacy-preserving data sharing for electronic medical records," in *2018 IEEE Global Communications Conference (GLOBECOM)*, Abu Dhabi, United Arab Emirates, December 2018.

[18] S. F. Niu, L. X. Chen, J. F. Wang, and F. Yu, "Electronic health record sharing scheme with searchable attribute-based encryption on blockchain," *IEEE Access*, vol. 8, pp. 7195–7204, 2020.

[19] W. J. Luo, S. L. Wen, and Y. Cheng, "Blockchain-based electronic medical record sharing scheme," *Computer Applications*, vol. 40, no. 1, pp. 157–161, 2020.

[20] S. Nakamoto, "Bitcoin: a peer-to-peer electronic cash system," 2018, http://bitcoin.org/bitcoin.pdf.

[21] E. Chukwu and L. Garg, "A systematic review of blockchain in healthcare: frameworks, prototypes, and implementations," *IEEE Access*, vol. 8, pp. 21196–21214, 2020.

[22] D. X. Song, "Practical techniques for searches on encrypted data," in *Proc. 2000 IEEE Symposium on Security and Privacy (SP'00). IEEE Computer Society*, pp. 44–55, 2000.

[23] D. Boneh, G. D. Crescenzo, R. Ostrovsky, and G. Persiano, "Public key encryption with keyword search," in *Advances in Cryptology - EUROCRYPT 2004, International Conference on the Theory and Applications of Cryptographic Techniques*, Proceedings. Springer-Verlag, Interlaken, Switzerland, 2004.

[24] M. Blaze, G. Bleumer, and M. Strauss, *Divertible protocols and atomic proxy cryptography*, International Conference on the Theory and Applications of Cryptographic Techniques, 1998.

[25] L. Li, Q. X. Zeng, Y. H. Wen, and S. C. Wang, "Data sharing scheme based on blockchain and proxy re-encryption," *Information Security*, vol. 20, no. 8, pp. 16–24, 2020.

[26] J. Xu, Z. M. Chen, P. Gong, and K. K. Wang, "A secure storage and access scheme for medical records based on blockchain network," *Computer Applications*, vol. 39, no. 5, pp. 1500–1506, 2019.

[27] L. Chen, W. K. Lee, C. C. Chang, K. K. R. Choo, and N. Zhang, "Blockchain based searchable encryption for electronic health record sharing," *Future generation computer systems*, vol. 95, pp. 420–429, 2019.

[28] H. Wang, Y. X. Liu, S. Y. Cao, and M. M. Zhou, "Medical data storage mechanism incorporating blockchain technology," *Computer Science*, vol. 47, no. 4, pp. 285–291, 2020.